CHAPTER
2

MOS Transistors

CHAPTER OUTLINE

2.1 Introduction

This chapter focuses on the semiconductor devices used in mainstream integrated circuits. MOS devices, *pn* junctions, and device capacitances are all covered here. We will spend much of this chapter examining basic physics and modeling of MOS transistors. It is an important chapter in that it lays the foundation for many of the design equations to be used throughout the book. Of course, it is difficult to have complete coverage of the background needed to grasp the device physics in one chapter. Readers who are unfamiliar with the terms and concepts presented here should consult the references for the prerequisite material. It is expected that many students embarking upon a study of this text will have had a course in semiconductor devices, including the MOS transistor. However, even these students should study this subject matter because the emphasis is on the specific characteristics of MOS devices that are important to VLSI digital circuit design.

We begin with MOS device characteristics that lead to the derivation of the threshold voltage equation, current equations, and capacitance models. MOS technology is the basis for most of the very large-scale integrated (VLSI) digital memory and microprocessor circuits. It is the dominant technology in the IC industry today, with bipolar technology a distant second. The most important advantage of MOS circuits over bipolar circuits for VLSI is that more transistors and more circuit functions may be successfully integrated on a single chip with MOS technology. An individual MOS transistor occupies less chip area and transistor scaling continues to increase the chip density by a factor of two with each new generation. As a result, MOS VLSI circuits are significantly cheaper to manufacture than bipolar circuits of equivalent function and, consequently, MOS VLSI circuits make up a dominant percentage of the total market for digital ICs.

The first MOS circuits, made in metal-gate *p*-channel (PMOS) technology in the early 1970s, required special supply voltages and functioned only at very low digital data rates. The change to *n*-channel (NMOS) silicon-gate technology and other improvements resulted in LSI circuits that required only a single standard supply and operated at much higher data rates. In NMOS technology, all gates on the chip were constructed from two types of *n*-channel transistors. One type of device had a turn-on voltage above 0 V. This device was referred to as an *enhancement-mode* device. The other type of device had a turn-on voltage that was less than 0 V, which implied that it was always *on*. This was the *depletion-mode* device that played a supporting role on the chip since it could be used as a load resistor for the logic gates. Many chips were designed in NMOS technology and, in fact, this technology dominated the decade of the 1970s.

Since the early 1980s, complementary MOS (CMOS) has been the most prevalent MOS technology. This technology provides both *n*- and *p*-channel devices in one chip at the expense of some increase in fabrication complexity and chip area compared to basic NMOS. The great advantage of CMOS digital circuits is that they may be designed with low static power consumption in the steady-state condition. Power is consumed primarily when circuits switch between the two logic states; the average CMOS power consumption is much smaller than for NMOS circuits. In fact, the power consumption problems of NMOS led to a wholesale conversion of the industry from NMOS to CMOS. Today, CMOS is widely used in almost every type of microelectronic application including personal computers, personal digital assistants, cell phones, Internet applications, and a variety of other communication equipment.

The analysis and design of integrated circuits depends heavily on the use of suitable mathematical models for the devices. This is true in hand calculations, where fairly simple models are generally used, and in computer analysis, where much more complex models are utilized for high accuracy. The importance of "back-of-the-envelope" or hand calculations cannot be overemphasized. Designers must be able to gain quick insight into a new circuit configuration in terms of speed, power, and area, as well as many other important characteristics. If a simple model is not available, then design becomes an iterative process with a CAD tool.

Since any analysis or design is only as accurate as the models, it is essential that the circuit designer have a thorough understanding of the models commonly used

and the degree of approximation involved in each. If we know the detailed models, and the approximations used to derive the simple models, we can safely use the simplified models and gain a tremendous insight into the behavior of complex devices. Keep this in mind as you work your way through the material in this chapter.

2.2 Structure and Operation of the MOS Transistor

A simplified view of an *n*-channel polysilicon-gate MOS transistor is shown in Figure 2.1. In simple terms, there are two operating modes for this transistor: *on* and *off*. This is true of all transistors. We must use the terminals provided on the device to place it in these two possible conditions. The four terminals are the *gate, drain, source,* and *bulk* (or *body,* or *substrate*). The schematic symbol for the NMOS device is shown with these four terminals labeled as *G*, *D*, *S*, and *B*, respectively. In the *on* condition, electron current flows from source to drain. In the *off* condition, no current flows in the device. To turn it on, a voltage is applied to the gate node to set up an electric field that creates a conductive channel between heavily doped *n*-type (n^+) source and drain regions,[1] and current flows when a potential difference exists between these two nodes.

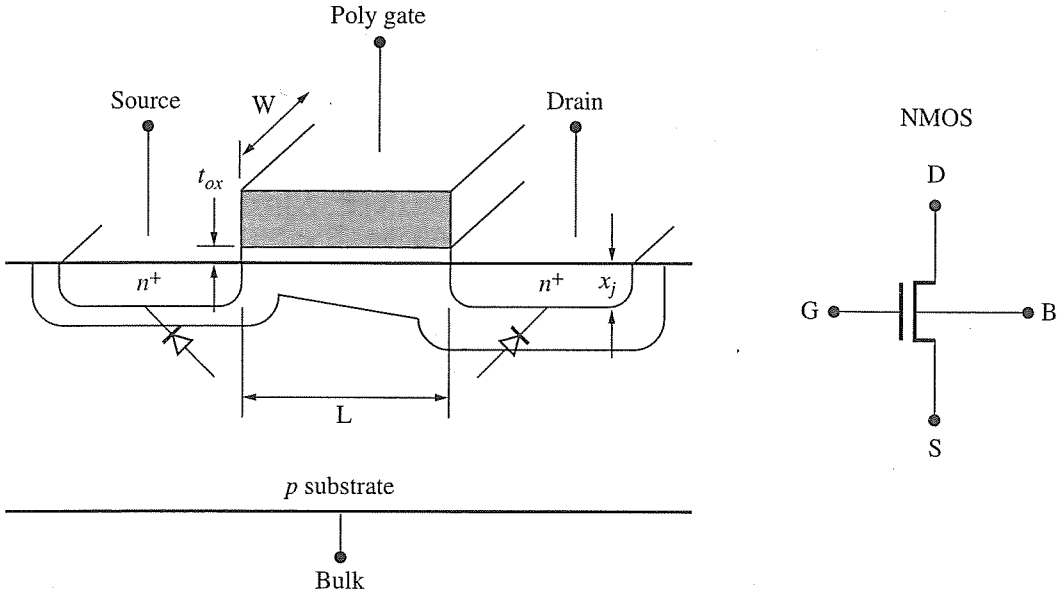


Figure 2.1
NMOS transistor structure and symbol.

[1] The notation n^+ implies doping $>10^{18}/cm^3$, n^- implies doping $<10^{15}/cm^3$, and n is not specific. The same convention applies for designations of *p*-type material.

Since it is based on the use of an electric field, the device is one form of field-effect transistor (FET). Note that the gate is completely insulated from the other electrodes. This fact leads to the designation insulated-gate field-effect transistor (IGFET). Another name, although much older, is unipolar transistor. This name arises from the fact that only a single type of charge carrier (electrons in NMOS) is necessary for device operation. The mobile holes in the *p*-type substrate of an NMOS transistor are not involved in normal transistor operation. A variety of terms have been used to reference this device including FET, IGFET, MOST, and MOSFET. We will use the more recent terms, *MOS transistor* or *MOS device, NMOS device,* and *n-channel* transistor, when referring to the device in Figure 2.1.

As mentioned in Chapter 1, the acronym *MOS* is derived from the metal-oxide-semiconductor structure that forms the gate of the device. In the early days of PMOS technology, a metal gate was used but it was difficult to align the metal over the channel precisely. An offset in one direction or the other would create a non-functioning transistor (either a short or an open). To overcome these alignment problems, a polycrystalline silicon material was introduced to serve as the gate. This so-called *poly* gate would be deposited before the source and drain material. When the source and drain diffusions were created, they would align with the gate, rather than the other way around. The yield of such devices went up significantly compared to metal-gate technology, and today, polysilicon is used almost exclusively as the gate material. It is heavily doped to keep its resistance low since it is supposed to behave like a metal material.

The device structure shown in Figure 2.1 is formed by a complex sequence of steps including oxidation, pattern definition, diffusion, ion implantation, and material deposition and removal processes. In the fabrication process, horizontal device dimensions are made as small as possible with the available technology in order to maximize both circuit density and high-speed performance. A corresponding reduction in the vertical dimensions is also necessary to maintain field levels. A description of the processing sequence is presented in Chapter 3.

The final structure has a number of features that deserve mention. The most important horizontal dimension is channel length L, shown in the figure. Typical values of L today are in the range of 350 nm to 90 nm. This dimension will continue to scale according to Moore's Law in the years ahead. Perpendicular to the plane of the figure is the channel width W, typically much larger than the minimum length, depending on the desired current-handling capability. Gate oxide thickness t_{ox}, the most important vertical dimension, is typically less than 5 nm (50 Å). Today it is below 25 Å. As we shall see, gate length and width and gate oxide thickness are major parameters in determining the electrical characteristics of the MOS transistor. The other vertical parameter shown is the junction depth, x_j, which is of the order of 70 nm to 150 nm today.

The *body*, or *substrate,* is a single-crystal silicon wafer that is the starting material for circuit fabrication and provides physical support for the final circuit. The *p*-type substrate doping density is a factor in device electrical behavior, as described in the next section. The silicon surface is comprised of *active* and *field* regions. The active region contains the transistor, while the field region serves to isolate transis-

tors. The main requirement on the field regions is that they should never permit conduction between separate active regions. In NMOS, all conduction is via electrons, so the field region fulfills its purpose if electrons can never pass through it. In addition, a thick layer of silicon dioxide over the field regions is used to minimize unwanted capacitance from interconnecting metal to the body.

The transistor regions in the body comprise n^+ source and n^+ drain regions separated by p-type material in the channel region. Typical devices are symmetrical, just as the one shown in Figure 2.1; source and drain are interchangeable. *Strictly speaking, the source and drain can be only identified after the polarities of applied voltages are established.* In NMOS, the more positive node is defined as the drain. With no external voltages applied, the path from drain to source has two pn^+ junction diodes in series, back to back, with the body as a p region common to them. These junctions should never be forward-biased. In fact, the subtrate should be connected to the lowest possible potential in the circuit, typically a ground potential, called V_{SS} or Gnd. The only current that is permitted to flow across the junctions is diode reverse leakage current.

Now consider the result when source, drain, and body are all tied to ground and a positive voltage is applied to the gate. From simple concepts of electrostatics, a positive gate voltage will tend to draw electrons from the substrate into the channel region. The n^+ source and drain regions also provide nearby copious sources of electrons. Once electrons are present in the channel region, a conducting path is present between drain and source. Current will flow from drain to source if there is a voltage difference between them. That is, electrons will flow from source to drain in this condition.

To first-order, the conducting channel does not form for very small positive gate voltages. The electrostatic potential at the surface of the p-type material in the channel region must be made positive by application of a larger gate-source voltage. The gate voltage needed to initiate formation of a conducting channel is termed the *threshold voltage* V_T. This important device parameter is analyzed in the next section.

Now consider the structure of the device in Figure 2.2, called the p-channel or PMOS device. It has complementary characteristics to the NMOS device. While the structure is the same as the NMOS device, the doping is opposite. The heavily doped regions in the body comprise p^+ source and p^+ drain regions separated by lightly doped n-type material in the channel region. Again, the source and drain are interchangeable and can be identified only after the polarities of applied voltages are established. In PMOS, *the more positive node is defined as the source.* This is represented in the schematic of Figure 2.2 by reversing the positions of the source and drain.

The *substrate* is an n-type material that must be created in advance of placing the source and drain regions. It is often referred to as an *n-well* or *n-tub,* especially if it is created in a silicon wafer that is initially doped with a p-type background material. In this case, it would be called an n-well process. If both n-channel and p-channel devices are created in their own wells, it is referred to as a *double-well* or *twin-tub* process. Again, the path from drain to source has two p^+n junction diodes in series, back to back, with the body as an n region common to them. The

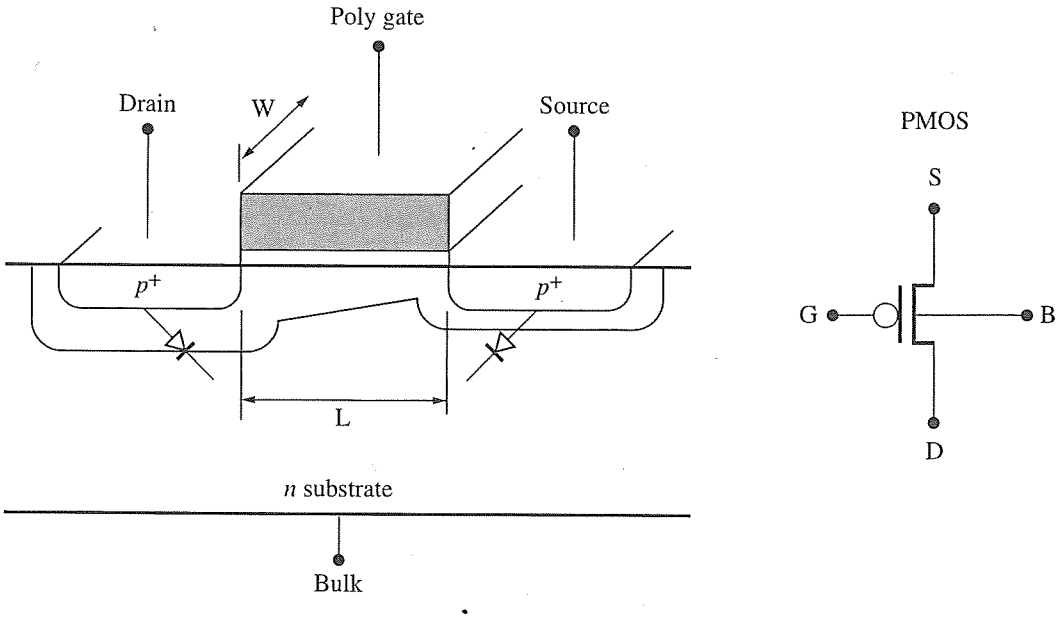


Figure 2.2
PMOS transistor structure and symbol.

only current that can flow is diode reverse leakage. They should never be forward-biased. Therefore, the *n*-well must be tied to the highest voltage in the system, called V_{DD}. In a twin-tub process, the *p*-well must be tied to Gnd or V_{SS} for the same reason. These connections of the wells to V_{DD} or Gnd are sometimes referred to as *tubties* or *well-plugs*.

Consider the *p*-channel device when source, drain, and body are all tied to V_{DD}. If a low voltage is applied to the gate, it will tend to draw holes into the channel region. Once holes are present in the channel region, a conducting path is present between drain and source. Current will flow from source to drain if there is a voltage difference between them. That is, holes will flow from the source to the drain if the drain has a lower potential than the source. By definition, the source is the terminal with the higher potential in PMOS devices. In effect, the PMOS device has a negative threshold since the gate must be at a lower voltage than the source to invert the channel.

The fact that the *p*-channel device has a negative threshold, and is therefore opposite to the NMOS device, is captured in the symbol for the device by placing a circle on the gate input. Many books use a slightly different symbol for this device, but we will use this one consistently throughout the book.

NMOS and PMOS transistors that have no conducting channel at zero gate-source voltage are termed *enhancement-mode* devices, meaning that gate-source voltage of the same polarity as drain-source voltage (positive for NMOS, negative for PMOS) is required to initiate conduction. NMOS *depletion-mode* devices have

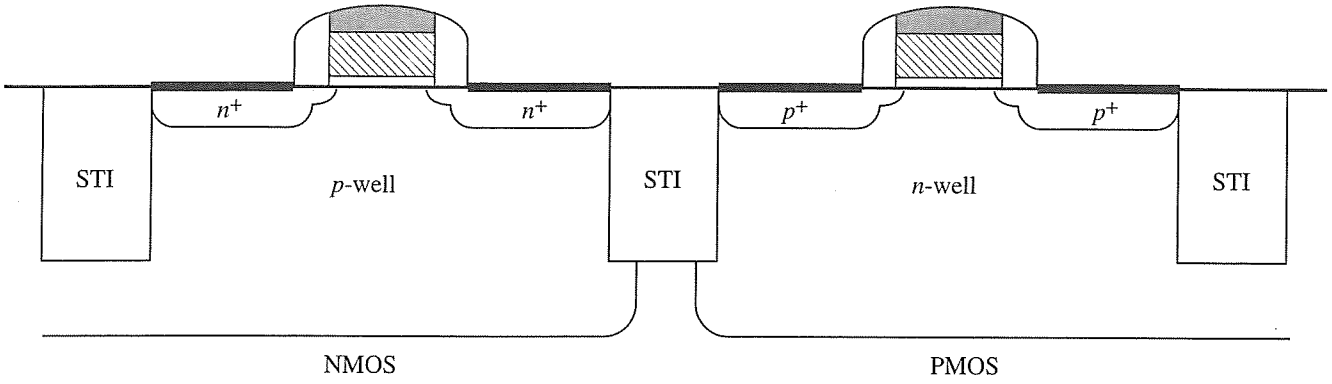


Figure 2.3
CMOS transistors structure.

negative thresholds, but in CMOS designs, we rarely (if ever) use depletion-mode devices today.

