CHAPTER

2

Basic MOS Device Physics

In studying the design of integrated circuits (ICs), one of two extreme approaches can be taken, (1) begin with quantum mechanics and understand solid-state physics, semiconductor device physics, device modeling, and finally the design of circuits; or (2) treat each semiconductor device as a black box whose behavior is described in terms of its terminal voltages and currents and design circuits with little attention to the internal operation of the device. Experience shows that neither approach is optimum. In the first case, the reader cannot see the relevance of all the physics to designing circuits, and in the second, he or she is constantly mystified by the contents of the black box.

In today's IC industry, a solid understanding of semiconductor devices is essential—more so in analog design than in digital design, because in the former, transistors are not considered to be simple switches, and many of their second-order effects directly impact the performance. Furthermore, as each new generation of IC technologies scales the devices, these effects become more significant. Since the designer must often decide which effects can be neglected in a given circuit, insight into device operation proves invaluable.

In this chapter, we study the physics of MOSFETs at an elementary level, covering the bare minimum that is necessary for basic analog design. The ultimate goal is still to develop a circuit model for each device by formulating its operation, but this is accomplished through a good understanding of the underlying principles. After studying many analog circuits in Chapters 3 through 14 and gaining motivation for a deeper understanding of devices, we return to the subject in Chapter 17 and deal with other aspects of MOS operation.

We begin our study with the structure of MOS transistors and derive their I/V characteristics. Next, we describe second-order effects such as body effect, channel-length modulation, and subthreshold conduction. We then identify the parasitic capacitances of MOSFETs, derive a small-signal model, and present a simple SPICE model. We assume that the reader is familiar with such basic concepts as doping, mobility, and *pn* junctions.

2.1 ■ General Considerations

2.1.1 MOSFET as a Switch

Before delving into the actual operation of the MOSFET, we consider a simplistic model of the device so as to gain a feel for what the transistor is expected to be and which aspects of its behavior are important.

Shown in Fig. 2.1 is the symbol for an *n*-type MOSFET, revealing three terminals: gate (G), source (S), and drain (D). The latter two are interchangeable because the device is symmetric. When operating

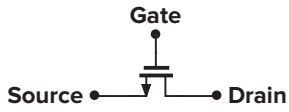


Figure 2.1 Simple view of a MOS device.

as a switch, the transistor "connects" the source and the drain together if the gate voltage, V_G, is "high" and isolates the source and the drain if V_G is "low."

Even with this simplified view, we must answer several questions. For what value of V_G does the device turn on? In other words, what is the "threshold" voltage? What is the resistance between S and D when the device is on (or off)? How does this resistance depend on the terminal voltages? Can we always model the path between S and D by a simple linear resistor? What limits the speed of the device?

While all of these questions arise at the circuit level, they can be answered only by analyzing the structure and physics of the transistor.

2.1.2 MOSFET Structure

Figure 2.2 shows a simplified structure of an n-type MOS (NMOS) device. Fabricated on a p-type substrate (also called the "bulk" or the "body"), the device consists of two heavily-doped n regions forming the source and drain terminals, a heavily-doped (conductive) piece of polysilicon[1] (simply called "poly") operating as the gate, and a thin layer of silicon dioxide (SiO_2) (simply called "oxide") insulating the gate from the substrate. The useful action of the device occurs in the substrate region under the gate oxide. Note that the structure is symmetric with respect to S and D.

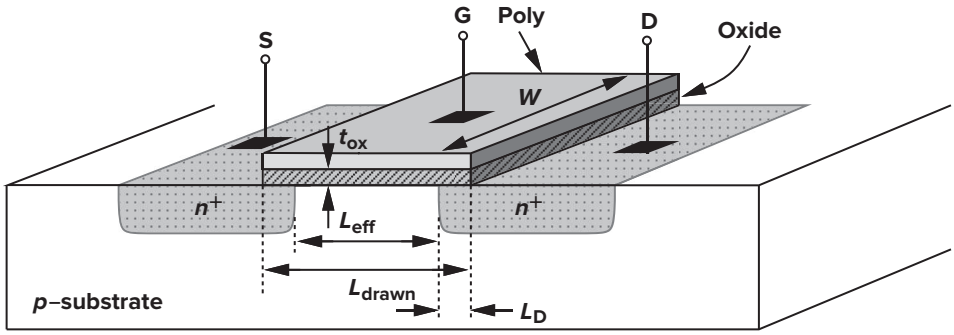


Figure 2.2 Structure of a MOS device.

The lateral dimension of the gate along the source-drain path is called the length, L, and that perpendicular to the length is called the width, W. Since the S/D junctions "side-diffuse" during fabrication, the actual distance between the source and the drain is slightly less than L. To avoid confusion, we write, $L_{eff} = L_{drawn} - 2L_D$, where L_{eff} is the "effective" length, L_{drawn} is the total length,[2] and L_D is the amount of side diffusion. As we will see later, L_{eff} and the gate oxide thickness, t_{ox}, play an important role in the performance of MOS circuits. Consequently, the principal thrust in MOS technology development is to reduce both of these dimensions from one generation to the next without degrading other parameters of the device. Typical values at the time of this writing are $L_{eff} \approx 10$ nm and $t_{ox} \approx 15$ Å. In the remainder of this book, we denote the effective length by L unless otherwise stated.

[1] Polysilicon is silicon in amorphous (non crystal) form. As explained in Chapter 18, when the gate silicon is grown on top of the oxide, it cannot form a crystal. The gate was originally made of metal [hence the term "metal-oxide-semiconductor" (MOS)] and is returning to metal in recent generations.

[2] The subscript "drawn" is used because this is the dimension that we draw in the layout of the transistor (Sec. 2.4.1).

If the MOS structure is symmetric, why do we call one n region the source and the other the drain? This becomes clear if the source is defined as the terminal that provides the charge carriers (electrons in the case of NMOS devices) and the drain as the terminal that collects them. Thus, as the voltages at the three terminals of the device vary, the source and the drain may exchange roles. These concepts are practiced in the problems at the end of the chapter.

We have thus far ignored the substrate on which the device is fabricated. In reality, the substrate potential greatly influences the device characteristics. That is, the MOSFET is a *four*-terminal device. Since in typical MOS operation, the S/D junction diodes must be reverse-biased, we assume that the substrate of NMOS transistors is connected to the most negative supply in the system. For example, if a circuit operates between zero and 1.2 volts, $V_{sub,NMOS} = 0$. The actual connection is usually provided through an ohmic p^+ region, as depicted in the side view of the device in Fig. 2.3.

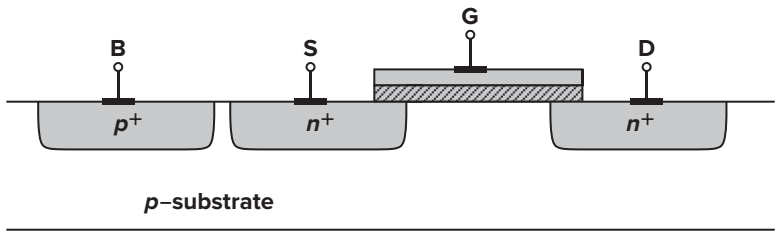


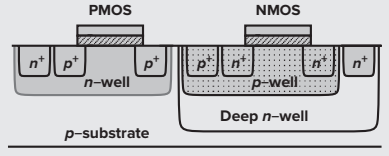
Figure 2.3 Substrate connection.