Figure 2.3 shows a cross-sectional view of a modern silicon-gate CMOS structure. While it is not drawn to scale, it represents the typical profile after fabrication of a twin-tub CMOS process, the details of which are left to Chapter 3. We now know that the PMOS devices are symmetrical but opposite to NMOS devices in terms of doping and voltage polarities. In CMOS circuits, all devices are assumed to be enhancement-mode devices. They are separated by oxide regions using a method called shallow-trench isolation (STI). The NMOS device has a positive threshold voltage while the PMOS device has a negative threshold voltage. The equations to be derived in the sections to follow are based on the NMOS device, but apply equally to the PMOS device after the appropriate sign changes.

2.3 Threshold Voltage of the MOS Transistor

In order to fully understand the mechanisms behind the threshold voltage derivation, we must examine the energy-band diagram of a material as shown in Figure 2.4. This diagram is a representation of the allowable energy states for electrons. Two distinct levels are defined: the conduction band, E_C, and the valence band, E_V. In order to have electrical conduction, an electron must move from the valence band to the conduction band and, in doing so, must surmount the *bandgap* associated with the material. In *metals,* the two energy levels overlap making the bandgap effectively zero; electrons can freely move from one band to another. In *insulators,* the energy levels are very far apart and the barrier is insurmountable. In *semiconductors,* the gap is relatively small, 1.1 eV for silicon. Since there are no allowed states in the bandgap region, electrons must have enough energy to jump from the valence band to the conduction band. However, since the gap is relatively small, energetic

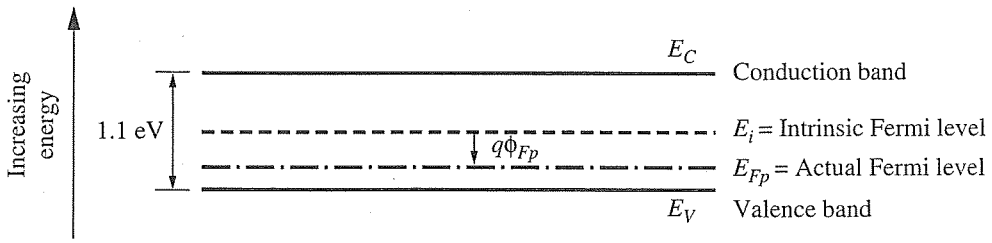


Figure 2.4
Energy-band diagram for doped *p*-type silicon.

electrons are able to surmount this barrier from time to time due to thermal excitation.

In an undoped semiconductor, the total number of electrons that are able to surmount this barrier is termed the *intrinsic carrier concentration*, n_i. This depends on the number of available states in the conduction band and the probability of occupancy of these states, which depends on the temperature. For example, n_i at room temperature is

$$n_i = 1.45 \times 10^{10}/\text{cm}^3 \tag{2.1}$$

The relationship between the intrinsic carrier concentration and the holes and electrons at equilibrium is given by the *mass action law:*

$$np = n_i^2 \tag{2.2}$$

where n is the mobile electron concentration and p is the mobile hole concentration. This equation tells us that the mobile electrons and mobile holes must have equal concentration in undoped silicon at equilibrium. To represent this on the energy band diagram, we define the *intrinsic Fermi level* as a line where the probability of electron occupancy of an empty state is 50%. This is denoted E_i in Figure 2.4 and is located roughly midway between the conduction band and the valence band.

More generally, the probability of electron occupancy of an allowable state is based on the doping level in the semiconductor. As we increase the doping level, N_A, of silicon with *p*-type acceptor impurities such as boron, the probability of electrons in the conduction band is reduced and the actual Fermi level (50% probability point) shifts below the intrinsic Fermi level. This is because we are adding holes to the semiconductor and reducing the likelihood that an electron will surmount the barrier. The mass action law can be used to support this fact. If $p = N_A >> n_i$, then the mobile electron concentration is reduced to

$$n = \frac{n_i^2}{N_A}$$

which is much smaller than n_i. Since there are fewer mobile electrons, the probability that the empty states in the conduction band will be occupied decreases.

The amount of shift in the Fermi level due to the doping is called the electrostatic potential. Conventionally, the equilibrium electrostatic potential ϕ_F in a semiconductor is defined as the difference between the intrinsic Fermi level and the actual Fermi level in the p-type or n-type doped semiconductor:[2]

$$\phi_{Fp} = \frac{kT}{q} \ln \frac{n_i}{p} \tag{2.3a}$$

or

$$\phi_{Fn} = \frac{kT}{q} \ln \frac{n}{n_i} \tag{2.3b}$$

where the equilibrium majority mobile carrier concentration is assumed to be equal to the doping concentration $p = N_A$ for p-type material, or $n = N_D$ for n-type material.

The kT/q term is called the thermal voltage, V_{th}, which is approximately 26 mV at room temperature. (Do not confuse this with the threshold voltage, V_T, which is a completely different quantity!) By convention, ϕ_F is negative for p-type semiconductor material and positive for n-type material. For example, in a p-type material, we would have:

$$\phi_{Fp} = \frac{kT}{q} \ln \frac{n_i}{N_A} \tag{2.4}$$

Since N_A is the doping level in the p-substrate, and the value of $N_A >> n_i$, the electrostatic potential is negative.

Now that the energy band diagram has been reviewed, its application to the threshold voltage derivation can be described. Recall that the MOS structure is comprised of a metal material (polysilicon), a gate oxide, and the silicon body. The gate and body of the MOS transistor form the plates of a capacitor with silicon dioxide, SiO_2, as a dielectric. The gate-oxide capacitance per unit area is defined as:

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \tag{2.5}$$

where ε_{ox} is the permittivity of oxide and t_{ox} is the oxide thickness.

The threshold voltage for a long-channel enhancement-mode NMOS device can be derived using the transistor cross-section shown in Figure 2.5. Consider the body, source, and drain connected to ground and a voltage V_{GS} (initially zero) applied to the gate. As V_{GS} changes from zero to a positive value, positive charge accumulates on top of the gate and negative charge accumulates as electrons in the substrate under the gate. This actually happens gradually as the gate voltage moves

[2] p-type silicon is obtained by adding acceptor impurities such as boron. n-type silicon is obtained by adding donor impurities such as phosphorus or arsenic.

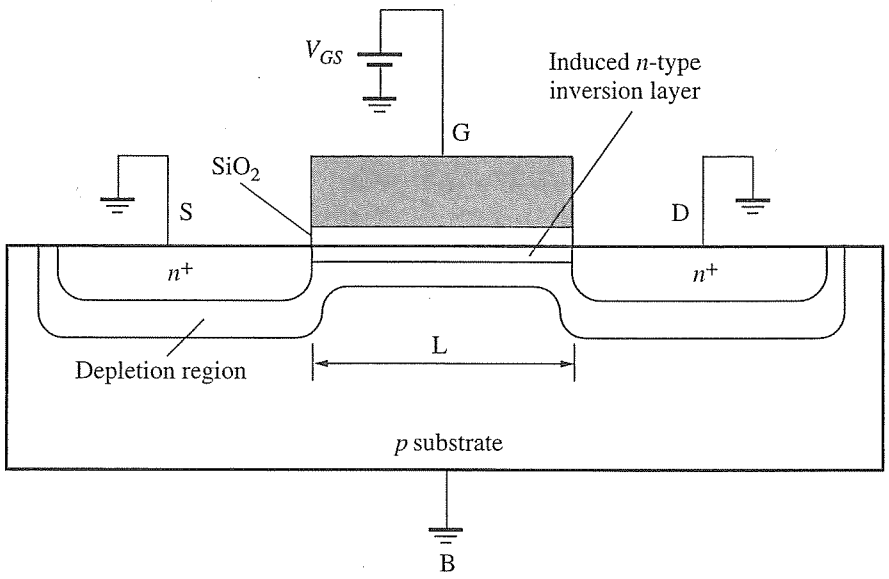


Figure 2.5
Idealized NMOS device cross section with positive voltage applied showing depletion regions and the induced channel.

from 0 V to the threshold voltage. Initially, the negative charge in the p-type body is manifested by creation of a *depletion region*[3] in which mobile holes are pushed down under the gate, leaving behind negatively charged immobile acceptor ions.

As the gate voltage continues to increase, the depletion-layer thickness increases and eventually an initial layer of mobile electrons appears at the surface of the silicon in the so-called *weak inversion* condition. Further increases in the gate voltage increases the concentration of the mobile carriers in the channel until the concentration of electrons at the surface equals the concentration of holes in the substrate, a condition known as *strong inversion.*[4] For gate voltages above this point, the depletion-layer thickness remains approximately constant while the additional charge on the gate is matched by additional mobile carriers in the channel region that are drawn from the source and drain regions—where an abundant supply of such carriers exists.

The value of gate voltage V_{GS} required to produce strong inversion is called the *threshold voltage,* V_T. We will present the main components of this voltage and then go back and describe the origin of each component. There are three main terms:

1. A voltage term, ϕ_{GC}, represents the difference in work functions between the gate (G) material and the silicon substrate on the channel (C) side. The work function for a material is the amount of energy needed to move an electron

[3] This is the same type of depletion region that appears in a *pn* junction.
[4] This is a somewhat arbitrary but important definition of strong inversion. The surface is inverted relative to the substrate since the carrier concentrations are equal and opposite.

from the Fermi level to free space level E_0, as shown in Figure 2.6. When two materials with different Fermi levels are brought together into one system, the Fermi levels are aligned at equilibrium. The work function difference tells us how "misaligned" they are to begin with. When equilibrium is reached, the conduction and valence bands will bend to accommodate the work function difference. For silicon-gate devices

$$\phi_{GC} = \Phi_{G\ (polysilicon\ gate)} - \Phi_{C\ (substrate)}$$

2. There is always an undesired positive charge Q_{ox} present in the oxide and at the interface between the oxide and the bulk silicon. This charge is due to impurities and/or imperfections such as sodium ions in the oxide, Q_{Na^+}, and dangling bonds at the interface, Q_{SS}. We combine these two terms into one term called Q_{ox} and place the equivalent charge at the oxide-silicon interface as in Figure 2.6. Since there is positive charge effectively on the bottom plate of the MOS capacitor, the top plate must provide negative charge to compensate for it. Therefore, it contributes a negative quantity to the threshold voltage of $-Q_{ox}/C_{ox}$.
3. A gate voltage $(-2\phi_F - Q_B/C_{ox})$ is needed to change the surface potential to the strong inversion condition and to offset the induced depletion-layer charge Q_B. This is shown in Figure 2.7.

Now that the terms are defined, we explain the inversion process in detail. Consider what happens when the MOS system is first brought together. At equilibrium, there is a certain amount of band-bending that occurs to make the Fermi levels constant throughout the entire system. We need to know exactly how much band-bending has taken place in order to provide a meaningful reference point for the strong inversion condition. Therefore, the bands are first flattened by a *flat-band* voltage, V_{FB}. The name comes from the fact that application of this voltage at the gate produces flat

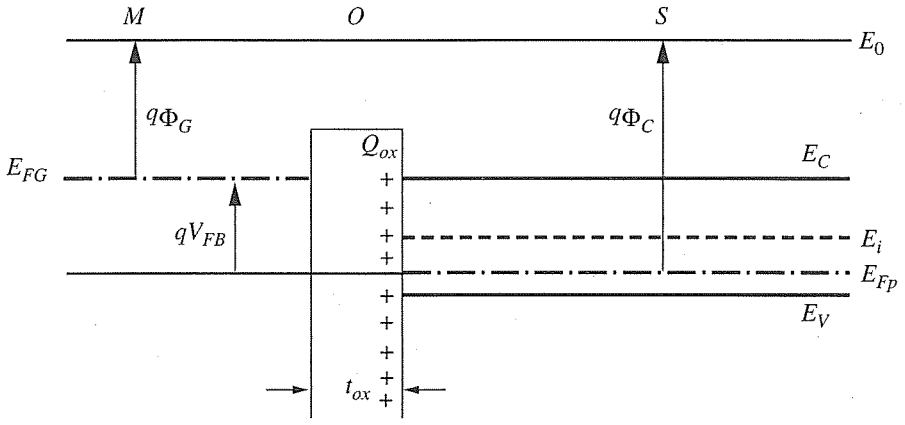


Figure 2.6
Flat-band condition in the MOS system.

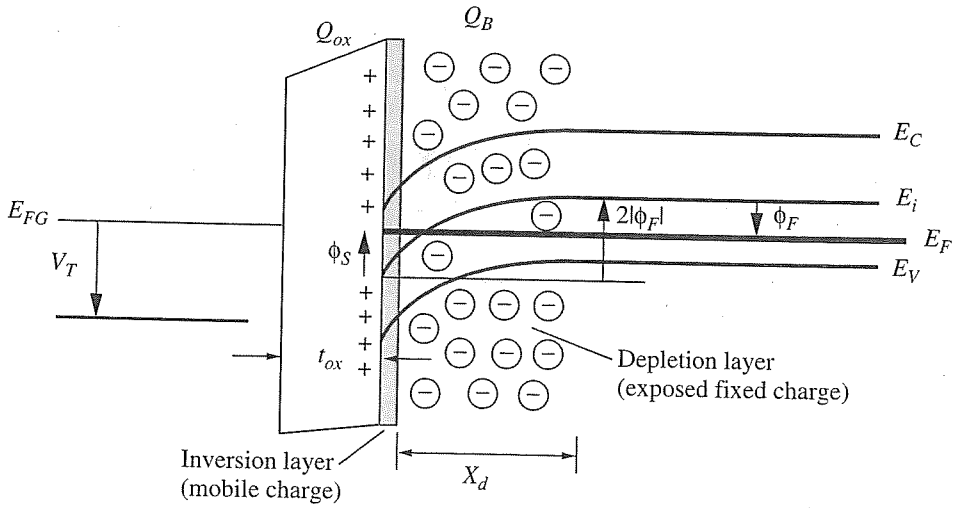


Figure 2.7
Energy bands in the strong inversion condition.

energy bands in the metal-oxide-semiconductor system, as shown in Figure 2.6. This figure shows the MOS energy band diagrams depicting the conduction band, E_C, valence band, E_V, the intrinsic Fermi level, E_i, and the actual Fermi level in the p-type semiconductor, E_{Fp}. The work functions associated with the gate, $q\Phi_G$, and the semiconductor, $q\Phi_C$, are also shown here. The Fermi level in the gate material, E_{FG}, depends on the doping level in the polysilicon gate material. If the gate is heavily doped with n^+ material, the position of E_{FG} is near E_C. However, if it is heavily doped with p^+ material, E_{FG} is near E_V. For an NMOS device, usually an n^+ doping is used in the poly gate whereas a p^+ doping is used for PMOS devices. This polarity of doping is consistent with the doping of the drain and source regions, so they can be doped at the same time.

The amount of band "unbending" that is needed to reach this flat-band condition is given by

$$V_{FB} = \phi_{GC} - \frac{Q_{ox}}{C_{ox}} \tag{2.6}$$

That is, if we apply the difference between the work functions, and we compensate for the oxide charges, we will achieve the flat-band condition shown in Figure 2.6. We now have a known reference point from which to operate.

Next, we need to bend the bands in the channel region by $2|\phi_F|$ relative to the substrate,[5] as shown in Figure 2.7. That is, the surface potential ϕ_S must be made equal to ϕ_F. This will achieve the desired strong inversion condition. Then, we need to compensate for the exposed depletion-layer charge, Q_B, due to the band-bending itself. The fact that we bent the bands means that the fixed charge is exposed in the depletion layer of thickness, X_d.

[5] We drop the electronic charge q for convenience in Figure 2.7.

A simple analysis can be applied to find the depletion-layer charge by first determining the thickness X_d of the depletion region between the channel and the substrate as a function of the electrostatic potential ϕ_S at the silicon surface. To create the shaded depletion region shown in Figure 2.7, mobile holes must be pushed back into the substrate. The thickness of the depletion layer in the p-type material is given by

$$X_d = \left(\frac{2\varepsilon_{si}|\phi_S - \phi_F|}{qN_A} \right)^{1/2} \tag{2.7}$$

The doping density in the p-type substrate is denoted as N_A, and ε_{si} is the permittivity of silicon. The immobile charge per unit area due to acceptor ions that have been stripped of their mobile holes is given by

$$Q_B = qN_A X_d = -\sqrt{2qN_A\varepsilon_{si}|\phi_S - \phi_F|} \quad \text{for } |\phi_S - \phi_F| \geq 0 \tag{2.8}$$

In order to invert the surface of the p-type semiconductor, the voltage V_{GS} is used to attract mobile electrons to the channel region. As V_{GS} is increased in the positive direction,[6] the potential ϕ_S at the silicon surface increases from its original (negative) equilibrium value of ϕ_{Fp}, through zero, until $\phi_S = -\phi_{Fp}$. Under this condition the density of mobile electrons at the surface is equal to the density of mobile holes in the original substrate or body. The surface potential has changed by $-2\phi_{Fp}$ relative to the substrate.

The value of V_{GS} needed to cause this $-2\phi_{Fp}$ change in surface potential is defined as the *threshold voltage* V_T for a MOS transistor. This condition is known as *strong inversion*. In essence, the semiconductor surface that is normally p-type becomes n-type. Further increases in gate voltage produce only slight changes in ϕ_S and the depletion-layer thickness. However, it increases the electric field across the gate oxide and increases the electron concentration in the channel. The additional electrons are drawn into the inversion layer from the strongly n-type source for drain, with the positive surface potential as the attractive force.