In complementary MOS (CMOS) technologies, both NMOS and PMOS transistors are available. From a simplistic viewpoint, the PMOS device is obtained by negating all of the doping types (including the substrate) [Fig. 2.4(a)], but in practice, NMOS and PMOS devices must be fabricated on the same wafer, i.e., the same substrate. For this reason, one device type can be placed in a "local substrate," usually called a "well." In today's CMOS processes, the PMOS device is fabricated in an n-well [Fig. 2.4(b)]. Note that the n-well must be connected to a potential such that the S/D junction diodes of the PMOS transistor remain reverse-biased under all conditions. In most circuits, the n-well is tied to the most positive supply voltage. For the sake of brevity, we sometimes call NMOS and PMOS devices "NFETs" and "PFETs," respectively.

Figure 2.4(b) indicates an interesting difference between NMOS and PMOS transistors: while all NFETs share the same substrate, each PFET can have an independent n-well. This flexibility of PFETs is exploited in some analog circuits.

Nanometer Design Notes

Some modern CMOS processes offer a "deep n-well," an n-well that contains an NMOS device and its p-type bulk. As shown below, the NMOS transistor's bulk is now localized and need not be tied to that of other NMOS devices. But the design incurs substantial area overhead because the deep n-well must extend beyond the p-well by a certain amount and must maintain a certain distance to the regular n-well.



2.1.3 MOS Symbols

The circuit symbols used to represent NMOS and PMOS transistors are shown in Fig. 2.5. The symbols in Fig. 2.5(a) contain all four terminals, with the substrate denoted by "B" (bulk) rather than "S" to avoid confusion with the source. The source of the PMOS device is positioned on top as a visual aid because it has a higher potential than its gate. Since in most circuits the bulk terminals of NMOS and PMOS devices are tied to ground and V_{DD}, respectively, we usually omit these connections in drawing [Fig. 2.5(b)]. In digital circuits, it is customary to use the "switch" symbols depicted in Fig. 2.5(c) for the two types, but we prefer those in Fig. 2.5(b) because the visual distinction between S and D proves helpful in understanding the operation of circuits.

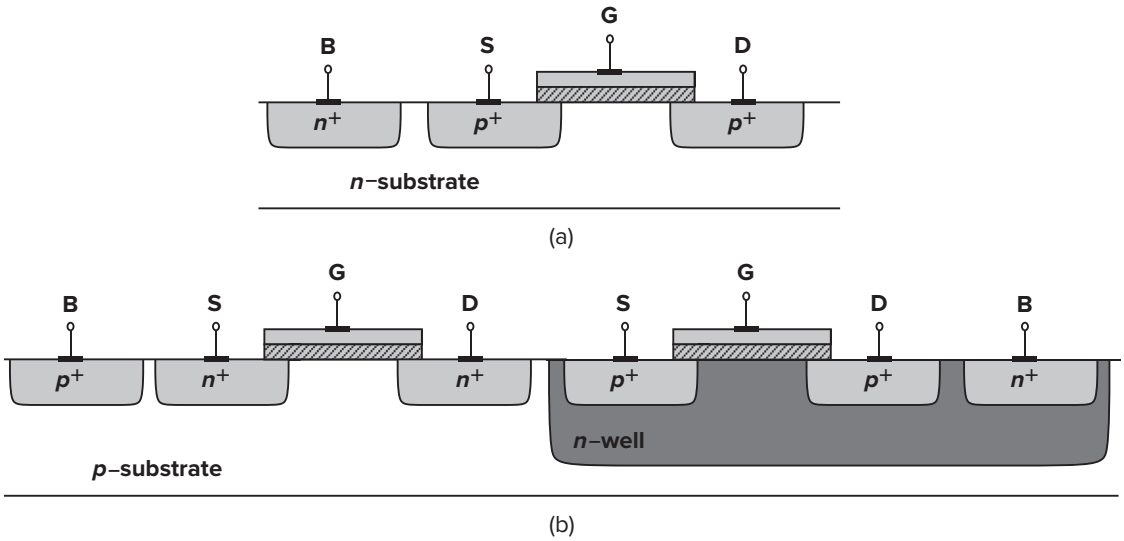


Figure 2.4 (a) Simple PMOS device; (b) PMOS inside an n-well.

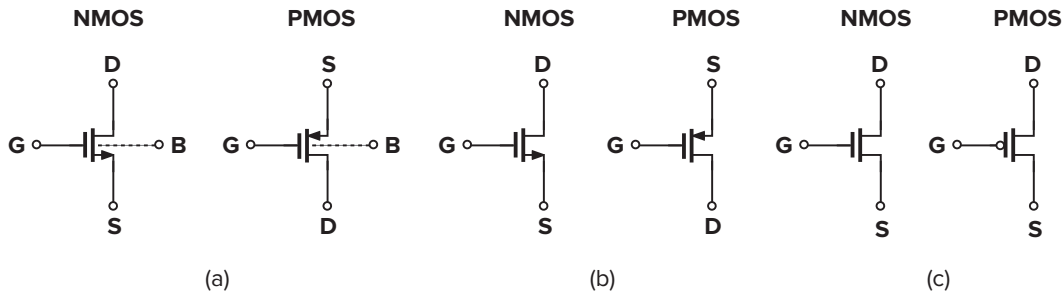


Figure 2.5 MOS symbols.

2.2 ■ MOS I/V Characteristics

In this section, we analyze the generation and transport of charge in MOSFETs as a function of the terminal voltages. Our objective is to derive equations for the I/V characteristics such that we can elevate our abstraction from device physics level to circuit level.

2.2.1 Threshold Voltage

Consider an NFET connected to external voltages as shown in Fig. 2.6(a). What happens as the gate voltage, V_G, increases from zero? Since the gate, the dielectric, and the substrate form a capacitor, as V_G becomes more positive, the holes in the p-substrate are repelled from the gate area, leaving negative ions behind so as to mirror the charge on the gate. In other words, a depletion region is created [Fig. 2.6(b)]. Under this condition, no current flows because no charge carriers are available.

As V_G increases, so do the width of the depletion region and the potential at the oxide-silicon interface. In a sense, the structure resembles a voltage divider consisting of two capacitors in series: the gate-oxide capacitor and the depletion-region capacitor [Fig. 2.6(c)]. When the interface potential reaches a sufficiently positive value, electrons flow from the source to the interface and eventually to the drain.

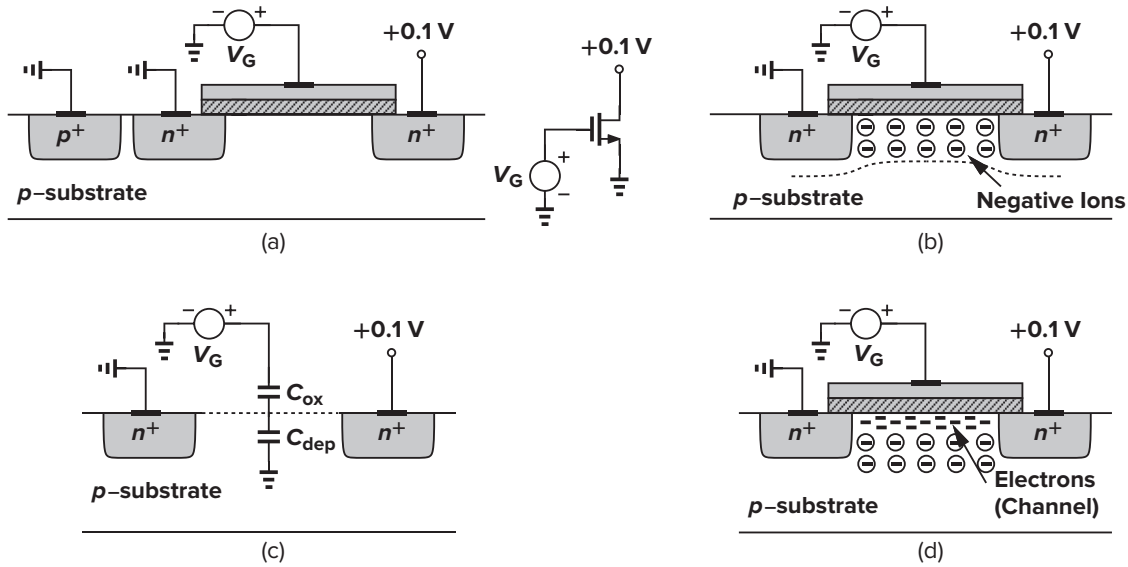


Figure 2.6 (a) A MOSFET driven by a gate voltage; (b) formation of depletion region; (c) onset of inversion; (d) formation of inversion layer.

Thus, a "channel" of charge carriers is formed under the gate oxide between S and D, and the transistor is "turned on." We say the interface is "inverted." For this reason, the channel is also called the "inversion layer." The value of V_G for which this occurs is called the "threshold voltage," V_{TH}. If V_G rises further, the charge in the depletion region remains relatively constant while the channel charge density continues to increase, providing a greater current from S to D.

In reality, the turn-on phenomenon is a gradual function of the gate voltage, making it difficult to define V_{TH} unambiguously. In semiconductor physics, the V_{TH} of an NFET is usually defined as the gate voltage for which the interface is "as much n-type as the substrate is p-type." It can be proved [1] that[3]

$$V_{TH} = \Phi_{MS} + 2\Phi_F + \frac{Q_{dep}}{C_{ox}} \tag{2.1}$$

where Φ_{MS} is the difference between the work functions of the polysilicon gate and the silicon substrate, $\Phi_F = (kT/q)\ln(N_{sub}/n_i)$, k is Boltzmann's constant, q is the electron charge, N_{sub} is the doping density of the substrate, n_i is the density of electrons in undoped silicon, Q_{dep} is the charge in the depletion region, and C_{ox} is the gate-oxide capacitance per unit area. From pn junction theory, $Q_{dep} = \sqrt{4q\epsilon_{si}|\Phi_F|N_{sub}}$, where ϵ_{si} denotes the dielectric constant of silicon. Since C_{ox} appears very frequently in device and circuit calculations, it is helpful to remember that for $t_{ox} \approx 20$ Å, $C_{ox} \approx 17.25$ fF/μm^2. The value of C_{ox} can then be scaled proportionally for other oxide thicknesses.

In practice, the "native" threshold value obtained from the above equation may not be suited to circuit design, e.g., $V_{TH} = 0$ and the device does not turn off for $V_G \geq 0$.[4] For this reason, the threshold voltage is typically adjusted by implantation of dopants into the channel area during device fabrication, in essence altering the doping level of the substrate near the oxide interface. For example, as shown in Fig. 2.7, if a thin sheet of p^+ is created, the gate voltage required to deplete this region increases.

[3]Charge trapping in the oxide is neglected here.

[4]Called a "depletion-mode" FET, such a device was used in old technologies. NFETs with a positive threshold are called "enhancement-mode" devices.

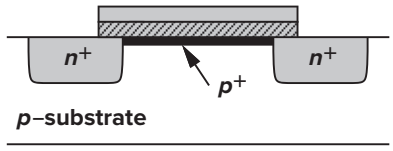


Figure 2.7 Implantation of $p+$ dopants to alter the threshold.

The above definition is not directly applicable to the *measurement* of V_{TH}. In Fig. 2.6(a), only the drain current can indicate whether the device is "on" or "off," failing to reveal at what V_{GS} the interface is as much n-type as the bulk is p-type. As a result, the calculation of V_{TH} from I/V measurements is somewhat ambiguous. We will return to this point later, but assume for now that the device turns on *abruptly* for $V_{GS} \geq V_{TH}$.

The turn-on phenomenon in a PMOS device is similar to that of NFETs, but with all the polarities reversed. As shown in Fig. 2.8, if the gate-source voltage becomes sufficiently *negative*, an inversion layer consisting of holes is formed at the oxide-silicon interface, providing a conduction path between the source and the drain. That is, the threshold voltage of a PMOS device is typically negative.

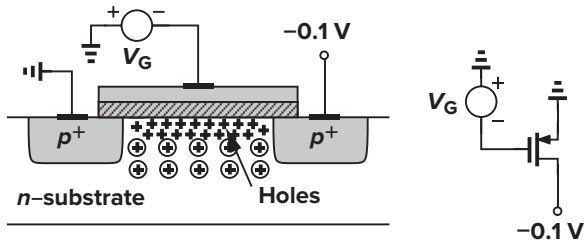


Figure 2.8 Formation of inversion layer in a PFET.

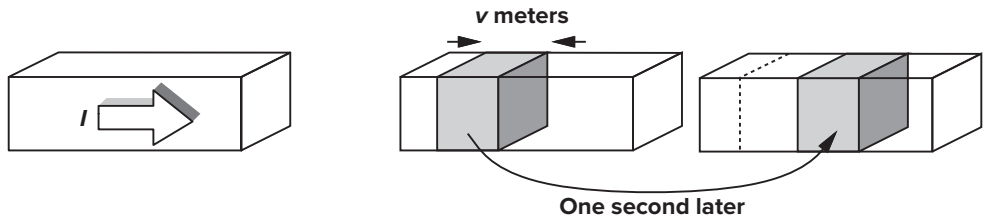
2.2.2 Derivation of I/V Characteristics

In order to obtain the relationship between the drain current of a MOSFET and its terminal voltages, we make two observations.

First, consider a semiconductor bar carrying a current I [Fig. 2.9(a)]. If the mobile charge density along the direction of current is Q_d coulombs per meter and the velocity of the charge is v meters per second, then

$$I = Q_d \cdot v \tag{2.2}$$

To understand why, we measure the total charge that passes through a cross section of the bar in unit time. With a velocity v, all of the charge enclosed in v meters of the bar must flow through the cross section in



(a) (b)

Figure 2.9 (a) A semiconductor bar carrying a current I; (b) snapshots of the carriers one second apart.

one second [Fig. 2.9(b)]. Since the charge density is Q_d, the total charge in v meters equals $Q_d \cdot v$. This lemma proves useful in analyzing semiconductor devices.