Band-bending at Strong Inversion **Example 2.1**

Problem:

A 130 nm technology employs carrier concentrations in the p-well in the range of 3×10^{17} cm^{-3}. Estimate the degree of band-bending required for strong inversion at room temperature, relative to the flat-band condition.

[6] The positive direction for the gate voltage is actually pointing downward in the energy band diagram of Figure 2.7. It bends the bands downwards. The flat-band voltage is negative since it will "unbend" the bands from its equilibrium condition.

Solution[7]:

$$2|\phi_{Fp}| = \frac{2kT}{q}\left|\ln\frac{n_i}{p}\right| = 2(0.026)\left|\ln\frac{1.45\times10^{10}}{3\times10^{17}}\right| \cong 0.88\ \text{V}$$

As long as there is only a small voltage difference between drain and source, the induced layer of electrons extends continuously from source to drain, producing a continuous region with mobile electrons. The conductivity of the channel thus formed can be increased or decreased (*modulated*) by increasing or decreasing the gate voltage. In the presence of an inversion layer, and with no body bias ($V_{SB} = 0$), the depletion region contains a fixed negative charge that may be found by using (2.8) together with the fact that $\phi_S = -\phi_F$.

$$Q_{B0} = -\sqrt{2qN_A\varepsilon_{si}\,|-2\phi_F|} \tag{2.9a}$$

If there is a voltage V_{SB} between source and body (V_{SB} is normally positive for n-channel and negative for p-channel devices), a slight modification is required. The surface potential required to produce inversion becomes $|-2\phi_F + V_{SB}|$ and the charge stored in the depletion region in this case is

$$Q_B = -\sqrt{2qN_A\varepsilon_{si}\,|-2\phi_F + V_{SB}|} \tag{2.9b}$$

Example 2.2 **Depletion Layer Fixed Charge Calculation**

Problem:

A p-type well in a 130 nm technology has $N_A = 3\times10^{17}$ cm^{-3}. Find the limiting value of depletion-layer width and the total charge contained in the depleted region.

Solution:

From Example 2.1, we already know that

$$2|\phi_{Fp}| = \frac{2kT}{q}\left|\ln\frac{n_i}{p}\right| = 0.88\ V$$
$$= |\phi_s - \phi_F|$$

From Equation (2.7),

$$X_d = \sqrt{\frac{2(11.7)8.85\times10^{-14}(0.88)}{1.6(10^{-19})\times3(10^{17})}} = 6\times10^{-6}\text{cm} = 60\,\text{nm}$$

[7] Note that ϕ_{Fp} is negative for n-channel devices and ϕ_{Fn} is positive for p-channel devices. This can be confusing so we usually take the absolute value of this quantity.

From Equation (2.9a),

$$Q_{B0} = -(2 \times 1.6 \times 10^{-19} \times (3)10^{17} \times 1.0 \times 10^{-12} \times |-0.88|)^{1/2}$$
$$\approx -3 \times 10^{-7}\,\text{C/cm}^2$$

The complete expression for the threshold voltage V_T is given by

$$V_T = V_{FB} - 2\phi_F - \frac{Q_B}{C_{ox}} \tag{2.10}$$

Expanding V_{FB} and rearranging, we obtain

$$\begin{aligned} V_T &= \phi_{GC} - 2\phi_F - \frac{Q_B}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} \\ &= \phi_{GC} - 2\phi_F - \frac{Q_{B0}}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_B - Q_{B0}}{C_{ox}} \\ &= V_{T0} + \gamma\left(\sqrt{V_{SB} + |2\phi_F|} - \sqrt{|2\phi_F|}\right) \end{aligned} \tag{2.11}$$

where Equation (2.9a) has been used to simplify the expression and V_{T0} is the threshold voltage with $V_{SB} = 0$ called the *zero-bias threshold voltage.* The parameter γ (gamma) is termed the *body-effect coefficient* or *body factor.* Comparing Equation (2.9) to Equation (2.11), we see that γ is given by

$$\gamma = \frac{1}{C_{ox}}\sqrt{2q\varepsilon_{si}N_A} \tag{2.12}$$

Computation of C_{ox} and γ **Example 2.3**

Problem:

Determine values of C_{ox} and γ, if $t_{ox} = 22$ Å and $N_A = 3 \times 10^{17}\,\text{cm}^{-3}$.

Solution:

To compute C_{ox} we use Equation (2.5):

$$\varepsilon_{ox} = 4\varepsilon_o$$

$$\therefore C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} = \frac{(4)(8.85 \times 10^{-14})}{22\ \text{Å}} = 1.6 \times 10^{-6}\,\text{F/cm}^2$$

Apply this value to the equation for γ:

$$\gamma = \sqrt{2q\varepsilon_{si}N_A}/C_{ox}$$

$$\gamma = \sqrt{(2)(1.6 \times 10^{-19})(11.7)(8.85 \times 10^{-14})(3 \times 10^{17})}/1.6 \times 10^{-6}$$

$$\approx 0.2\ \text{V}^{1/2}$$

Table 2.1
Signs in threshold voltage equation

Parameter	NMOS	PMOS
Substrate	*p*-type	*n*-type
V_{T0}	+	−
ϕ_{GC}:		
n^+ polysilicon gate	−	−
p^+ polysilicon gate	+	+
Φ_F	−	+
Q_{B0}, Q_B	−	+
Q_{ox}	+	+
γ	+	−
X_d, C_{ox}	+	+
V_{SB}	+	−

It is easy to become confused about the signs of the various terms in the threshold voltage equations (if not the terms themselves!). Equations (2.10–2.12) give correct results for NMOS and PMOS if the signs shown in Table 2.1 are used.

Example 2.4 **Threshold Voltage Calculation**

Problem:

Calculate the zero-bias threshold voltage (i.e., $V_{SB} = 0$) for an NMOS silicon-gate transistor that has well doping $N_A = 3 \times 10^{17}$ cm^{-3}, gate doping $N_D = 10^{20}$ cm^{-3}, gate-oxide thickness $t_{OX} = 22$ Å, and 2×10^{10} cm^{-2} singly charged positive ions per unit area at the oxide-silicon interface. Assume that the gate doping is n^+ and explain why it is appropriate.

Solution:

$$V_{T0} = \phi_{GC} - 2\phi_{FP} - \frac{Q_{B0}}{C_{ox}} - \frac{Q_{ox}}{C_{ox}}$$

$$\phi_{Fp} = \frac{kT}{q} \ln \frac{n_i}{N_A} = 0.026 \ln \frac{1.4 \times 10^{10}}{3 \times 10^{17}} = -0.44 \text{ V}$$

$$\phi_{GC} = \phi_{Fp} - \phi_{G(gate)} = -0.44 - 0.55 = -0.99 \text{ V}$$

$$\varepsilon_{ox} = 4\varepsilon_0 = 3.5 \times 10^{-13} \text{ F/cm}$$

$$C_{ox} = 1.6 \times 10^{-6} \text{ F/cm}^2$$

$$Q_{B0} = 3 \times 10^{-7} \text{ C/cm}^2$$

$$\frac{Q_{B0}}{C_{ox}} = \frac{3 \times 10^{-7}}{1.6 \times 10^{-6}} = 0.188 \text{ V}$$

$$\frac{Q_{ox}}{C_{ox}} = \frac{2 \times 10^{10} \times 1.6 \times 10^{-19}}{1.6 \times 10^{-6}} = 0.002\ \text{V}$$

$$V_{T0} = -0.99 - (-0.88) - (-0.188) - 0.002 = +0.08\ \text{V}$$

In the above solution, we assumed that the n^+ gate doping is so high that the Fermi level in the gate is coincident with the conduction band, which implies an electrostatic potential of $\phi_{FG(\text{gate})} = 0.55$. If the gate doping were p^+, the value for V_{T0} would be 1.18 V which is well above the desired levels. Therefore, the doping of the poly gate must be n^+ to keep it below the target value. For the same reason, the poly gate for p-channel devices is doped with p^+ material. Of course, the value computed above is not very satisfactory for an NMOS device. We require a value closer to $V_{T0} = 0.4$ V.

The small positive value of threshold voltage calculated above is not desirable for use in digital circuits. Although in principal the threshold voltage may be set to any value by proper choice of doping concentrations and oxide capacitance, considerations such as breakdown voltage and junction capacitance frequently dictate the desirable specifications for these variables. Therefore, the final value of V_{T0} is determined during circuit fabrication by ion implanting dopant atoms into the channel region.

A p-type threshold implant, using boron for example, will make the threshold voltage more positive. On the other hand, an n-type threshold implant, using phosphorus for example, makes the threshold voltage more negative. NMOS threshold voltages are adjusted using a p-type implant until the desired positive value is reached. The PMOS thresholds are all adjusted with n-type implants until they are at the desired negative value.

Extra p-type impurities are implanted to make $V_{T0N} = +0.4$ V for n-channel enhancement devices in our generic 130nm process. By implanting n-type impurities in the channel region of a p-channel device, we try to achieve a $V_{T0P} = -0.4$ V. If Q_I is the charge density per unit area in the channel region due to the implant, then the threshold voltage, V_{T0}, given by (2.10) and (2.11) is shifted by Q_I/C_{ox}. It is assumed that all implanted ions are electrically active.

Threshold Voltage Implant Dosage Calculation **Example 2.5**

Problem:

Continuing on with the previous example, since the NMOS threshold voltage is not close to the desired value, calculate the ion-implant doses N_I needed to achieve a threshold voltage of 0.4 V in unit of ions/cm^2.

Solution:

We know that V_T will be shifted by $Q_I/C_{ox} = qN_I / C_{ox}$.

$$\therefore N_I = \frac{Q_I}{q} = \frac{C_{ox}\Delta V}{q} = \frac{C_{ox}(0.4 - (+0.08))}{1.6 \times 10^{-19}} = 3.2 \times 10^{12} \frac{\text{ions}}{\text{cm}^2}$$

The above calculations of threshold voltage do not give exact quantitative results in practical cases. Reasons for this include the fact that body doping may vary near the oxide interface, the oxide thickness and dielectric constant may vary due to process variations, and oxide charge is not exactly controlled. Considerations such as hot-electron effects, junction breakdown voltages, junction capacitances, and punchthrough dictate the desirable specifications of the doping levels in the channel region. For our purposes, calculations of threshold voltage are useful for predicting how V_{T0} varies as a function of doping levels and dimensions. Furthermore, the values given above are for illustrative purposes only. As a practical matter for circuit design, nominal values and statistical variations of the threshold voltage and body-effect coefficient must be determined by direct measurements of actual devices.

2.4 First-Order Current-Voltage Characteristics

The next step is to derive the large-signal characteristics of long-channel MOS transistors for dc or slowly changing applied signals. We assume an NMOS device with source grounded and bias voltages V_{GS}, V_{DS}, and V_{BS} applied as shown in Figure 2.8. If V_{GS} is greater than V_T, a conducting channel is present and V_{DS} causes a drift current I_{DS} to flow from drain to source. The voltage V_{DS} causes a larger reverse bias from drain to body than that present from source to body, and thus a wider depletion layer exists at the drain. However, for simplicity we assume that the voltage drop along the channel is small so that the threshold voltage and depletion-layer width are approximately constant along the channel.

At a distance y along the channel, the voltage with respect to the source is $V(y)$, where $0 \leq V(y) \leq V_{DS}$, and the gate-to-channel voltage at that point is $V_{GS} - V(y)$. In order to have any inversion-layer charge, the gate-to-channel voltage must be higher than V_T. The amount of charge induced by a given voltage is $Q = CV$. We assume that this voltage $V_{GS} - V(y)$ exceeds the threshold voltage V_T, and thus the induced charge per unit area at the point y in the channel is the charge density:

$$Q_n(y) = C_{ox}[V_{GS} - V(y) - V_T] \tag{2.13}$$

In basic terms, current is simply charge in motion. Since Equation (2.13) determines how much charge exists in the channel, the drain-source current I_{DS} is given

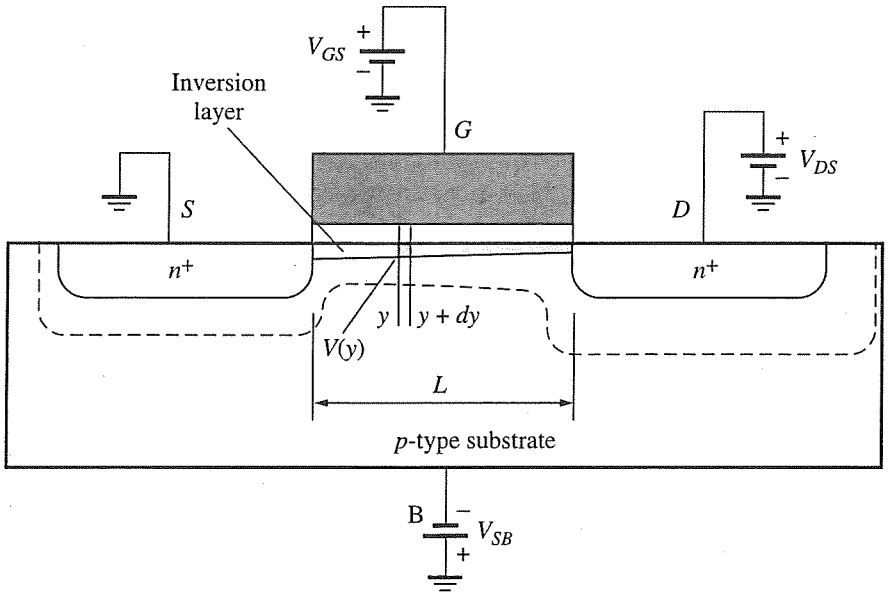


Figure 2.8
NMOS device with bias voltages applied.

by the charge density times the carrier velocity, v, times the channel width, W, perpendicular to the plane of Figure 2.8.

$$I_{DS} = Q_n \times v \times W \tag{2.14}$$

The carrier velocity in cm/s is determined by the horizontal electric field, E, which is in units of V/cm and is directed from drain to source. For the first-order model, we assume that the velocity is linearly proportional to the magnitude of the E field:

$$v = \mu E \quad \text{where } E = \frac{dV(y)}{dy} \tag{2.15}$$

and μ is the carrier mobility which is the ratio of carrier (electron or hole) velocity to electric field. Its dimensional units are cm/s over V/cm, or cm^2/V-s.

To continue with the analysis of MOS transistor conduction, we substitute Equation (2.15) into (2.14):

$$I_{DS} = C_{ox}\left[(V_{GS} - V(y)) - V_T\right] \times \mu_n E \times W$$

Further substitution and minor rearrangement produces:

$$I_{DS}\, dy = W\mu_n C_{ox}\,(V_{GS} - V(y) - V_T)\, dV$$

For simplicity, a *process transconductance parameter* k' is defined as:

$$k' = \mu_n C_{ox} = \frac{\mu_n \varepsilon_{ox}}{t_{ox}} \tag{2.16}$$

We assume that V_T does not vary significantly along the length of the channel. Integrating the left side along the channel from $y = 0$ to L and the right side from $V = 0$ to V_{DS} and substituting:

$$I_{DS}\int_0^L dy = Wk'\int_0^{V_{DS}} (V_{GS} - V - V_T)\, dV$$

$$I_{DS} = k'\frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right] \tag{2.17a}$$

The *device transconductance parameter* is defined as $k = k'(W/L)$. Substituting this in the above yields

$$I_{DS} = \frac{k}{2}\left[2(V_{GS} - V_T)V_{DS} - V_{DS}^2\right] \tag{2.17b}$$

for the so-called *linear* region of operation. This equation is an important one since it describes the current-voltage (I-V) characteristics of the long channel MOS transistor, assuming a continuous channel is present from source to drain.

As the value of V_{DS} is increased, the induced conducting channel charge Q_n decreases near the drain. Equation (2.13) indicates that Q_n at the drain end approaches zero as V_{DS} approaches $V_{GS} - V_T$. When V_{DS} equals or exceeds $V_{GS} - V_T$, the channel is said to be *pinched off*. Increases in V_{DS} above this critical voltage produce little change in I_{DS}, and Equation (2.17a) no longer applies. In fact, if Equation (2.17b) is plotted as I_{DS} versus V_{DS} with different values of V_{GS}, the curves would reach a peak and then "roll over" as V_{DS} is increased since it is a quadratic equation. This characteristic, shown in Figure 2.9a, is nonphysical and inaccurate. It is due to Equation (2.13) reversing its polarity when $V(y)$ is greater than $V_{GS} - V_T$. This produces positive charge in the channel after the pinch-off point, as shown in Figure 2.9b, something that is quite impossible!