Second, to utilize the above lemma, we must determine the mobile charge density in a MOSFET. To this end, consider an NFET whose source and drain are connected to ground [Fig. 2.10(a)]. What is the charge density in the inversion layer? Since we assume that the onset of inversion occurs at $V_{GS} = V_{TH}$, the inversion charge density produced by the gate-oxide capacitance is proportional to $V_{GS} - V_{TH}$. For $V_{GS} \geq V_{TH}$, any charge placed on the gate must be mirrored by the charge in the channel, yielding a uniform channel charge density (charge per unit length along the source-drain path) equal to

$$Q_d = WC_{ox}(V_{GS} - V_{TH}) \tag{2.3}$$

where C_{ox} is multiplied by W to represent the total capacitance per unit length.

Now suppose, as depicted in Fig. 2.10(b), that the drain voltage is greater than zero. Since the channel potential varies from zero at the source to V_D at the drain, the local voltage *difference* between the gate and the channel varies from V_G (near the source) to $V_G - V_D$ (near the drain). Thus, the charge density at a point x along the channel can be written as

$$Q_d(x) = WC_{ox}[V_{GS} - V(x) - V_{TH}] \tag{2.4}$$

where $V(x)$ is the channel potential at x. From (2.2), the current is given by

$$I_D = -WC_{ox}[V_{GS} - V(x) - V_{TH}]v \tag{2.5}$$

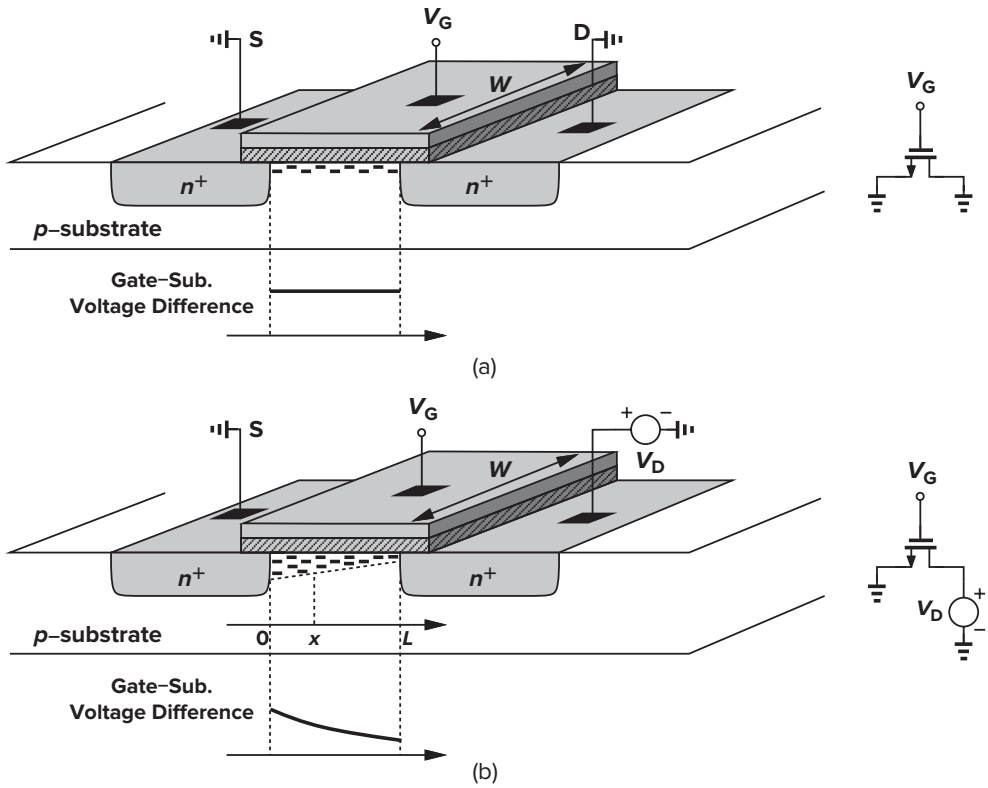


Figure 2.10 Channel charge with (a) equal source and drain voltages and (b) unequal source and drain voltages.

where the negative sign is inserted because the charge carriers are negative. Note that v denotes the velocity of the electrons in the channel. For semiconductors, $v = \mu E$, where μ is the mobility of charge carriers and E is the electric field. Noting that $E(x) = -dV/dx$ and representing the mobility of electrons by μ_n, we have

$$I_D = WC_{ox}[V_{GS} - V(x) - V_{TH}]\mu_n \frac{dV(x)}{dx} \tag{2.6}$$

subject to boundary conditions $V(0) = 0$ and $V(L) = V_{DS}$. While $V(x)$ can be easily found from this equation, the quantity of interest is in fact I_D. Multiplying both sides by dx and performing integration, we obtain

$$\int_{x=0}^{L} I_D dx = \int_{V=0}^{V_{DS}} WC_{ox}\mu_n[V_{GS} - V(x) - V_{TH}]dV \tag{2.7}$$

Since I_D is constant along the channel,

$$I_D = \mu_n C_{ox}\frac{W}{L}\left[(V_{GS} - V_{TH})V_{DS} - \frac{1}{2}V_{DS}^2\right] \tag{2.8}$$

Note that L is the effective channel length.

Figure 2.11 plots the parabolas given by (2.8) for different values of V_{GS}, indicating that the "current capability" of the device increases with V_{GS}. Calculating $\partial I_D/\partial V_{DS}$, the reader can show that the peak of each parabola occurs at $V_{DS} = V_{GS} - V_{TH}$ and the peak current is

$$I_{D,max} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2 \tag{2.9}$$

We call $V_{GS} - V_{TH}$ the "overdrive voltage" and W/L the "aspect ratio." If $V_{DS} \leq V_{GS} - V_{TH}$, we say the device operates in the "triode region."[5]

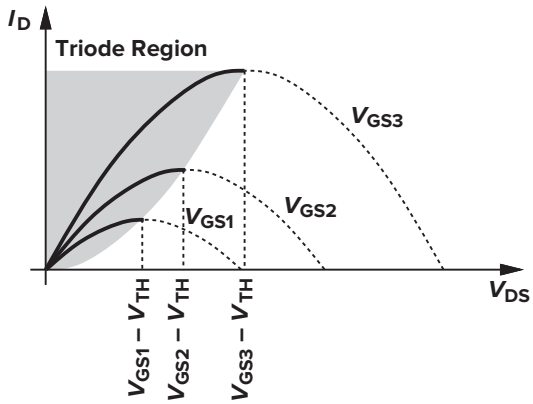


Figure 2.11 Drain current versus drain-source voltage in the triode region.

Equations (2.8) and (2.9) serve as our first step toward CMOS circuit design, describing the dependence of I_D upon the constant of the technology, $\mu_n C_{ox}$, the device dimensions, W and L, and the gate and drain potentials with respect to the source. Note that the integration in (2.7) assumes that μ_n and V_{TH} are independent of x and the gate and drain voltages, an approximation that we will revisit in Chapter 17.

[5] Also called the "linear region."