Since the current equation is only valid up to the pinch-off point, we will call this the saturation voltage:

$$V_{Dsat} = V_{GS} - V_T \tag{2.18}$$

Beyond this value, I_{DS} is obtained by substituting $V_{DS} = V_{GS} - V_T$ in Equation (2.17b), giving

$$I_{DS} = \frac{k}{2}(V_{GS} - V_T)^2 \tag{2.19}$$

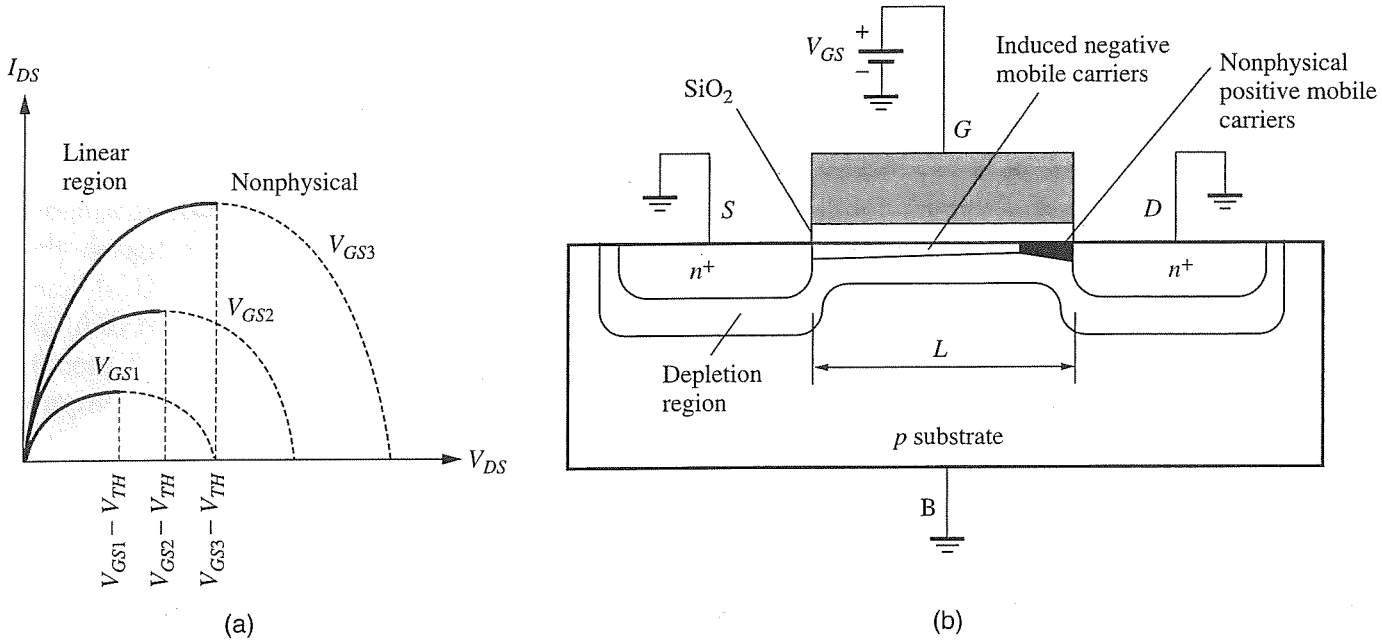


Figure 2.9
Nonphysical region of the drain current equation beyond pinch-off.

for the MOS transistor operating in this so-called *saturation* region.[8] The word *saturation* is used because I_{DS} reaches a limit, or saturates, at the level given by Equation (2.19). Figure 2.10 shows the currents in the linear and saturation regions of operation.

The drain current of an MOS transistor in the saturation region in fact is dependent on V_{DS}, because the depletion layer at the drain widens as V_{DS} increases, shortening the electrically effective value of L. Also, there is significant electrostatic coupling between the drain and the mobile charge in the channel, such that increasing the drain voltage increases Q_n above the value given by Equation (2.13). Each of these effects acts to increase the drain current as drain voltage increases. An empirical approximation to the actual drain current can be used to model this increase as

$$I_{DS} = \frac{k}{2}(V_{GS} - V_T)^2(1 + \lambda V_{DS}) \tag{2.20}$$

where the channel-length modulation parameter λ (lambda) represents the influence of V_{DS} on I_{DS} in saturation. To avoid discontinuities in the $I_{DS} - V_{DS}$ characteristic, the $1 + \lambda V_{DS}$ term may be included for both saturated and linear regions with negligible error. Usually the value of λ has little effect on the operating characteristics of digital MOS circuits.

[8] The consequences of the earlier assumption of V_T constant along the channel may now be understood. In fact, V_T increases with $V(y)$ due to body effect. The result is that saturated drain current I_D is 10 to 40% smaller than the value given by Equation (2.19b). However, good accuracy is obtained in circuit design if the value used for k' is determined from data taken with the transistor operating in saturation.

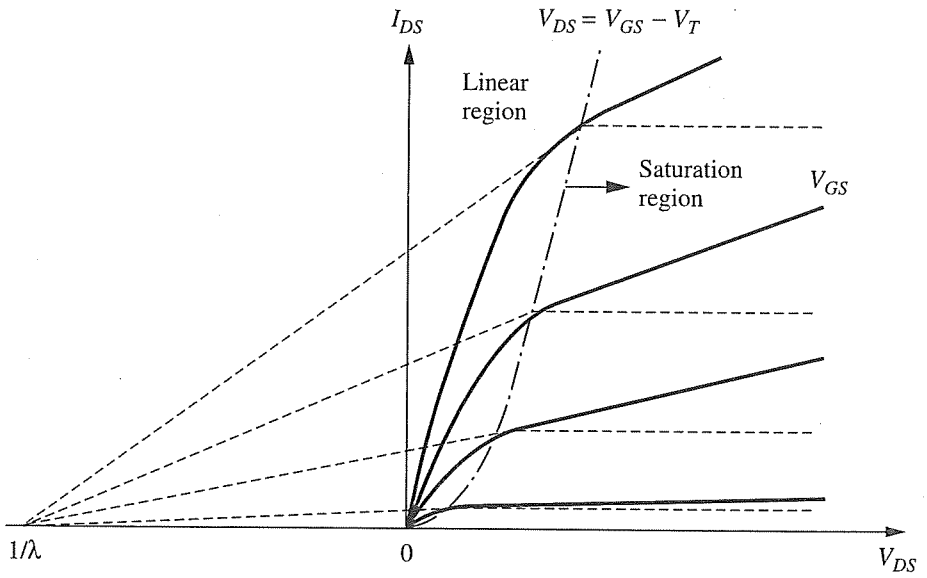


Figure 2.10
NMOS device $I_D - V_{DS}$ characteristics.

We summarize the results thus far using the I_{DS} versus V_{DS} plot for an NMOS transistor shown in Figure 2.10. Below pinch-off the device behaves as a nonlinear voltage-controlled resistor. This is termed the *linear, resistance, triode,* or *nonsaturation* region of operation. This book will refer to it as the linear region. Above pinch-off, the device approximates a voltage-controlled current source. The channel-length modulation parameter in the figure is much larger than usual, leading to a steeper slope for $I_D - V_{DS}$ in saturation than is usually observed. The slopes of all the curves in saturation converge at $1/\lambda$ along the V_{DS} axis.[9]

There is also a noticeable quadratic relationship between the current I_{DS} and V_{GS} in the saturation region. This makes sense since there is a $(V_{GS} - V_T)^2$ term in the current expression in saturation, so the current must increase quadratically with V_{GS}. This is an important characteristic associated with long-channel devices in saturation.

The separation between linear and saturation regions is also plotted in Figure 2.10. It is defined as the point where V_{DS} is equal to $V_{GS} - V_T$ for all the curves. Throughout the study of the material in this book, it will be important to quickly ascertain the region of operation of a device. As a rule of thumb, if V_{DS} is large relative to V_{GS}, then it is probably in saturation. If V_{GS} is large relative to V_{DS} then it is probably in the linear region. If the two are equal, then the device is definitely in saturation. This intuitive approach to determining the region of operation is useful as an initial guess, but it must be validated using Equation (2.18).

A similar set of plots can be generated for the PMOS device. As noted before, for PMOS devices, all polarities of voltages and current are reversed. When check-

[9] In bipolar transistors, this is referred to as the *Early* voltage.

ing for the region of operation, it is important to use the proper magnitudes of the voltages. Otherwise, incorrect results will be obtained.

MOS digital circuits are usually designed with MOS transistors as the only circuit elements. The values of V_T are usually given for the NMOS and PMOS devices but can be adjusted with body-bias, V_{BS}. All of the other parameters are process-dependent. As a consequence, the transistor width W and length L are typically the only design parameters available. Since L is usually selected as the minimum value possible, the only practical degree of freedom is the selection of W. Therefore, it is useful to relate most of the equations for the MOS transistor in terms of W. As we proceed through this chapter and subsequent chapters, useful coefficients will be computed relative to the W of the device.

2.5 Derivation of Velocity-Saturated Current Equations

The quadratic model derived above is valid for long-channel devices (typically $>$ 1 μm) and has served the IC industry well for many years, especially for hand calculations. However, it is not as suitable for today's deep submicron (DSM) technologies. The primary reason is that, in modern devices, the transistor channel lengths have been scaled to the point where the vertical and horizontal electric fields are large and they interact with one another. Furthermore, the assumption that saturation occurs at the pinch-off point is no longer correct. Saturation in DSM devices occurs when the carriers reach velocity saturation—that is, when they reach the speed limit of the carriers in silicon.

The effect on the current-voltage characteristics is illustrated in Figure 2.11 for a 0.18 μm device, for which V_{DS} and V_{GS} are swept from 0 to 1.8 V. Notice that the relationship between I_{DS} and V_{GS} is more linear than quadratic. This can be seen by visual inspection of the plot of I_{DS} versus V_{DS}, or directly in the plot of I_{DS} versus V_{GS}. This behavior is not predicted by the equations that we have derived thus far. As a result, a more suitable model is necessary for hand calculations.

Another observation is that the saturation voltage is smaller than what is predicted by the first-order model. The plot of I_{DS} versus V_{GS} indicates that the threshold voltage is around 0.5 V (x-intercept of plot). For the uppermost curve of I_{DS} versus V_{DS}, with $V_{GS} = 1.8$ V, the saturation condition of Equation (2.18) requires that $V_{DS} = V_{GS} - V_T = 1.8\text{ V} - 0.5\text{ V} = 1.3$ V. However, the curve saturates well before 1.3 V. It is closer to 0.6 V by inspection. This early saturation is true for all three curves in this plot and is due to velocity saturation. We now revisit the current derivations taking into account the high fields and velocity saturation.

2.5.1 Effect of High Fields

Consider the following trends in the horizontal and vertical fields. The horizontal field occurs in the y-direction which is in the direction of the channel pointing from drain to source. The vertical field occurs in the x-direction from the gate to the channel. The horizontal field is given by $E_y = V_{DS}/L$. Shown in the following are estimates of the horizontal field for three different years. We can see that the

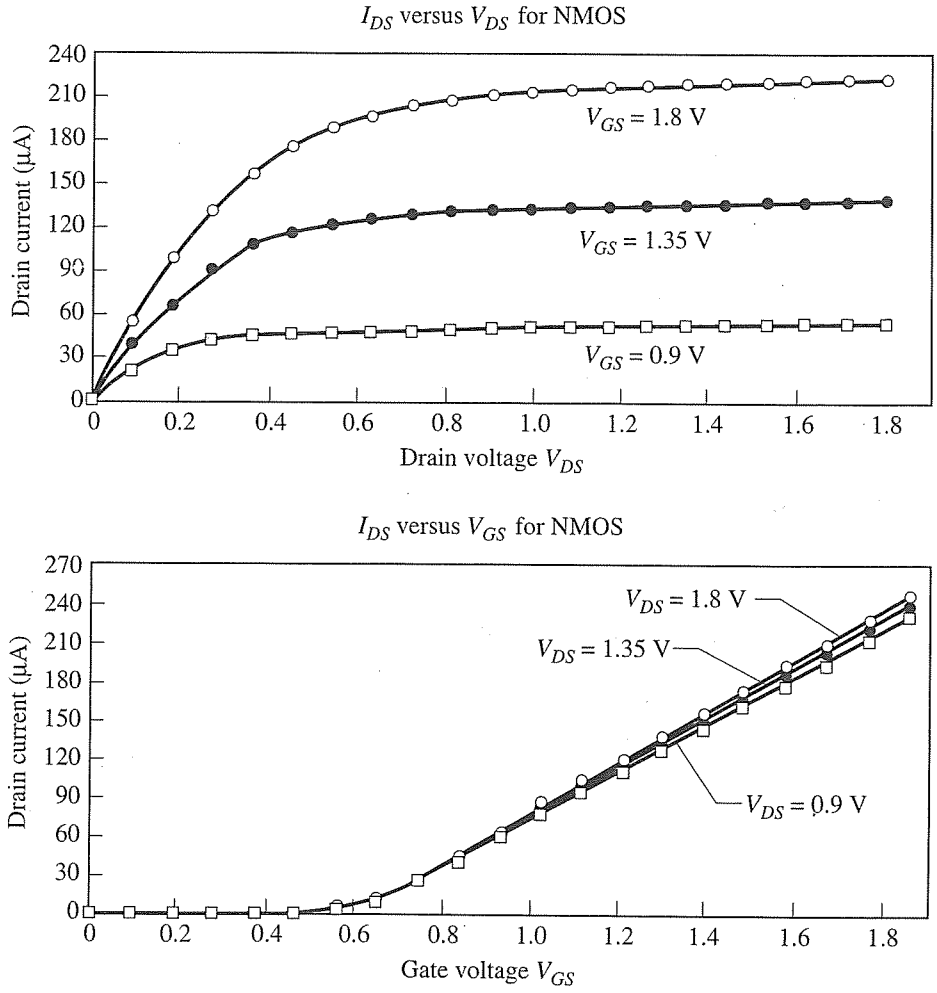


Figure 2.11
Current-voltage characteristics of a 0.18 μm process.

horizontal field has increased over the years by an order of magnitude and has remained approximately at 10^5 V/cm since 1995. The horizontal field acts to push the carriers to their velocity limit and this causes early saturation, which in turn degrades the mobility. The carrier velocity and the electric field no longer have a linear relationship as was assumed in the long-channel case.

1980

$$E_y = \frac{5\text{V}}{5\,\mu\text{m}} = 10^4\ \text{V/cm}$$

1995

$$E_y = \frac{3.3\text{V}}{0.35\,\mu\text{m}} = 9.4 \times 10^4\ \text{V/cm}$$

2001

$$E_y = \frac{1.2V}{0.1\,\mu\text{m}} = 1.2 \times 10^5\ \text{V/cm}$$

The vertical field can be approximated as $E_x = V_{DD}/t_{ox}$. Shown in the following are estimates of the vertical field for three different years. We can see that the vertical field has also increased by an order of magnitude and has remained approximately at 5×10^6 V/cm.

1980

$$E_x = \frac{5\text{V}}{1000\ \text{Å}} = 50 \times 10^4\ \text{V/cm}$$

1995

$$E_x = \frac{3.3\text{V}}{75\ \text{Å}} = 4.4 \times 10^6\ \text{V/cm}$$

2001

$$E_x = \frac{1.2\text{V}}{22\ \text{Å}} = 5.5 \times 10^6\ \text{V/cm}$$

For high gate voltages, a large number of mobile carriers are induced in the inversion layer near the interface. The mobility of these carriers decreases due primarily to electron scattering caused by dangling bonds at the Si-SiO_2 interface. The effect of the vertical field on mobility can be modeled to first-order as follows:

$$\mu_e = \frac{\mu_0}{1 + \left(\dfrac{V_{GS} - V_T}{\theta \cdot t_{ox}}\right)^{\eta}} \tag{2.21}$$

Here, μ_0 is the nominal mobility in the presence of low fields, typically 540 cm^2/V-sec for NMOS devices, and θ and η are empirical values. The equation uses a term based on the vertical electric field, $(V_{GS} - V_T)/t_{ox}$, to reduce the nominal mobility value. As an example of mobility degradation, let $\theta = 3.6 \times 10^6$ V/cm and $\eta = 1.85$ in a 0.13 μm technology. Then, with $t_{ox} = 22$ Å and $V_{GS} - V_T \approx 1.2\ \text{V} - 0.4\ \text{V} = 0.8$ V, we find that the mobility, $\mu_e \approx 270\ cm^2/V\text{-sec}$, is reduced by a factor of 2 relative to μ_0. This effective value of mobility can be used to account for the high vertical field. The effective mobility for PMOS devices is 70 cm^2/V-sec which is about four times smaller than NMOS devices.

The horizontal field acts to reduce the mobility even further. We assumed that as E_y goes up, the carriers continue to increase in speed. Actually their velocity saturates (or reaches a velocity limit) at approximately $v_{sat} = 10^7$ cm/s. Consider Figure 2.12, which shows the relationship between the carrier velocity and horizontal electric field. Initially, as we increase the E_y field, the carrier velocity also increases. The linear proportionality constant between the velocity and the electric field is, of course, the mobility. Note that as the field increases beyond a certain critical electrical field, E_c, the carrier velocity saturates at its limit in silicon. The fields are so high in DSM devices that they tend to saturate very quickly as V_{DS} increases. This is the basic difference between long- and short-channel devices.

Figure 2.13 shows the experimental results for drift velocities, of holes and electrons in silicon. The saturation velocity for both electrons and holes is 8×10^6

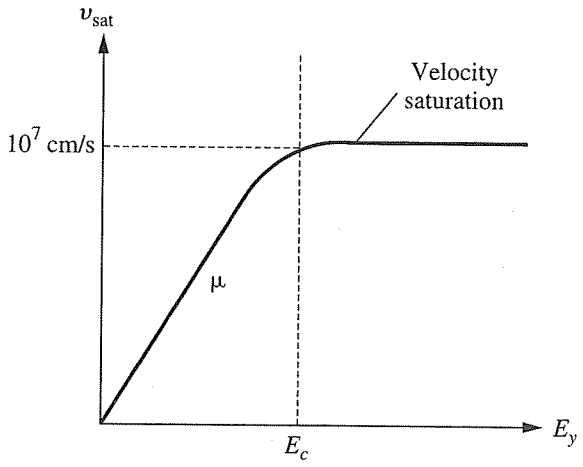


Figure 2.12
Carrier velocity versus electric field.

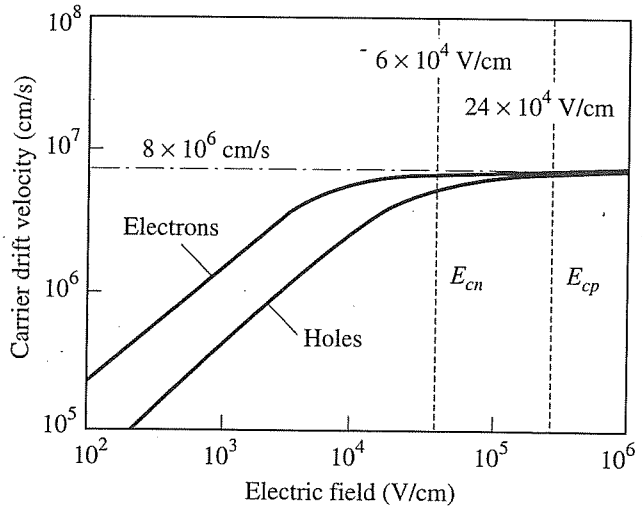


Figure 2.13
Plot of carrier drift velocity versus electric field at 400° K. (From Sze.)

cm/sec at $T = 400°$ K, and is independent of the doping level. This value is used in most of the analysis that follows. From the curves, we estimate the critical field values to be

$$E_{cn} = 6 \times 10^4 \quad \frac{\text{V}}{\text{cm}} \text{ for electrons}$$
$$E_{cp} = 24 \times 10^4 \quad \frac{\text{V}}{\text{cm}} \text{ for holes} \tag{2.22}$$

As the temperature decreases, the saturation velocity increases. For $T = 300°$ K, $v_{sat} \approx 10^7$ cm/s.

The next step is to somehow incorporate the effects of the high fields on mobility into a set of equations suitable for hand calculation. We first include the effect of the vertical field by using Equation (2.21). To account for the horizontal field, we can express the velocity as a piecewise continuous function of the horizontal electric field, E_y, as follows:

$$v = \mu_e \frac{E_y}{\left(1 + \frac{E_y}{E_c}\right)} \qquad E_y < E_c \tag{2.23a}$$

$$v = v_{sat} \qquad E_y \geq E_c \tag{2.23b}$$

That is, the relationship between the carrier velocity and the field is divided into two segments with a boundary defined by E_c, the critical field. Before the critical field is reached, the value of v is given by Equation (2.23a). Beyond E_c, the velocity saturates at v_{sat}. This captures the basic behavior of the curve but we should remember that it is only a model of the true behavior.[10] In order to ensure continuity at the boundary, where $E_y = E_c$, we can plug in this condition (2.23a) and set $v = v_{sat}$.

$$v = \mu_e \frac{E_y}{\left(1 + \frac{E_y}{E_c}\right)} = \frac{\mu_e E_c}{2} \tag{2.24}$$

$$\therefore E_c = \frac{2v_{sat}}{\mu_e}$$

This sets up the relationship between E_c and v_{sat}. Note that this relationship is the result of the modeling approach described above as opposed to a fundamental relationship between the critical field and the saturated velocity. It holds true only in the context of Equation (2.23).

2.5.2 Current Equations for Velocity-Saturated Devices

To derive more suitable MOS current equations, we use Equation (2.14) and set v based on the region of operation using the conditions specified in Equations (2.23a) and (2.23b).

[10] This model slightly underestimates the velocity below the critical field and slightly overestimates the velocity above the critical field. But the model leads to simple current equations that are accurate enough for hand calculations.

In the linear region of operation,

$$I_{DS} = W \times Q_n \times v$$

$$= W \times C_{ox}(V_{GS} - V_T - V(y))\left(\frac{\mu_e E_y}{1 + \frac{E_y}{E_c}}\right)$$

Reorganizing the equation and applying the relationship between field and potential, we obtain

$$I_{DS} = \left(\frac{W}{1 + \frac{E_y}{E_c}}\right) C_{ox}(V_{GS} - V_T - V(y))\mu_e E_y$$

$$\text{where } E_y = \frac{dV(y)}{dy}$$

Plugging in and re-arranging produces

$$I_{DS}\,dy = W\mu_e\left[C_{ox}(V_{GS} - V_T - V(y)) - \frac{I_{DS}}{W\mu_e E_c}\right]dV(y)$$

After integration, we obtain

$$I_{DS} = \frac{W}{L}\frac{\mu_e C_{ox}}{\left(1 + \frac{V_{DS}}{E_c L}\right)}\left(V_{GS} - V_T - \frac{V_{DS}}{2}\right)V_{DS} \tag{2.25}$$

This expression for the linear region is very similar to the previously derived equation except for the extra multiplier in the denominator. This is very convenient since we are only required to remember one additional term to obtain a more accurate current expression.

The next step is to determine the expression for the saturation region of operation. Saturation occurs when the carriers are moving at v_{sat}:

$$I_{DS} = W \times Q_n \times v_{sat} \tag{2.26}$$

Since the current is the same throughout the channel we can set $V(y) = V_{DS}$ and write

$$I_{DS} = W \times C_{ox}\,(V_{GS} - V_T - V_{DS})v_{sat} \tag{2.27}$$

We can further simplify it by equating the expressions of I_{DS} in linear (2.25) and saturation (2.27) regions to obtain V_{Dsat}:

$$V_{Dsat} = \frac{(V_{GS} - V_T)\,E_c L}{(V_{GS} - V_T) + E_c L} \tag{2.28}$$

The difference between this expression and the first-order expression for V_{Dsat} is the extra multiplier (recall that $V_{Dsat} = V_{GS} - V_T$ in the first-order model). The new factor $E_cL/(V_{GS} - V_T + E_cL) < 1$, so the value of the V_{Dsat} will always be lower than the first-order model. This is the effect of velocity saturation—devices saturate faster and deliver less current than the quadratic model would predict.

If we now substitute this new V_{Dsat} into our saturation current equation of (2.27) we obtain

$$I_{DS} = Wv_{sat}C_{ox}\frac{(V_{GS} - V_T)^2}{(V_{GS} - V_T) + E_cL} \tag{2.29}$$

Again, this equation is similar to the first-order current equation and will be easy to remember. It is different from Equation (2.19) but has a familiar quadratic form in the numerator.

Consider the limiting cases for this equation. What happens if $E_cL >> V_{GS} - V_T$ (i.e., a long-channel device)? The equation reduces to our quadratic expression:

$$I_{DS} = \frac{W}{2L}\mu_e C_{ox}(V_{GS} - V_T)^2$$

Now consider what happens if $E_cL << V_{GS} - V_T$, that is, if we have a very short channel. We obtain an equation where the current is linear in $V_{GS} - V_T$, as we would expect for very short-channel devices:

$$I_{DS} = Wv_{sat}C_{ox}(V_{GS} - V_T)$$

NMOS and PMOS Saturation Voltages for 0.18 μm Technology **Example 2.6**

Problem:

Consider a 0.18 μm technology. Compute the values of V_{Dsat} for the NMOS and PMOS device assuming $V_{GS} = 1.8$ V, $V_{TN} = 0.5$ V, $V_{TP} = -0.5$ V. Assume the channel length is 200 nm for convenience.

Solution:

Using (2.22), we find that $E_{cn}L_n = 6 \times 10^4$ (0.2 μm) $\approx$ 1.2 V and $E_{cp}L_p = 24 \times 10^4$ (0.2 μm) $\approx$ 4.8 V. Using (2.28),

NMOS: $$V_{Dsat} = \frac{(1.8 - 0.5)(1.2)}{(1.8 - 0.5 + 1.2)} \approx 0.6 \text{ V}$$

PMOS: $$V_{Dsat} = \frac{(1.8 - 0.5)(4.8)}{(1.8 - 0.5 + 4.8)} \approx 1.0 \text{ V}$$

The value computed for the NMOS device is consistent with the plot shown in Figure 2.11.

The results of Example 2.6 are important in device sizing. The fact that the PMOS device saturates at a higher voltage implies that it will be able to supply a

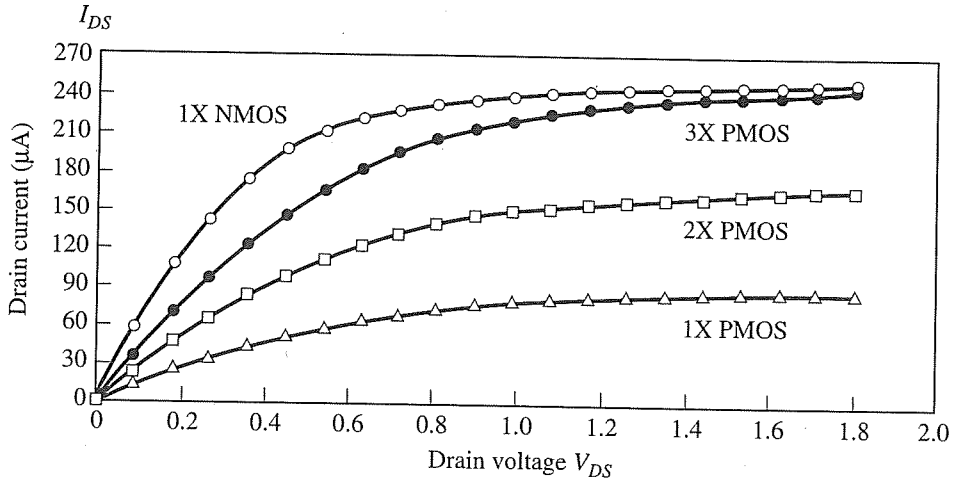


Figure 2.14
Effect of mobility and velocity saturation on device sizes.

higher than expected amount of current. To illustrate this, we plot the current for both NMOS and PMOS devices with the minimum (W/L) ratio in Figure 2.14. These two cases are labeled as 1X devices. In addition, we plot a 2X PMOS device and a 3X PMOS device. The 1X NMOS device and the 3X PMOS device deliver the same current in saturation even though mobility difference between the two 1X devices is 4:1. The device ratio is only 3:1 since the PMOS device saturates at a higher voltage.

To compare with hand analysis, consider the current ratio of the two 1X devices using Equation (2.29):

$$\frac{I_{DsatN}}{I_{DsatP}} = \frac{W_N \upsilon_{sat} C_{ox} (V_{GS} - V_{TN})^2/(V_{GS} - V_{TN} + E_{CN}L_N)}{W_P \upsilon_{sat} C_{ox} (V_{GS} - V_{TP})^2/(V_{GS} - V_{TP} + E_{CP}L_P)}$$

Plugging in our nominal values, we obtain

$$\frac{I_{DsatN}}{I_{DsatP}} = 2.4$$

Therefore, the required device ratio to deliver the same current in saturation is about 2.4X rather than 4X due to the difference in V_{Dsat}. Hand analysis underestimates the current ratio in saturation.

Example 2.7 **NMOS and PMOS Currents in Saturation for 0.13 μm Technology**

Problem:

Compute the saturation currents per micron of width for a 0.13 μm technology. Assume a channel length of 100 nm, $t_{ox} = 22$ Å, $V_{TN} = 0.4$ V, $V_{TP} = -0.4$ V, $V_{DD} = 1.2$ V. Use $\upsilon_{sat} = 8 \times 10^6$ cm/s.

Solution:

Using (2.22), we find that $E_{cn}L_n = 6 \times 10^4$ V/cm (0.1 μm) $\approx$ 0.6 V and $E_{cp}L_p = 24 \times 10^4$ V/cm (0.1 μm) $\approx$ 2.4 V.

Then, for the *n*-channel device:

$$I_{DS} = Wv_{sat}\, C_{ox} \frac{(V_{GS} - V_T)^2}{(V_{GS} - V_T) + E_cL}$$

$$\therefore \frac{I_{DS}}{W} = v_{sat}C_{ox} \frac{(V_{GS} - V_T)^2}{(V_{GS} - V_T) + E_cL}$$

$$= (8 \times 10^6)(1.6 \times 10^{-6}) \frac{(1.2 - 0.4)^2}{(1.2 - 0.4) + 0.6} = 585\ \mu A/\mu m$$

For the *p*-channel device,

$$\therefore \frac{I_{DS}}{W} = v_{sat}C_{ox} \frac{(V_{GS} - V_T)^2}{(V_{GS} - V_T) + E_cL}$$

$$= (8 \times 10^6)(1.6 \times 10^{-6}) \frac{(1.2 - 0.4)^2}{(1.2 - 0.4) + 2.4} = 256\ \mu A/\mu m$$

To summarize the equations for deep submicron devices, we should first decide whether a device is in the linear region or saturation using the following inequalities:

$$\text{if } V_{DS} \geq \frac{(V_{GS} - V_T)\, E_cL}{(V_{GS} - V_T) + E_cL} \Rightarrow \text{saturation}$$

$$\text{if } V_{DS} < \frac{(V_{GS} - V_T)\, E_cL}{(V_{GS} - V_T) + E_cL} \Rightarrow \text{linear}$$

If we are in the saturation region, we should use the equation

$$I_{DS} = Wv_{sat}C_{ox} \frac{(V_{GS} - V_T)^2}{(V_{GS} - V_T) + E_cL} \quad \text{saturation}$$

Note that there is still a channel length modulation effect, which we can capture as we did before. This involves the use of a parameter, λ, as follows:

$$I_{DS} = Wv_{sat}C_{ox} \frac{(V_{GS} - V_T)^2}{(V_{GS} - V_T) + E_cL} (1 + \lambda V_{DS}) \quad \text{saturation}$$

The use of this equation is perhaps too cumbersome for hand calculations. Therefore, we will drop the last term and ensure that the I_{DS} is computed using the average value in saturation. For convenience, we set λ=0.

If we are in the linear region, we should use the equation

$$I_{DS} = \frac{W}{L} \cdot \frac{\mu_e C_{ox}}{\left(1 + \dfrac{V_{DS}}{E_c L}\right)} \left(V_{GS} - V_T - \frac{V_{DS}}{2}\right) V_{DS} \quad \text{linear}$$

Example 2.8 **Current Ratio for 0.13 μm Process**

Problem:

Consider a 0.13 μm technology where the channel length is 100 nm. Using (2.22), $E_{cn}L_n = 0.6$ V and $E_{cp}L_p = 2.4$ V, and assuming that $V_{GS} = 1.2$ V and $V_{TN} = 0.4$ V, $V_{TP} = -0.4$ V, compute V_{Dsat} and the current ratio of two 1X devices in saturation.

Solution:

Using (2.28) we find that

$$\text{NMOS:} \quad V_{Dsat} = \frac{(1.2 - 0.4)(0.6)}{(1.2 - 0.4 + 0.6)} = 0.34 \text{ V}$$

$$\text{PMOS:} \quad V_{Dsat} = \frac{(1.2 - 0.4)(2.4)}{(1.2 - 0.4 + 2.4)} = 0.6 \text{ V}$$

The current ratio of the two 1X devices is

$$\frac{I_{DsatN}}{I_{DsatP}} = \frac{W_N v_{sat} C_{ox} (V_{GS} - V_{TN})^2/(V_{GS} - V_{TN} + E_{CN}L_N)}{W_P v_{sat} C_{ox} (V_{GS} - V_{TP})^2/(V_{GS} - V_{TP} + E_{CP}L_P)}$$

Plugging in our nominal values, we obtain

$$\frac{I_{DsatN}}{I_{DsatP}} = 2.3$$

*2.6 Alpha-Power Law Model

Another approach for a hand-calculation model is to empirically fit the real data to the following form of I_{DS} in saturation:

$$I_{DS} = K_S \frac{W}{L} (V_{GS} - V_T)^{\alpha} \tag{2.30a}$$

In this formulation, we can set K_S and α based on measured data. Clearly, α should be set to a value that is closer to 1 than 2. Today, α is approximately 1.25 but it will continue to approach 1 as technology scales.

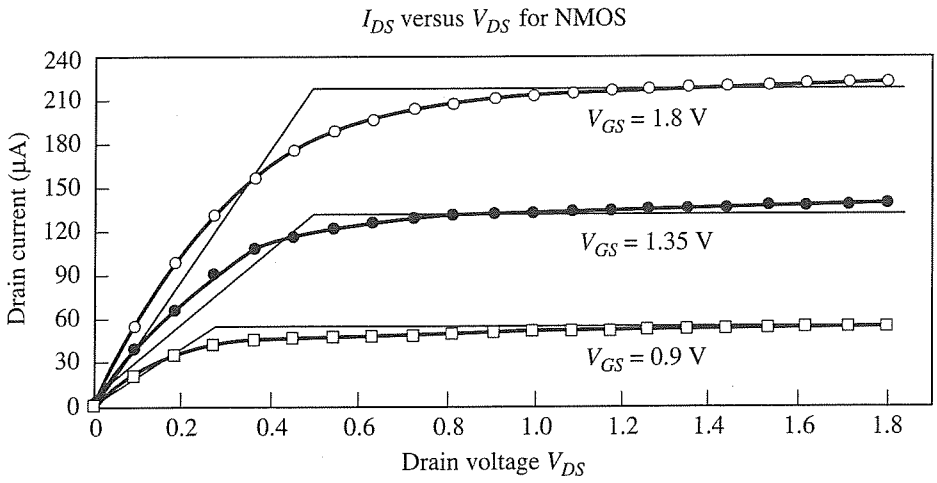


Figure 2.15
Alpha-power law modeling of MOS transistor.