If in (2.8), $V_{DS} \ll 2(V_{GS} - V_{TH})$, we have

$$I_D \approx \mu_n C_{ox} \frac{W}{L}(V_{GS} - V_{TH})V_{DS} \tag{2.10}$$

that is, the drain current is a *linear* function of V_{DS}. This is also evident from the characteristics of Fig. 2.11 for small V_{DS}: as shown in Fig. 2.12, each parabola can be approximated by a straight line. The linear relationship implies that the path from the source to the drain can be represented by a linear resistor equal to

$$R_{on} = \frac{1}{\mu_n C_{ox} \dfrac{W}{L}(V_{GS} - V_{TH})} \tag{2.11}$$

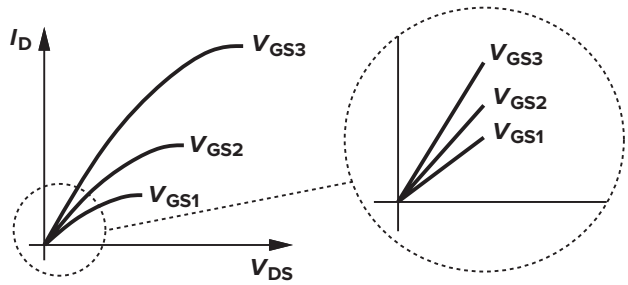


Figure 2.12 Linear operation in deep triode region.

A MOSFET can therefore operate as a resistor whose value is controlled by the overdrive voltage [so long as $V_{DS} \ll 2(V_{GS} - V_{TH})$]. This is conceptually illustrated in Fig. 2.13. Note that in contrast to bipolar transistors, a MOS device may be on even if it carries no current. With the condition $V_{DS} \ll 2(V_{GS} - V_{TH})$, we say the device operates in the deep triode region.

Figure 2.13 MOSFET as a controlled linear resistor.

▶ Example 2.1

For the arrangement in Fig. 2.14(a), plot the on-resistance of M_1 as a function of V_G. Assume that $\mu_n C_{ox} = 50\ \mu\text{A/V}^2$, $W/L = 10$, and $V_{TH} = 0.3$ V. Note that the drain terminal is open.

Solution

Since the drain terminal is open, $I_D = 0$ and $V_{DS} = 0$. Thus, if the device is on, it operates in the deep triode region. For $V_G < 1\text{ V} + V_{TH}$, M_1 is off and $R_{on} = \infty$. For $V_G > 1\text{ V} + V_{TH}$, we have

$$R_{on} = \frac{1}{50\ \mu\text{A/V}^2 \times 10(V_G - 1\text{ V} - 0.3\text{ V})} \tag{2.12}$$

The result is plotted in Fig. 2.14(b).

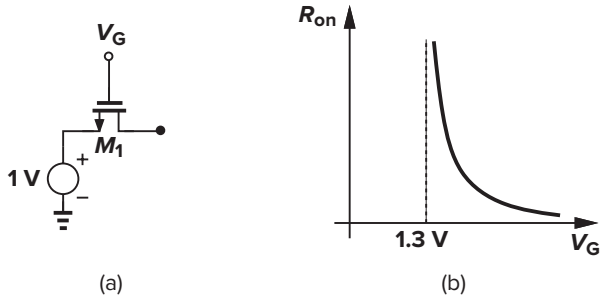


Figure 2.14

MOSFETs operating as controllable resistors play a crucial role in many analog circuits. For example, a voltage-controlled resistor can be used to adjust the frequency of the clock generator in a laptop computer if the system must go into a power saving mode. As studied in Chapter 13, MOSFETs also serve as switches.

What happens if the drain-source voltage in Fig. 2.11 exceeds $V_{GS} - V_{TH}$? In reality, the drain current does *not* follow the parabolic behavior for $V_{DS} > V_{GS} - V_{TH}$. In fact, as shown in Fig. 2.15, I_D becomes relatively constant, and we say the device operates in the "saturation" region.[6] To understand this phenomenon, recall from (2.4) that the local density of the inversion-layer charge is proportional to $V_{GS} - V(x) - V_{TH}$. Thus, if $V(x)$ approaches $V_{GS} - V_{TH}$, then $Q_d(x)$ drops to zero. In other words, as depicted in Fig. 2.16, if V_{DS} is slightly greater than $V_{GS} - V_{TH}$, then the inversion layer stops at $x \leq L$, and we say the channel is "pinched off." As V_{DS} increases further, the point at which Q_d equals zero gradually moves toward the source. Thus, at some point along the channel, the local potential difference between the gate and the oxide-silicon interface is not sufficient to support an inversion layer.

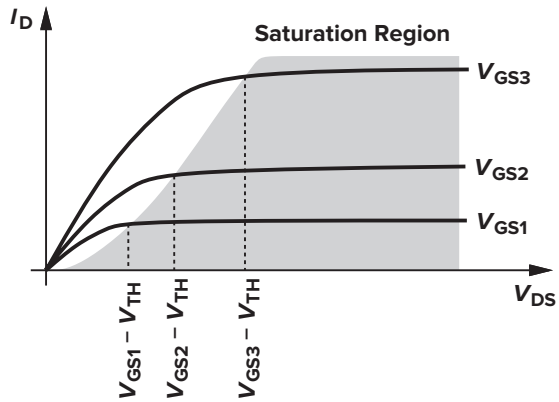


Figure 2.15 Saturation of drain current.

How does the device conduct current in the presence of pinch-off? As the electrons approach the pinch-off point (where $Q_d \to 0$), their velocity rises tremendously ($v = I/Q_d$). Upon passing the pinch-off point, the electrons simply shoot through the depletion region near the drain junction and arrive at the drain terminal.

[6]Note the difference between saturation in bipolar and MOS devices.

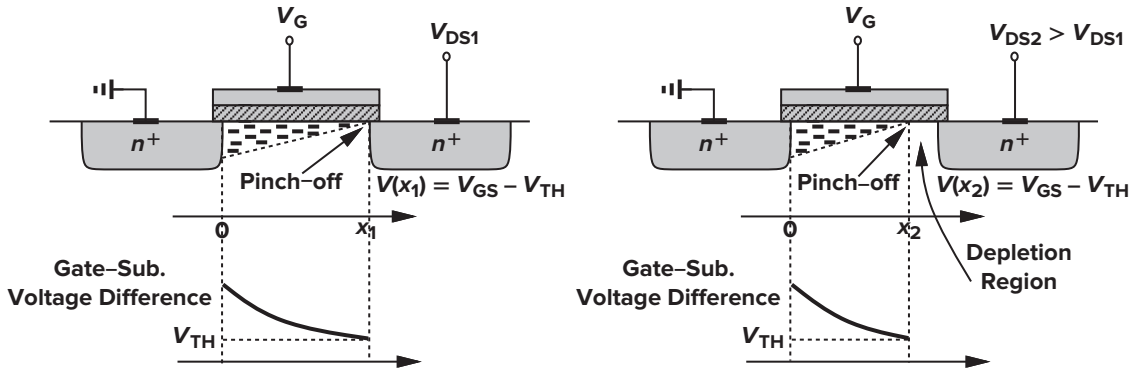


Figure 2.16 Pinch-off behavior.