Note that this model does not account for the behavior of the device in the linear region. Therefore, a simple model must be developed for the alpha-power law that fits both the linear and saturation regions, without a mismatch at the boundary between the two regions. An example of such a model for the linear region is to set

$$I_{DS} = K_L \frac{W}{L} (V_{GS} - V_T) V_{DS} \tag{2.30b}$$

V_{Dsat} can be obtained by equating (2.30a) to (2.30b):

$$V_{Dsat} = \frac{K_S}{K_L} (V_{GS} - V_T)^{\alpha - 1} \tag{2.31}$$

The current versus voltage plots based on alpha-power law modeling is shown in Figure 2.15. The transistor characteristics in the saturation region are described by (2.30a) while the linear region is given by (2.30b). The interface between the two regions is obtained from (2.31). This model is suitable for hand calculations only after curve fitting is performed with the data points in the figure. If a new technology is developed, a new set of parameters must be extracted. The model derived earlier based on velocity saturation is a more general model and will be the primary model used in the rest of this book.

Parameters for Alpha-Power Law **Example 2.9**

Problem:

Find K_S and α based on Figure 2.15 for the NMOS device using the saturation region alpha-power law model. Assume $(W/L) = 1$ and $V_T = 0.5$ V.

Solution:

From the plot of Figure 2.15, we first estimate the saturation value of I_{DS} at $V_{GS} = 1.35$ V to be 130 μA. Then we estimate the value of I_{DS} at $V_{GS} = 1.8$ V at the same value of V_{DS} to be 220 μA. We can write the current equation for each measurement as

$$I_{DS} = K_S \frac{W}{L} (V_{GS} - V_T)^{\alpha}$$

and take their ratio, as follows:

$$\frac{220\ \mu A}{130\ \mu A} = \frac{K_S(1.8 - 0.5)^{\alpha}}{K_S(1.35 - 0.5)^{\alpha}}$$

Solving for α, we obtain roughly 1.25. Applying this value back into the current equation, we find that $K_S = 160\ \mu A/V^{1.25}$. It is interesting to note that the exponent is close to 1. This is expected since the long-channel device has an exponent of 2 and a very short-channel device would have an exponent of 1.

One way to check the results would be to compute the current for the case when $V_{GS} = 0.9$ V:

$$I_{DS} = 160 \frac{\mu A}{V^{1.25}} (1)(0.9 - 0.5)^{1.25} \approx 50\ \mu A$$

This value is close to the expected value based on Figure 2.15.

2.7 Subthreshold Conduction

So far, there has been an implicit assumption that the current I_{DS} in the cutoff region is zero. In actual fact, there is appreciable transistor current even if $V_{GS} < V_T$, especially for deep submicron transistors. This section explores this previously negligible component of current that is increasingly important as technology scales. The magnitude of the current is small compared to the current when $V_{GS} > V_T$. However, if we have millions of such devices all leaking current, the standby power dissipation will be large. This is becoming a big concern for deep submicron technologies. As we will see in later chapters, this leakage current will prove to be problematic for dynamic logic circuits and dynamic memories, as it will discharge capacitances that store logic values. In technologies prior to 0.18 μm, this current was considered a second-order effect, but it is now a first-order concern and deserves some attention in this book.

In the derivation of the threshold voltage expression, the strong inversion condition was defined (somewhat arbitrarily) as the point at which the surface potential is $2|\phi_F|$ below the level in bulk silicon. However, in practice, surface inversion occurs well before this point and there are mobile carriers in the channel region capable of conducting current. The term *subthreshold region* is preferred to describe the case where $V_{GS} < V_T$, rather than *cutoff*, since there is current flow. As shown in Figure 2.16, I_{DS} is a continuous function of V_{GS} and drops off in a logarithmic fashion in the subthreshold region as V_{GS} is decreased. Figure 2.16 illustrates all three regions of operation and the associated current characteristics of the NMOS and PMOS devices.

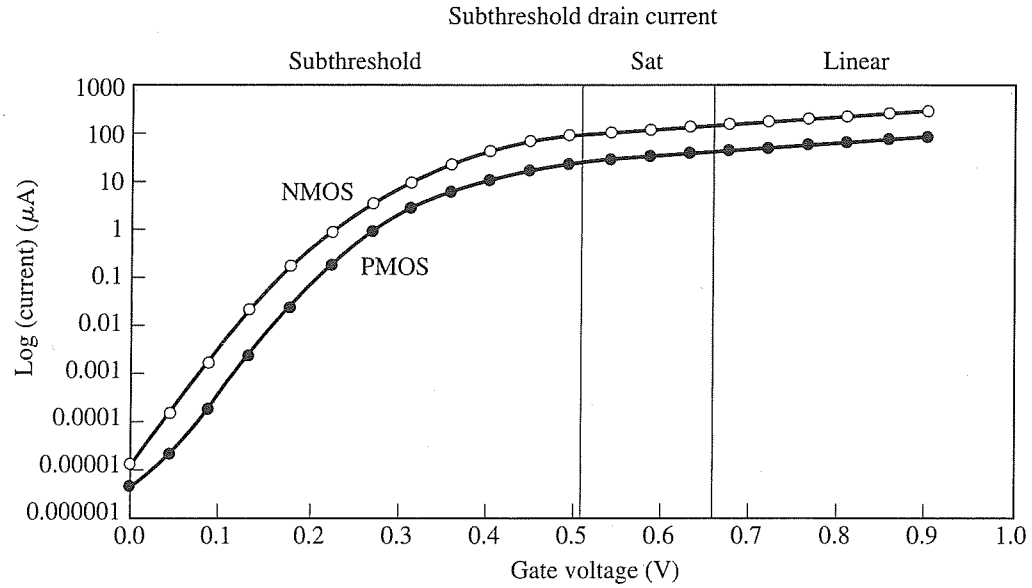


Figure 2.16
MOS current versus V_{GS}.

The mechanism for subthreshold current flow is due to the diffusion of minority carriers when the gate voltage is several thermal voltages less than V_T. In the subthreshold region, the MOS transistor behaves more like a (lateral) bipolar transistor (see Appendix B). The substrate is the base region, while the source and drain act as the emitter and collector, respectively. Therefore, modeling of the current can be carried out using a derivation based on bipolar modeling. In particular, the current equation in this region takes the form:

$$I_{sub} = I_s e^{\frac{q(V_{GS} - V_T - V_{offset})}{nkT}} \left(1 - e^{\frac{-qV_{DS}}{kT}}\right) \tag{2.32}$$

where I_s represents a current coefficient, V_{offset} is the sum of a number of voltage terms and lies in the range -0.1 to 0.1, and the factor n is a subthreshold swing parameter, typically ranging from 1 to 2. We can determine the value of n by considering how much voltage change in V_{GS} produces an *order of magnitude* change in the subthreshold current:

$$\frac{10\, I_{sub}}{I_{sub}} = \frac{I_s e^{\frac{q(V_{GS1} - V_T - V_{offset})}{nkT}} \left(1 - e^{\frac{-qV_{DS}}{kT}}\right)}{I_s e^{\frac{q(V_{GS2} - V_T - V_{offset}}{nkT}} \left(1 - e^{\frac{-qV_{DS}}{kT}}\right)}$$

$$\therefore 10 = \frac{e^{\frac{q(V_{GS1} - V_T - V_{offset})}{nkT}}}{e^{\frac{q(V_{GS2} - V_T - V_{offset})}{nkT}}} = e^{\frac{q(V_{GS1} - V_{GS2})}{nkT}}$$

Taking the natural logarithm of both sides and solving for $\Delta V_{GS} = V_{GS} - V_{GS2}$, we obtain a metric for the quality of the subthreshold region, called the *slope factor:*

$$S = \Delta V_{GS} = \frac{nkT}{q} \ln(10) \tag{2.33}$$

which is specified in units of mV/decade. The thermal voltage at room temperature is approximately 26 mV.

In the subthreshold region, it is desirable to have the current drop off significantly as we reduce V_{GS}. This implies a small n in Equation (2.33). Ideally, $n = 1$ which leads to a slope factor of 60 mV/decade at room temperature. However, n is usually between 1.5 or 2 making the slope factor about 90–120 mV/decade. The proper value of n must be determined from device measurements.

Example 2.10 **Subthreshold Swing Parameter Calculation**

Problem:

From the plot of $\log(I_{DS})$ versus V_{GS} in Figure 2.16 with $V_{DS} = 1.8$ V, find n and the slope factor in the subthreshold region for the PMOS device.

Solution:

By measuring the subthreshold current at 10 nA and 100 nA the values for V_{GS} are

	V_{GS} (V)
10 nA	0.140
100 nA	0.212

From these points, $n = 1.2$ for PMOS. The slope factor is approximately $S = 60$ mV/decade $\times$ 1.2 = 72 mV/decade, which is close to the ideal value.

2.8 Capacitances of the MOS Transistor

The switching speed of MOS digital circuits is limited by the time required to charge and discharge the capacitances at internal nodes. Within VLSI circuits most of these capacitances are so small that they are difficult to measure directly. For circuit analysis, these capacitances must be calculated from device dimensions and dielectric constants. The capacitance values are usually specified in femto-Farads per μm of width (i.e., fF/μm). In the description to follow, the subscripts will be lowercase to imply these units. Total capacitance values use uppercase subscripts.

Figure 2.17 shows the significant capacitances between nodes of an MOS transistor. There are two basic types of nonlinear or voltage-dependent capacitances in the structure: thin-oxide capacitances and junction capacitances. The thin-oxide capacitances comprised of C_{gs}, C_{gd}, and C_{gb} are represented by C_g. The junction

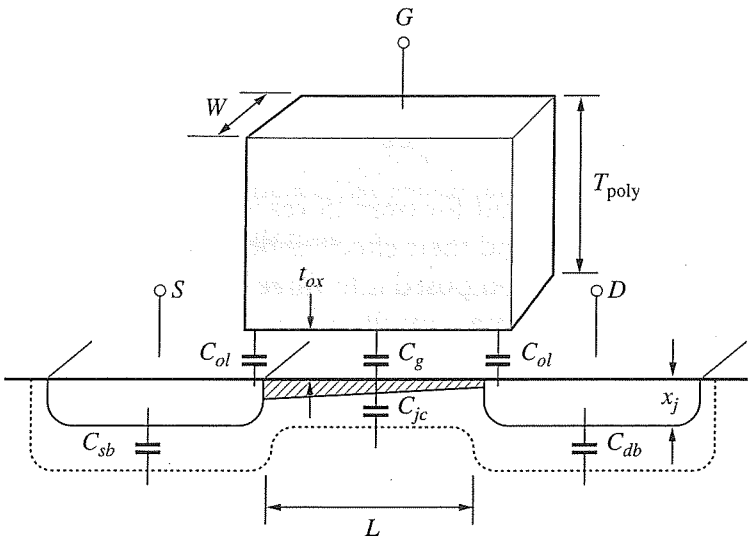


Figure 2.17
Capacitances of the MOS transistor.

capacitances are shown as C_{sb}, and C_{db}. In addition, there are two overlap capacitances, C_{ol}, which are linear and voltage-independent. Finally, the depletion layer capacitance under the channel, C_{jc}, is associated with C_{gb}, although a small part of this capacitance is associated with the drain and source on either edge. These types of capacitances are described in more detail in the sections below.

2.8.1 Thin-Oxide Capacitance

The thin-oxide capacitance is perhaps the most important capacitance in the MOS system. The two plates of the capacitance are defined as the gate and the channel. The dielectric material is the oxide sandwiched between these two plates. The total capacitance of the thin-oxide is:

$$C_G = WLC_{ox} = WL\frac{\varepsilon_{ox}}{t_{ox}} = WC_g \tag{2.34}$$

where C_{ox} is the capacitance per unit area of the gate dielectric as defined in Equation 2.5. It is interesting to examine the factor, C_g. In a 5 μm technology, the oxide thickness was approximately 1100 Å. Therefore,

$$C_g = C_{ox}L = \frac{\varepsilon_{ox}}{t_{ox}}L = \frac{(4)(8.85 \times 10^{-14})}{1100}(5\ \mu\text{m}) \cong 1.6\ \text{fF}/\mu\text{m}$$

In a 0.35 μm process, with $t_{ox} = 75$Å,

$$C_g = C_{ox}L = \frac{\varepsilon_{ox}}{t_{ox}}L = \frac{(4)(8.85 \times 10^{-14})}{75}(0.35\ \mu\text{m}) \cong 1.6\ \text{fF}/\mu\text{m}$$

In a 0.13 μm process, with $L = 0.1$ μm and $t_{ox} = 22$Å,

$$C_g = C_{ox}L = \frac{\varepsilon_{ox}}{t_{ox}}L = \frac{(4)(8.85 \times 10^{-14})}{22}(0.1\ \mu\text{m}) \cong 1.6\ \text{fF}/\mu\text{m}$$

This factor has remained constant for over 25 years! The reason is that both L and t_{ox} are scaled at the same rate, and their effects cancel each other out.

The gate capacitance is decomposed into three capacitances: the gate-to-source capacitance, C_{gs}; the gate-to-drain capacitance, C_{gd}; and, the gate-to-bulk capacitance, C_{gb}. The division of C_g into its three elements is fairly complex. The components vary depending on whether the device is in cutoff, linear, saturation, or accumulation, as shown in Figure 2.18. This a plot of the three capacitances as a function of V_{GS}. In the linear region, C_{gs} and C_{gd} are approximately equal to $(1/2)C_g$ since the channel extends from source to drain. In the saturation region, the channel extends most of the way from source to drain (recall that the saturation condition is defined by the carriers reaching velocity saturation), so most of the gate capacitance can be attributed to the source node, and a negligible amount to the drain node. For this region, a detailed analysis would show the capacitances to be $C_{gs} = (2/3)C_g$ and $C_{gd} = 0$.

When the device is in cutoff, then $C_{gs} = C_{gd} = 0$. All of the capacitance is attributed to C_{gb}, the gate-to-bulk capacitance. However, in this regime, there is a depletion region under the gate and we have C_g in series with C_{jc}, which is the channel junction capacitance. As a result, the total capacitance is less than C_g (since it is in series with C_{jc}) until we reach the accumulation region. At this point, the channel is positively charged while the gate is negatively charged producing a total capacitance

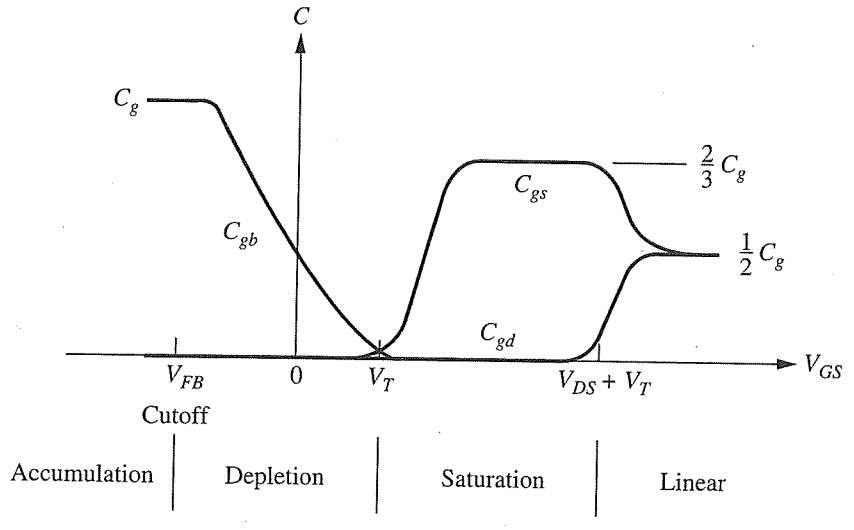


Figure 2.18
Capacitances based on region of operation.

of C_g. In normal operation, we are not usually concerned with the accumulation region of operation, but it is included for completeness. Note that when $V_{GS} = 0$, the value of C_{gb} is about 1/2 C_g. It does not reach C_g until the gate voltage is equal to the flat-band voltage.

Thin-Oxide Capacitance Calculation **Example 2.11**

Problem:

Compute the gate capacitance in the cutoff, linear, and saturation regions for a PMOS device with t_{ox} = 22 Å and device dimensions of W = 400 nm and L = 100 nm. Assume that $V_{GS} = 0$ in the cutoff region.

Solution:

The total capacitance is $C_gW = 1.6$ fF/μm $\times$ 0.4 μm = 0.64 fF.
In cutoff: $C_{GS} = 0, C_{GD} = 0, C_{GB} \cong 0.64$ fF/2 = 0.32 fF
In linear: $C_{GS} = 0.32$ fF, $C_{GD} = 0.32$ fF, $C_{GB} = 0$
In saturation: $C_{GS} = 0.43$ fF, $C_{GD} = 0, C_{GB} = 0$

2.8.2 *pn* Junction Capacitance

The source and drain regions and the substrate form *pn* junctions that give rise to two additional capacitances. The capacitances C_{sb} and C_{db} are n^+p source/drain junction capacitances for NMOS devices and are readily calculated using layout information. For PMOS devices, the capacitances are due to p^+n source/drain junctions. In addition, there is a junction capacitance between the inverted channel and the substrate, C_{jc}. Since the *pn* junction is a diode, it is worth revisiting some of the basic physics for this diode circuit element.

The current-voltage characteristic of a diode is given by

$$I_D = I_S(e^{V_J/V_{th}} - 1) \tag{2.35}$$

where I_S is the reverse saturation current of the diode, V_J is the voltage drop across the diode, and V_{th} is the thermal voltage. In normal operation, the *pn* junctions are all reverse-biased in a MOS transistor. When specifying the voltages at the source/drain and bulk terminals, we must ensure that the bulk is connected to the lowest voltage for NMOS devices. Similarly, for PMOS devices, the bulk voltage must be the highest potential in the system. Otherwise, we risk forward-biasing the diode.

Since $V_J < 0$, the exponential term in Equation (2.35) is small. In essence, for our MOS devices,

$$I_D = -I_S \tag{2.36}$$

This represents the *leakage* current of the MOS transistor, which is unwanted current flow from source/drain regions into the substrate. The actual value of leakage

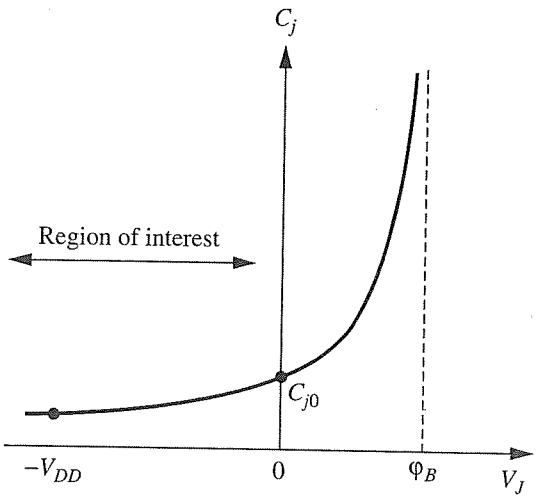


Figure 2.19
Plot of junction capacitance versus applied voltage.

is relatively small and depends on the area and the doping levels on each side of the junction.

The built-in junction potential for a diode is another important quantity in *pn* junction theory. It is computed using the equation:

$$\phi_B = \frac{kT}{q} \ln \frac{N_A N_D}{n_i^2} \tag{2.37}$$

Now we return to the more important discussion of junction capacitance. The depletion region of the diode has a capacitance effect associated with it since the modulation of diode voltage V_J changes the charge that is exposed or "covered up" in this *space-charge region.* This is equivalent to the action of a capacitor. The basic expression for junction capacitance is given by

$$C_J = \frac{C_{j0} A}{\left(1 - \frac{V_J}{\phi_B}\right)^m} \tag{2.38}$$

where C_{j0} is the zero-bias junction capacitance, A is the area of the junction, ϕ_B is the built-in junction potential, and m is the junction grading coefficient, which is approximately 1/2 for abrupt junctions (p^+n or n^+p).This function is plotted in Figure 2.19. The built-in junction potential defines the point where C_j asymptotically approaches infinity.[11] We can see that C_{j0} is an important quantity since it

[11] This would not occur in practice. The capacitance would reach some level and begin to "roll-over" as V_J increases. In any case, we are mostly interested in the reverse-biased region of operation where $V_J < 0$.

represents the value of junction capacitance when the external bias is 0 V. Its value is computed using the equation

$$C_{j0} = \sqrt{\frac{\varepsilon_{si} q}{2\phi_B} \frac{N_A N_D}{N_A + N_D}} \tag{2.39}$$

For a one-sided step junction, where one region has a much higher doping level than the other region, we can simplify the expression. For an n^+p junction of the NMOS device, the equation is

$$C_{j0} = \sqrt{\frac{\varepsilon_{si} q N_A}{2\phi_B}} \tag{2.40}$$

Since the junctions in a MOS transistor are normally reverse-biased, the junction capacitances are usually less than C_{j0}. This quantity is in units of fF/μm^2 and therefore it must be multiplied by the area of the junction to obtain the actual capacitance. The denominator of Equation (2.38) can be viewed as an adjustment factor due to the applied voltage. In our case, this is V_{BS} or V_{BD}, which is either zero or a negative quantity but never positive (otherwise we would have a forward-biased source/drain junction). Note that we are operating in the range of $V_J = 0$ to $-V_{DD}$ in normal mode for NMOS transistors, as indicated in Figure 2.19.

In order to compute the capacitance associated with the source and drain, we need to examine the transistor layout. Shown in Figure 2.20 is a very simple layout

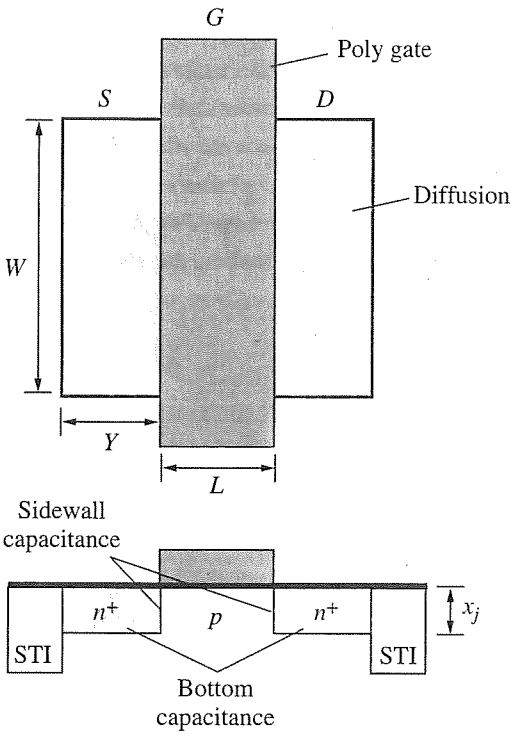


Figure 2.20
Junction capacitances from layout data.

of a MOS transistor in plan view and, below it, the corresponding cross-sectional view. In the layout view, there is a shaded polysilicon gate on top of a diffusion region. The intersection of polysilicon and diffusion forms a transistor. The gate (G), drain (D), and source (S) terminals are shown, along with the W, L, and Y dimensions of the transistor. The source/drain junction capacitances we seek are associated with the diffusion region, and so it is sometimes called the *diffusion* capacitance.

There are two types of junction capacitances that need to be computed for the NMOS device, namely the *bottom* capacitance and the *sidewall* capacitance. In the case of the bottom capacitance, we need to compute the area of the bottom region and the zero-bias value, C_{jb}, from the parameters for the n^+p junction. Looking at the layout, the area calculation for the bottom capacitance is as follows

$$A_b = WY$$

where W is the device width and Y is the diffusion extension from the poly gate edge. The junction is an n^+p junction so the numerator of Equation (2.38) would be equal to $C_{jb}A_b$.

For the sidewall capacitance, we need to take each edge of the junction and then multiply by the junction depth, x_j, to obtain the sidewall areas. A total of four faces need to be considered. Three edges are associated with the interface of the n^+ region and the shallow trench isolation,[12] and a fourth edge is due to an n^+p junction for the sidewall facing the channel.

Since the three edges that abut the shallow trench isolation have a small capacitance in modern technologies, the channel-facing sidewall edge is the only one we need to consider, as shown in Figure 2.20. Therefore, we compute the area for the channel-facing sidewall as

$$A_{sw} = Wx_j$$

The numerator of Equation (2.38) would be $C_{jsw}A_{sw}$, where the zero-bias junction capacitance for the channel sidewall is C_{jsw}. Now, the detailed junction capacitance equation can be stated

$$C_J = \frac{C_{jb}\,A_b}{\left(1 - \dfrac{V_J}{\phi_{Bb}}\right)^{mj}} + \frac{C_{jsw}\,A_{sw}}{\left(1 - \dfrac{V_J}{\phi_{Bsw}}\right)^{mjsw}} \tag{2.41}$$

where $mjsw$ and ϕ_{Bsw} are the capacitance terms for the channel-facing sidewall. This equation reflects the model used in circuit simulators, but we need to simplify the equation for use in hand calculations. First we would like to combine the two terms together. We will see in the next chapter that the channel-facing edge (i.e., the second term) has a very complex doping profile. However, if we consider it to

[12] Shallow-trench isolation will be described in more detail in Chapter 3.

be primarily an n^+p junction, then we can combine the two terms together. In fact, only a small error is made with this simplification

$$C_J = \frac{C_{jb}A_b + C_{jb}A_{sw}}{\left(1 - \frac{V_J}{\phi_B}\right)^{mj}} = \frac{C_{jb}(A_b + A_{sw})}{\left(1 - \frac{V_J}{\phi_B}\right)^{mj}} \tag{2.42}$$

Next, we would like to remove the voltage dependence of the term in the denominator of Equation (2.42) to produce a large-signal equivalent value. To do this, we can look back at Figure 2.19 and realize that the only region of interest is when $V_J \leq 0$, since the junctions are normally reverse-biased. If $V_J = 0$, then the denominator of Equation (2.42) is equal to one. If there is a reverse-bias of $-V_{DD}$, the denominator is between 1 and 2. One way to handle the voltage dependence is to define a new parameter, K_{eq}, which adjusts the zero-bias capacitance, C_{jb}, based on the expected bias voltage across the junction.

Since the voltage in digital circuits is expected to switch from low-to-high and vice-versa, we can use this fact to compute K_{eq}. To this end, we define an equivalent voltage-independent capacitance C_{eq} that requires the same change in charge as the nonlinear capacitance for a transition between two voltages V_1 and V_2 applied to the junction:

$$C_{eq} = \frac{Q_j(V_2) - Q_j(V_1)}{V_2 - V_1} = \frac{\Delta Q}{\Delta V}$$

where

$$\Delta Q = \int_{V_1}^{V_2} C(V)dV = \int_{V_1}^{V_2} C_{jb}\left(1 - \frac{V}{\phi_B}\right)^{-m} dV$$

Thus

$$C_{eq} = -\frac{C_{jb}\phi_B}{(V_2 - V_1)(1 - m)}\left[\left(1 - \frac{V_2}{\phi_B}\right)^{1-m} - \left(1 - \frac{V_1}{\phi_B}\right)^{1-m}\right]$$

We are now able to define K_{eq}, the dimensionless constant used to relate C_{eq} to C_{jb} for specified values of V_1 and V_2. In the case of an abrupt junction, $m = 1/2$, and therefore

$$K_{eq} = \frac{C_{eq}}{C_{jb}} = \frac{-2\phi_B^{1/2}}{V_2 - V_1}\left[(\phi_B - V_2)^{1/2} - (\phi_B - V_1)^{1/2}\right] \tag{2.43}$$

Fortunately, this complicated equation is required in only one case. Once computed, the simplified equation for hand calculations is as follows:

$$C_J = K_{eq}(C_{jb}WY + C_{jb}x_jW) = K_{eq}(C_{jb}Y + C_{jb}x_j)\,W = K_{eq}C_{jb}(Y + x_j)W \tag{2.44}$$

All the leading terms of the capacitance can be combined to produce a relatively straightforward equation for hand calculations:

$$C_J = K_{eq} C_{jb}(Y + x_j)W = C_j W \tag{2.45}$$

Note that if the voltage switches from one level to the other, $K_{eq} \cong 0.75$. However, if there is a fixed voltage across the junction then we do not need to use Equation (2.43) to determine K_{eq}. In NMOS devices, if a source or drain node is close to Gnd, then $V_{SB} = 0$ so we can set $K_{eq} = 1.0$. If a node is close to V_{DD} we should use the capacitance value at $V_J = -V_{DD}$. For this case, we should set $K_{eq} \cong 0.5$ since there is a large voltage drop across the junction. Similar considerations apply for PMOS devices with similar capacitance values.

One final note on junction capacitance: the channel junction capacitance, C_{jc}, shown in Figure 2.18 uses the same formulation as given in Equation (2.38). This term is associated mainly with the gate-to-bulk capacitance. There is a small portion of it that is charged from the source and drain, but it is negligible and usually ignored.

Example 2.12 **Junction Capacitance Calculations**

Problem:

(a) Find ϕ_B and C_{jb} for an n^+p junction diode with $N_D = 10^{20}$ cm^{-3} and $N_A = (3)10^{17}$ cm^{-3}.

Solution:

From Equation (2.37),

$$\phi_B = \frac{kT}{q} \ln \frac{N_A N_D}{n_i^2} = 0.026 \ln \left(\frac{3(10^{17})(10^{20})}{(1.45(10^{10}))^2} \right) = 1 \text{ V}$$

From Equation (2.39)

$$C_{jb} = \sqrt{\frac{\varepsilon_{si} q}{2\phi_B} \frac{N_A N_D}{N_A + N_D}} \approx \sqrt{\frac{\varepsilon_{si} q N_A}{2\phi_B}}$$

$$= \sqrt{\frac{11.7 * (8.85)(10^{-14}) * 1.6(10^{-19})(3)(10^{17})}{2(1.0)}} \approx 1.6 \frac{\text{fF}}{\mu\text{m}^2}$$

Problem:

(b) For a 0.13 μm process, $W = 400$ nm, $L = 100$ nm, $x_j = 50$ nm, and the diffusion extension is $Y = 300$ nm. Using the layout of Figure 2.20, find C_J in units of fF for and $V_J = 0$ and $V_J = -1.2$ V.

Solution:

For $V_J = 0$, the value is obtained by multiplying C_{jb} with $(Y+x_j)W$

$C_J = C_{jb}(Y + x_j)W = 1.6 \text{ fF}/\mu\text{m}^2 \times (0.3\mu\text{m} + 0.05\mu\text{m}) \times 0.4\mu\text{m} \approx 0.22 \text{ fF}$

For $V_J = -1.2$,

$$C_J = \frac{C_{jb}(Y + x_j)W}{(1 - V_J/\phi_B)^m}$$

$$= \frac{1.6\ \text{fF}/\mu\text{m} \times (0.3\ \mu\text{m} + 0.05\ \mu\text{m}) \times 0.4\ \mu\text{m}}{(1 + 1.2/1.0)^{1/2}} = 0.16\ \text{fF}$$

Problem:

(c) Find K_{eq} for $V_1 = -1.2$ V, $V_2 = 0$ and then compute the large-signal effective C_J.

Solution:

From Equation (2.43),

$$K_{eq} = \frac{-2(1)^{1/2}}{0 - (-1.2)}[(1 - 0)^{1/2} - (1 - (-1.2))^{1/2}] = 0.8$$

To compute C_J, we use

$$\therefore C_J = K_{eq}C_{jb}(Y + x_j)W$$

$$= 0.8 \times 1.6\ \text{fF}/\mu\text{m}^2 \times (0.3\mu\text{m} + 0.05\mu\text{m}) \times 0.4\mu\text{m} \approx 0.18\ \text{fF}$$

2.8.3 Overlap Capacitance

The overlap capacitance, C_{ol}, shown on both sides of the gate in Figure 2.17 is due to *lateral diffusion* and *fringing* components. This voltage-independent capacitance is connected from gate-to-drain and from gate-to-source. In the older technologies, there was significant diffusion of the heavily doped source and drain regions under the gate. This gave rise to the overlap capacitance, C_{ov}. However, over the years, the fringing capacitance, C_f, between the sidewall of the polysilicon and the surface of the drain and source has increased and must also be taken into account. It is difficult to separate out the components due to lateral diffusion and fringing so the combination of the two is referred to as C_{ol}.

$$C_{ol} = C_{ov} + C_f \tag{2.46}$$

The two components are shown in Figure 2.21. They should always be included in the capacitance calculation to obtain accurate results. We can try to estimate this overlap capacitance for hand calculations as follows. The fringing component can be approximated by the formulation

$$C_f = \frac{2\varepsilon_{ox}}{\pi} \ln\left(1 + \frac{T_{poly}}{t_{ox}}\right) \tag{2.47}$$

where T_{poly} is the thickness of the polysilicon material that sits on the oxide.

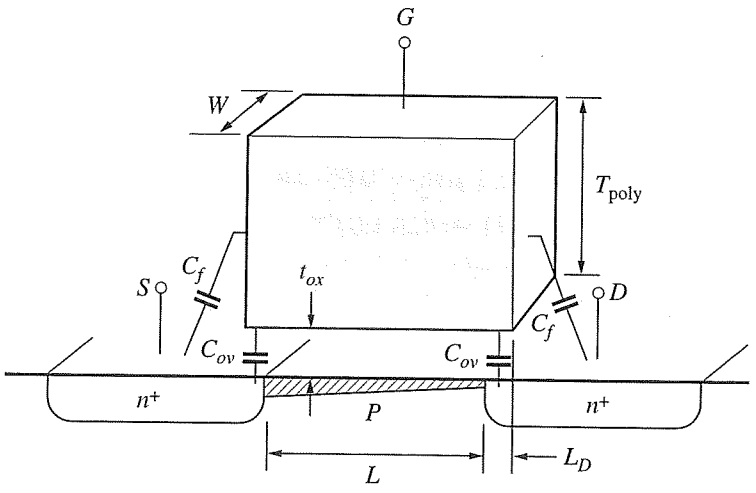


Figure 2.21
Overlap capacitance components.

The capacitance due to lateral diffusion is computed as

$$C_{ov} = C_{ox} \times L_D$$

where L_D is a lateral diffusion term illustrated in Figure 2.21.

Example 2.13 **Overlap Capacitance Calculations**

Problem:

Compute the overlap capacitance, C_{ol}, for a 0.13 μm technology with $T_{poly}/t_{ox} = 100$ and a lateral diffusion of 10 nm. Specify the solution in units of fF/μm of width.

Solution:

If we apply the ratio T_{poly}/t_{ox} to (2.47), then

$$C_f = \frac{2(4)(8.85 \times 10^{-14})}{3.14} \ln(1 + 100) \approx 0.1 \text{ fF}/\mu\text{m}$$

If $L_D = 10$ nm, then $C_{ov} = C_{ox}L_D \approx 0.15$ fF/μm. Therefore, for convenience we will use

$$C_{ol} = 0.1 \text{ fF}/\mu\text{m} + 0.15 \text{ fF}/\mu\text{m} = 0.25 \text{ fF}/\mu\text{m}$$

This value is multiplied by the width of the device to obtain the total overlap capacitance.