With the above observations, we reexamine (2.7) for a saturated device. Since Q_d is the density of *mobile* charge, the integral on the left-hand side of (2.7) must be taken from $x = 0$ to $x = L'$, where L' is the point at which Q_d drops to zero (e.g., x_2 in Fig. 2.16), and that on the right from $V(x) = 0$ to $V(x) = V_{GS} - V_{TH}$. As a result,

$$I_D = \frac{1}{2}\mu_n C_{ox}\frac{W}{L'}(V_{GS} - V_{TH})^2 \tag{2.13}$$

indicating that I_D is relatively independent of V_{DS} if L' remains close to L. We say the device exhibits a "square-law" behavior. If I_D is known, then V_{GS} is obtained as

$$V_{GS} = \sqrt{\frac{2I_D}{\mu_n C_{ox}\dfrac{W}{L'}}} + V_{TH} \tag{2.14}$$

We must emphasize that for the transistor to remain in saturation (as is the case in many analog circuits), the drain-source voltage must be equal to or greater than the overdrive voltage. For this reason, some books write $V_{D,sat} = V_{GS} - V_{TH}$, where $V_{D,sat}$ denotes the minimum V_{DS} necessary for operation in saturation. As seen later in this book, if the signal swings at the drain or the gate cause V_{DS} to fall below $V_{GS} - V_{TH}$, then a number of undesirable effects occur. For this reason, the choice of the overdrive and hence $V_{D,sat}$ translates to a certain voltage "headroom" for the signal swings in the circuit: the larger the $V_{D,sat}$, the less headroom is available for the signals.

Equations (2.8) and (2.13) represent the "large-signal" behavior of NMOS devices; i.e., they can predict the drain current for arbitrary voltages applied to the gate, source, and drain (but only if the device is on). Since the nonlinear nature of these equations makes the analysis difficult, we often resort to linear approximations ("small-signal" models) so as to develop some understanding of a given circuit. This point becomes clear in Sec. 2.4.3.

For PMOS devices, Eqs. (2.8) and (2.13) are respectively written as

$$I_D = -\mu_p C_{ox}\frac{W}{L}\left[(V_{GS} - V_{TH})V_{DS} - \frac{1}{2}V_{DS}^2\right] \tag{2.15}$$

and

$$I_D = -\frac{1}{2}\mu_p C_{ox}\frac{W}{L'}(V_{GS} - V_{TH})^2 \tag{2.16}$$

The negative sign appears here because we assume that I_D flows from the drain to the source, whereas holes flow in the reverse direction. Note that V_{GS}, V_{DS}, V_{TH}, and $V_{GS} - V_{TH}$ are negative for a PMOS transistor that is turned on. Since the mobility of holes is about one-half the mobility of electrons, PMOS devices suffer from lower "current drive" capability.

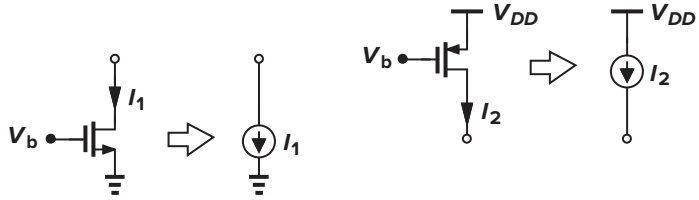


Figure 2.17 Saturated MOSFETs operating as current sources.

With L assumed constant, a saturated MOSFET can be used as a current source connected between the drain and the source (Fig. 2.17), an important component in analog design. Note that the NMOS current source injects current into ground and the PMOS current source draws current from V_{DD}. In other words, only one terminal of each current source is "floating." (It is difficult to design a current source that flows between two arbitrary nodes of a circuit.)

▶ Example 2.2

On a V_{DS}-V_{GS} plane, show the regions of operation of an NMOS transistor.

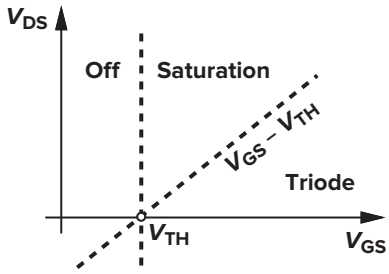


Figure 2.18 V_{DS}-V_{GS} plane showing regions of operation.

Solution

Since the value of V_{DS} with respect to $V_{GS} - V_{TH}$ determines the region of operation, we draw the line $V_{DS} = V_{GS} - V_{TH}$ in the plane, as shown in Fig. 2.18. If $V_{GS} > V_{TH}$, then the region above the line corresponds to saturation, and that below the line corresponds to the triode region. Note that for a given V_{DS}, the device eventually leaves saturation as V_{GS} increases. The minimum allowable V_{DS} for operation in saturation is also called $V_{D,sat}$. It is important to bear in mind that $V_{D,sat} = V_{GS} - V_{TH}$. ◀

The distinction between saturation and triode regions can be confusing, especially for PMOS devices. Intuitively, we note that the channel is pinched off if the difference between the gate and drain voltages is not sufficient to create an inversion layer. As depicted conceptually in Fig. 2.19, as $V_G - V_D$ of an NFET drops below V_{TH}, pinch-off occurs. Similarly, if $V_D - V_G$ of a PFET is not large enough ($< |V_{THP}|$), the device is saturated. Note that this view does not require knowledge of the source voltage. This means that we must know a priori which terminal operates as the drain. The drain is defined as the terminal with a higher (lower) voltage than the source for an NFET (PFET).

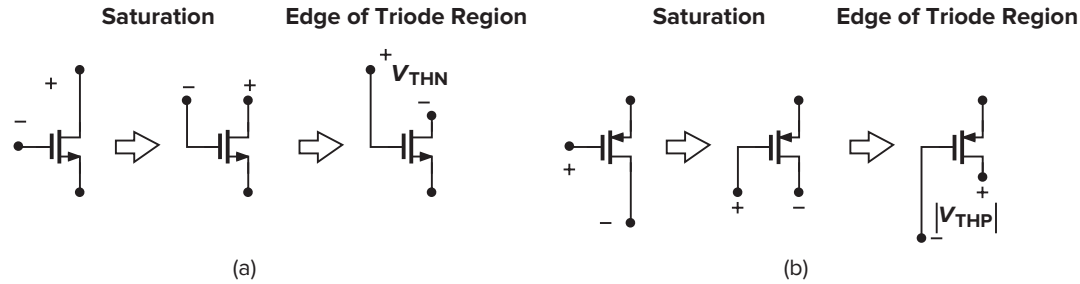


Figure 2.19 Conceptual visualization of saturation and triode regions.

2.2.3 MOS Transconductance

Since a MOSFET operating in saturation produces a current in response to its gate-source overdrive voltage, we may define a figure of merit that indicates how well a device converts a voltage to a current. More specifically, since in processing signals, we deal with the *changes* in voltages and currents, we define the figure of merit as the change in the drain current divided by the change in the gate-source voltage. Called the "transconductance" (and usually defined in the saturation region) and denoted by g_m, this quantity is expressed as

$$g_m = \frac{\partial I_D}{\partial V_{GS}}\bigg|_{VDS \text{ const.}} \tag{2.17}$$

$$= \mu_n C_{ox} \frac{W}{L}(V_{GS} - V_{TH}) \tag{2.18}$$

In a sense, g_m represents the sensitivity of the device: for a high g_m, a small change in V_{GS} results in a large change in I_D. We express g_m in $1/\Omega$ or in siemens (S); e.g., $g_m = 1/(100\,\Omega) = 0.01$ S. In analog design, we sometimes say a MOSFET operates as a "transconductor" or a "V/I converter" to indicate that it converts a voltage change to a current change. Interestingly, g_m in the saturation region is equal to the inverse of R_{on} in the deep triode region.

The reader can prove that g_m can also be expressed as

$$g_m = \sqrt{2\mu_n C_{ox} \frac{W}{L} I_D} \tag{2.19}$$

$$= \frac{2I_D}{V_{GS} - V_{TH}} \tag{2.20}$$

Plotted in Fig. 2.20, each of the above expressions proves useful in studying the behavior of g_m as a function of one parameter while other parameters remain constant. For example, (2.18) suggests that

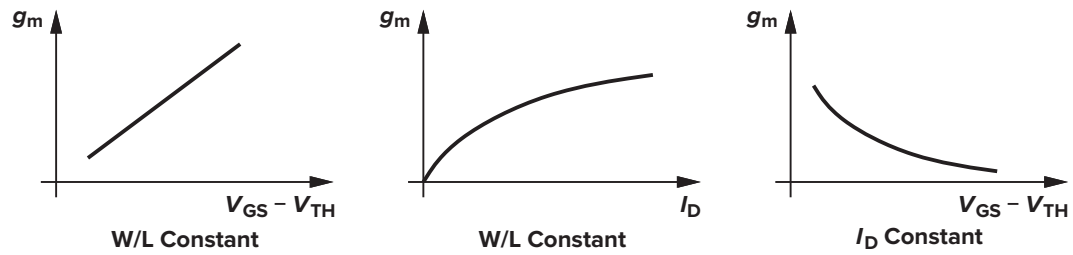


Figure 2.20 Approximate MOS transconductance as a function of overdrive and drain current.

g_m increases with the overdrive if W/L is constant, whereas (2.20) implies that g_m decreases with the overdrive if I_D is constant.

The I_D and $V_{GS} - V_{TH}$ terms in the above g_m equations are *bias* values. For example, a transistor with $W/L = 5\ \mu\text{m}/0.1\ \mu\text{m}$ and biased at $I_D = 0.5$ mA may exhibit a transconductance of $(1/200\ \Omega)$. If a signal is applied to the device, then I_D and $V_{GS} - V_{TH}$ and hence g_m *vary*, but in small-signal analysis, we assume that the signal amplitude is small enough that this variation is negligible.

Equation (2.19) implies that the transconductance can be raised arbitrarily if we increase W/L and keep I_D constant. This result is incorrect and will be revised in Sec. 2.3.

The concept of transconductance can also be applied to a device operating in the triode region, as illustrated in the following example.

▶ Example 2.3

For the arrangement shown in Fig. 2.21, plot the transconductance as a function of V_{DS}.

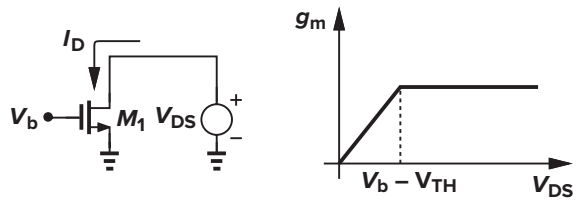


Figure 2.21

Solution

It is simpler to study g_m as V_{DS} decreases from infinity. So long as $V_{DS} \geq V_b - V_{TH}$, M_1 is in saturation, I_D is relatively constant, and, from (2.19), so is g_m. If the drain voltage falls below the gate voltage by more than one threshold, M_1 enters the triode region, and

$$g_m = \frac{\partial}{\partial V_{GS}} \left\{ \frac{1}{2}\mu_n C_{ox} \frac{W}{L} \left[2(V_{GS} - V_{TH})V_{DS} - V_{DS}^2 \right] \right\} \tag{2.21}$$

$$= \mu_n C_{ox} \frac{W}{L} V_{DS} \tag{2.22}$$

Thus, as plotted in Fig. 2.21, the transconductance drops in the triode region. For amplification, therefore, we usually employ MOSFETs in saturation. ◀

For a PFET, the transconductance in the saturation region is expressed as $g_m = -\mu_p C_{ox}(W/L)(V_{GS} - V_{TH}) = -2I_D/(V_{GS} - V_{TH}) = \sqrt{2\mu_p C_{ox}(W/L)I_D}$.

2.3 ■ Second-Order Effects

Our analysis of the MOS structure has thus far entailed various simplifying assumptions, some of which are not valid in many analog circuits. In this section, we describe three second-order effects that are essential in our subsequent circuit analyses. Other phenomena that appear in nanometer devices are studied in Chapter 17.

Body Effect In the analysis of Fig. 2.10, we tacitly assumed that the bulk and the source of the transistor were tied to ground. What happens if the bulk voltage of an NFET drops below the source voltage (Fig. 2.22)? Since the S and D junctions remain reverse-biased, we surmise that the device continues to operate properly, but some of its characteristics may change. To understand the effect, suppose

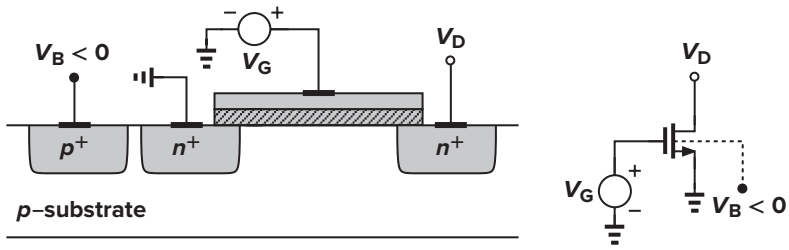


Figure 2.22 NMOS device with negative bulk voltage.

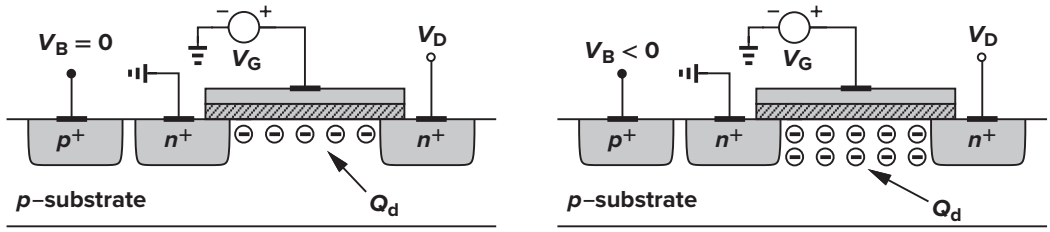


Figure 2.23 Variation of depletion region charge with bulk voltage.

$V_S = V_D = 0$, and V_G is somewhat less than V_{TH}, so that a depletion region is formed under the gate but no inversion layer exists. As V_B becomes more negative, more holes are attracted to the substrate connection, leaving a larger negative charge behind; i.e., as depicted in Fig. 2.23, the depletion region becomes wider. Now recall from Eq. (2.1) that the threshold voltage is a function of the total charge in the depletion region because the gate charge must mirror Q_d before an inversion layer is formed. Thus, as V_B drops and Q_d increases, V_{TH} also increases. This phenomenon is called the "body effect" or the "back-gate effect."

It can be proved that with body effect,

$$V_{TH} = V_{TH0} + \gamma \left(\sqrt{2\Phi_F + V_{SB}} - \sqrt{|2\Phi_F|} \right) \tag{2.23}$$

where V_{TH0} is given by (2.1), $\gamma = \sqrt{2q\epsilon_{si}N_{sub}}/C_{ox}$ denotes the body-effect coefficient, and V_{SB} is the source-bulk potential difference [1]. The value of γ typically lies in the range of 0.3 to 0.4 $V^{1/2}$.