2.9 Summary

Operating voltages for the NMOS and PMOS devices:

NMOS Voltages	PMOS Voltages
$V_{TN} \geq 0$ V	$V_{TP} \leq 0$ V or $\lvert V_{TP}\rvert \geq 0$ V
$V_{GS} \geq 0$ V	$V_{GS} \leq 0$ V or $\lvert V_{GS}\rvert \geq 0$ V
$V_{DS} \geq 0$ V	$V_{DS} \leq 0$ V or $\lvert V_{DS}\rvert \geq 0$ V
$V_{BS} \leq 0$ V or $V_{SB} \geq 0$ V	$V_{BS} \geq 0$ V

The NMOS threshold voltage can be computed as follows (with appropriate changes for PMOS device):

$$V_T = V_{T0} + \gamma\left(\sqrt{|2\phi_F| + V_{SB}} - \sqrt{|2\phi_F|}\right)$$

$$V_{T0} = \phi_{GC} - 2\phi_F - \frac{Q_{B0}}{C_{ox}} - \frac{Q_{ox}}{C_{ox}} + \frac{Q_I}{C_{ox}}$$

$$\phi_F = \frac{kT}{q} \ln \frac{n_i}{p} \quad (p\text{-type substrate})$$

$$\gamma = \frac{1}{C_{ox}} \sqrt{2q\varepsilon_{si}N_A}$$

For long-channel devices, the current equations are as follows:

If $V_{GS} \geq V_T$

if $V_{DS} \geq (V_{GS} - V_T)$ saturation region $I_{DS} = \frac{k}{2}(V_{GS} - V_T)^2(1 + \lambda V_{DS})$

if $V_{DS} < (V_{GS} - V_T)$ linear region $I_{DS} = \frac{k}{2}[2(V_{GS} - V_T)V_{DS} - V_{DS}^2]$

If $V_{GS} < V_T$

subthreshold conduction region $I_{DS} = 0$

For the velocity saturated short-channel devices, use

If $V_{GS} < V_T$

if $V_{DS} \geq \frac{(V_{GS} - V_T)E_cL}{(V_{GS} - V_T) + E_cL}$ saturation region $I_{DS} = Wv_{sat}C_{ox}\frac{(V_{GS} - V_T)^2}{(V_{GS} - V_T) + E_cL}$

if $V_{DS} < \frac{(V_{GS} - V_T)E_cL}{(V_{GS} - V_T) + E_cL}$ linear region $I_{DS} = \frac{W}{L}\frac{\mu_e C_{ox}}{\left(1 + \frac{V_{DS}}{E_cL}\right)}\left(V_{GS} - V_T - \frac{V_{DS}}{2}\right)V_{DS}$

If $V_{GS} \geq V_T$

subthreshold conduction region $I_{DS} = I_s e^{\frac{q(V_{GS} - V_T - V_{offset})}{nkT}} \left(1 - e^{\frac{-qV_{DS}}{kT}}\right)$

Simple expressions for total MOS capacitances associated with thin oxide:

	Cutoff	Linear	Saturation
C_{GS}	0	$\frac{1}{2} C_{ox}WL$	$\frac{2}{3} C_{ox}WL$
C_{GD}	0	$\frac{1}{2} C_{ox}WL$	0
C_{GB}	$C_{ox}WL$	0	0

Equations for junction capacitances:

$$C_J = \frac{C_{j0}A}{\left(1 - \frac{V_J}{\phi_B}\right)^m}$$

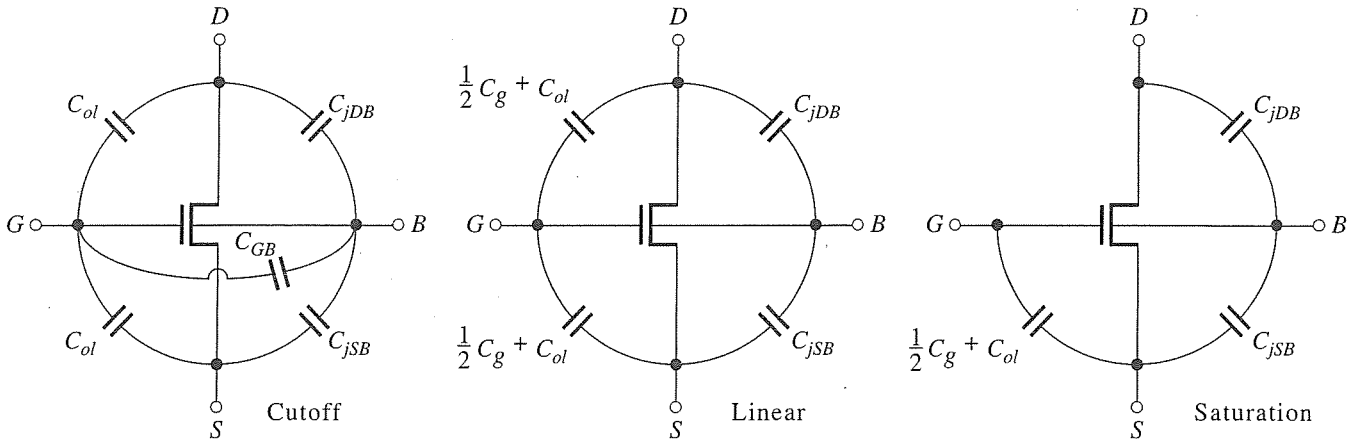
$$C_J = \frac{C_{jb}(A_b + A_{sw})}{\left(1 - \frac{V_J}{\phi_B}\right)^{mj}}$$

$$C_J = K_{eq}C_{jb}(Y + x_j)W = C_jW$$

where $C_j = K_{eq}C_{jb}(Y + x_j)$

Complete table of MOS transistor capacitances (including C_{ol} and C_{jc}) in fF/μm of width:

	Cutoff	Linear	Saturation
C_{gs}	C_{ol}	$C_{ol} + \frac{1}{2} C_g$	$C_{ol} + \frac{2}{3} C_g$
C_{gd}	C_{ol}	$C_{ol} + \frac{1}{2} C_g$	C_{ol}
C_{gb}	$1/C_g + 1/C_{jc} < C_{gb} < C_g$	0	0
C_{sb}	C_{jSB}	$C_{jSB} + \alpha_1 C_{jc}$	$C_{jSB} + \beta_1 C_{jc}$ (α, β small)
C_{db}	C_{jDB}	$C_{jDB} + \alpha_2 C_{jc}$	$C_{jDB} + \beta_2 C_{jc}$ (α, β small)



REFERENCES

1. A. S. Grove, *Physics and Technology of Semiconductor Devices,* Wiley, New York, 1967.
2. R. S. Muller and T. Kamins, *Device Electronics for Integrated Circuits,* 2nd ed., Wiley, New York, 1986.
3. S. M. Sze, *Physics of Semiconductor Devices,* 2nd ed., Wiley-Interscience, 1981.
4. K-Y Toh, P-K Ko, and R. G. Meyer, "An Engineering Model for Short-Channel MOS Devices," *IEEE Journal of Solid-State Circuits,* vol. 23, no. 4, August 1988.
5. T. Sakurai and A. R. Newton, "Alpha-Power Law MOSFET Model and Its Applications to CMOS Inverter Delay and Other Formulas," *IEEE Journal of Solid-State Circuits,* vol. 25, no. 2, pp. 584–594, April 1990.
6. J. Meyer, "MOS Models for Circuit Simulation," *RCA Review,* vol. 32, pp. 42–63, 1971.
7. D. Ward and R. Dutton, "A Charge-Oriented Model for MOS Transistor Capacitances," *IEEE Journal of Solid-State Circuits,* vol. SC-13, pp. 703–708, 1978.
8. N. Arora, *MOSFET Models for VLSI Circuit Simulation,* Springer-Verlag, 1993.

PROBLEMS

P2.1 Below is a table of process parameters for a generic 0.13 μm technology.

Table P2.1
0.13 μm Process technology parameters

Typical Technology Parameters	Symbol	NMOS	PMOS
Gate thickness	t_{ox} [Å]	22	22
Poly doping level	N_D [cm^{-3}]	3×10^{20}	—
Poly doping level	N_A [cm^{-3}]	—	3×10^{20}
Substrate doping level	N_A [cm^{-3}]	3×10^{17}	—
N-well doping level	N_D [cm^{-3}]	—	3×10^{17}
Number of surface state charges	N_{SS} [cm^{-2}]	6×10^{11}	6×10^{11}

Using the data in Table P2.1, find the following threshold voltage values:

(a) Compute V_{T0}, the unimplanted, zero-bias threshold voltages for both NMOS and PMOS devices. Assume that there is no charge in the oxide itself, but there does exist a sheet charge at the Si-SiO_2 interface (computed as qN_{SS}).

(b) Normally, NMOS gates are doped with donors (n^+) while PMOS gates are doped with acceptors (p^+).

(i) How would V_{T0} be affected if we doped the PMOS poly gate with donors rather than acceptors?

(ii) Calculate the new $V_{T0.}$

(c) We now want to adjust the threshold voltages of both the NMOS and PMOS devices so that we achieve the following threshold voltages: $V_{T0N} = 0.4$ V and $V_{T0P} = -0.4$ V. Calculate the threshold implant levels for the two cases in parts (a) and (b).

(d) Why do modern technologies have an n^+ poly gate for NMOS and a p^+ gate for PMOS devices?

P2.2 Find the effective mobility, (μ_e), of a PMOS transistor due to the vertical field if $t_{ox} = 22$ Å and $|V_{GS} - V_T| = 0.8$ V. Let $\theta = 4 \times 10^6$ V/cm and $\eta = 1.85$ in a 0.13 μm technology. Assume that $\mu_0 = 130$ cm^2/V-sec.

P2.3 This problem involves plotting several I-V characteristics of NMOS and PMOS transistors in a 0.13 μm technology. The 0.13 μm technology uses a 1.2 V power supply. The transistors are unit sized where $W =$ 100 nm, $L = 100$ nm.

(a) Plot I_{DS} versus V_{DS} as a function of $V_{GS} = 0, 0.4, 0.8, 1.2$ V for both NMOS and PMOS devices.

(b) Plot I_{DS} versus V_{GS} with $V_{DS} = 1.2$ V for the NMOS device. Does the quadratic model hold for 0.13 μm devices, or is it closer to linear with V_{GS}?

P2.4 Find the region of operation for each of the transistors in Figure P2.4. Assume $V_{T0} = 0.4$ V.

P2.5 In the circuits of Figure P2.5, determine the voltage across the capacitor after the circuit reaches steady state. Assume the capacitor is initially discharged ($V_x = 0$ V) and $V_{T0} = 0.4$ V.

P2.6 In the circuits of Figure P2.6, draw V_X versus V_{in} for $0 < V_{in} < 1.2$.

P2.7 Calculate the gate and junction capacitance of the transistor shown in Figure P2.7 if $L = 100$ nm, $W = 400$ nm, $Y = 300$ nm, $K_{eq} = 0.8$, $C_{jb} =$ 1.6 fF/μm^2, $C_{ox} = 1.6 \times 10^{-6}$ F/cm^2, and $x_j = 65$ nm.

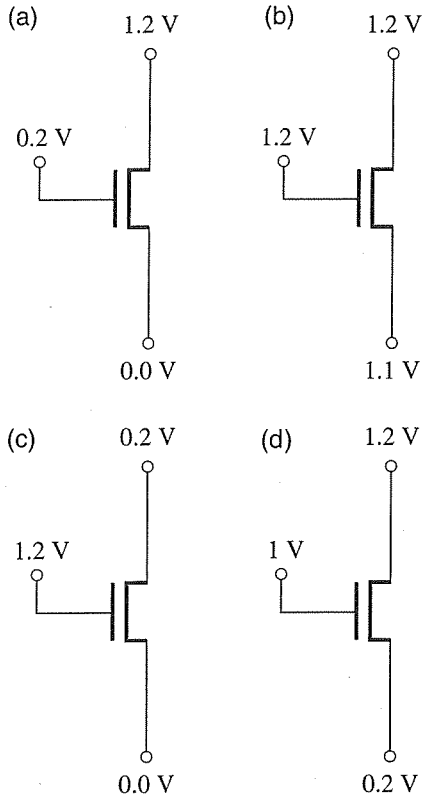


Figure P2.4

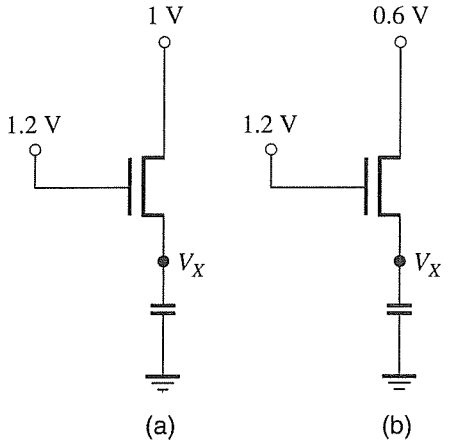


Figure P2.5

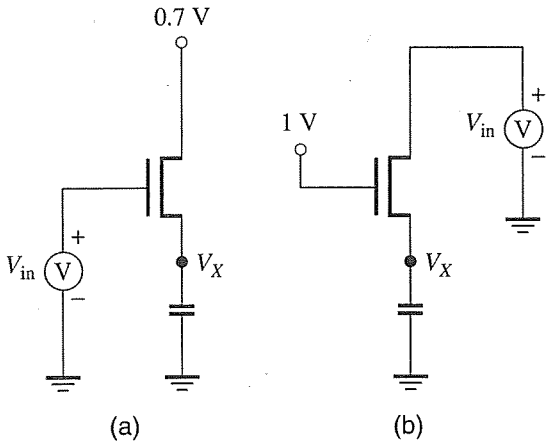


Figure P2.6

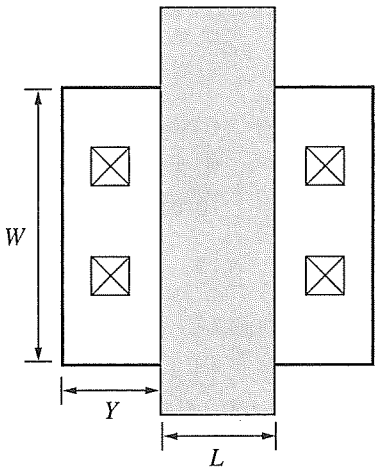


Figure P2.7

P2.8 Calculate the NMOS transistor current in each of the cases in Figure P2.8:

Use the following parameters if needed: $V_{T0} = 0.4$ V, $E_c = 6$ V/μm, $L = 100$ nm, $W = 400$ nm, $v_{sat} = 8{*}10^6$ cm/sec, $C_{ox} = 1.6 \times 10^{-6}$ F/cm^2, $\mu_e = 270$ cm^2/V-sec, $\gamma = 0.2$ (V$^{1/2}$), $2|\phi_F| = 0.88$ V

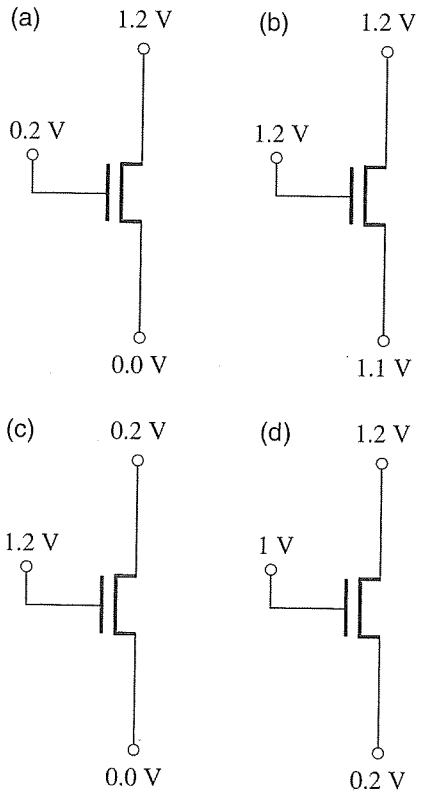


Figure P2.8

P2.9 Compare I_{DS} versus V_{DS} for the NMOS transistor if $V_{GS} = 1.2$ V, $V_{BS} = 0$, and $0 < V_{DS} < 1.2$ V for the following two cases. What is the main difference between the two curves?

(a) $L = 100$ nm $W = 200$ nm

(b) $L = 100$ nm $W = 400$ nm

Use the following parameters if needed:

$V_{T0} = 0.4$ V, $E_c = 6$ V/μm, $v_{sat} = 8 \times 10^6$ cm/sec, $C_{ox} = 1.6 \times 10^{-6}$ F/cm^2, $\mu_e = 270$ cm^2/V-sec, $\gamma = 0.2$ (V$^{1/2}$), $2|\phi_F| = 0.88$ V

P2.10 Compared to the velocity saturation model, the alpha-power law is a more empirically based model where the I-V curves are fitted to the data. The equation for the saturation region current in the alpha-power law model is as follows:

$$I_{DS} = K(W/L)(V_{GS} - V_T)^{\alpha}$$

(a) For the plots in Problem P2.3, compute the values for K and α for the NMOS and PMOS devices.

(b) For the linear region, develop a model where the current is linearly proportional to V_{DS} until the saturation region is reached. Derive a formula for V_{Dsat}.