▶ Example 2.4

In Fig. 2.24(a), plot the drain current if V_X varies from $-\infty$ to 0. Assume $V_{TH0} = 0.3$ V, $\gamma = 0.4$ $V^{1/2}$, and $2\Phi_F = 0.7$ V.

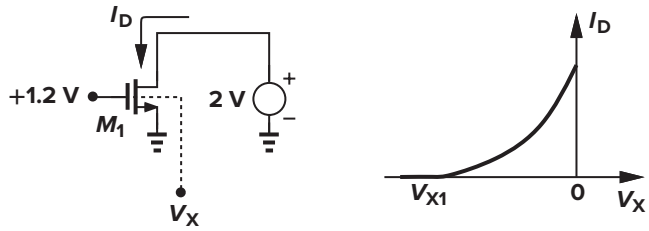


Figure 2.24

Solution

If V_X is sufficiently negative, the threshold voltage of M_1 exceeds 1.2 V and the device is off. That is,

$$1.2\ \text{V} = 0.3 + 0.4\left(\sqrt{0.7 - V_{X1}} - \sqrt{0.7}\right) \tag{2.24}$$

and hence $V_{X1} = -8.83$ V. For $V_{X1} < V_X < 0$, I_D increases according to

$$I_D = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}\left[V_{GS} - V_{TH0} - \gamma\left(\sqrt{2\Phi_F - V_X} - \sqrt{2\Phi_F}\right)\right]^2 \tag{2.25}$$

Fig. 2.24(b) shows the resulting behavior.

◀

For body effect to manifest itself, the bulk potential, V_{sub}, need not change: if the source voltage varies with respect to V_{sub}, the same phenomenon occurs. For example, consider the circuit in Fig. 2.25(a), first ignoring body effect. We note that as V_{in} varies, V_{out} closely follows the input because the drain current remains equal to I_1. In fact, we can write

$$I_1 = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{in} - V_{out} - V_{TH})^2 \tag{2.26}$$

concluding that $V_{in} - V_{out}$ is constant if I_1 is constant [Fig. 2.25(b)].

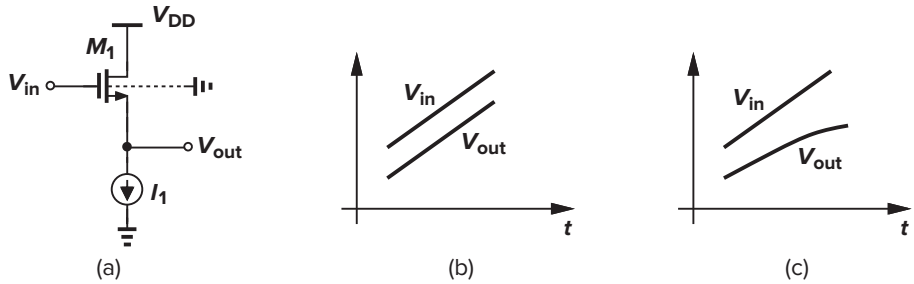


Figure 2.25 (a) A circuit in which the source-bulk voltage varies with input level; (b) input and output voltages with no body effect; (c) input and output voltages with body effect.

Now suppose that the substrate is tied to ground and body effect is significant. Then, as V_{in} and hence V_{out} become more positive, the potential difference between the source and the bulk increases, raising the value of V_{TH}. Equation (2.26) implies that $V_{in} - V_{out}$ must increase so as to maintain I_D constant [Fig. 2.25(c)].

Body effect is usually undesirable. The change in the threshold voltage, e.g., as in Fig. 2.25(a), often complicates the design of analog (and even digital) circuits. Device technologists balance N_{sub} and C_{ox} to obtain a reasonable value for γ.

▶ **Example 2.5**

Equation (2.23) suggests that if V_{SB} becomes *negative*, then V_{TH} *decreases*. Is this correct?

Solution

Yes, it is. If the bulk voltage of an NMOS device rises above its source voltage, V_{TH} falls below V_{TH0}. This observation proves useful in low-voltage design, where the performance of a circuit may suffer due to a high threshold voltage; one can bias the bulk to reduce V_{TH}. Unfortunately, this is not straightforward for NFETs because they typically share one substrate, but it can readily be applied to individual PFETs.

◀

Channel-Length Modulation In the analysis of channel pinch-off in Sec. 2.2, we noted that the actual length of the channel gradually decreases as the potential difference between the gate and the drain decreases. In other words, in (2.13), L' is in fact a function of V_{DS}. This effect is called "channel-length modulation." Writing $L' = L - \Delta L$, i.e., $1/L' \approx (1 + \Delta L/L)/L$, and assuming a first-order relationship between $\Delta L/L$ and V_{DS}, such as $\Delta L/L = \lambda V_{DS}$, we have, in saturation,

$$I_D \approx \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2(1 + \lambda V_{DS}) \tag{2.27}$$

where λ is the "channel-length modulation coefficient." Illustrated in Fig. 2.26, this phenomenon results in a nonzero slope in the I_D/V_{DS} characteristic and hence a nonideal current source between D and S in saturation. The parameter λ represents the *relative* variation in length for a given increment in V_{DS}. Thus, for longer channels, λ is smaller.

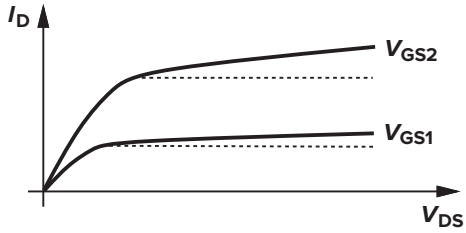


Figure 2.26 Finite saturation region slope resulting from channel-length modulation.

▶ **Example 2.6**

Is there channel-length modulation in the triode region?

Solution

No, there is not. In the triode region, the channel continuously stretches from the source to the drain, experiencing no pinch-off. Thus, the drain voltage does not modulate the length of the channel.

The reader may then observe a discontinuity in the equations as the device goes from the triode region to saturation:

$$I_{D,tri} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}\left[2(V_{GS} - V_{TH})V_{DS} - V_{DS}^2\right] \tag{2.28}$$

$$I_{D,sat} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2(1 + \lambda V_{DS}) \tag{2.29}$$

The former yields $(1/2)\mu_n C_{ox} W/L(V_{GS} - V_{TH})^2$ at the edge of the triode region, whereas the latter exhibits an additional factor of $1+\lambda V_{DS}$. This discrepancy is removed in more complex models of MOSFETs (Chapter 17). ◀

With channel-length modulation, some of the expressions derived for g_m must be modified. Equations (2.18) and (2.19) are respectively rewritten as

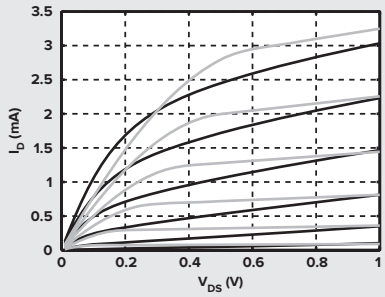
$$g_m = \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})(1 + \lambda V_{DS}) \tag{2.30}$$

$$= \sqrt{2\mu_n C_{ox}(W/L)I_D(1 + \lambda V_{DS})} \tag{2.31}$$

while Eq. (2.20) remains unchanged.

Nanometer Design Notes

Nanometer transistors suffer from various imperfections and markedly depart from square-law behavior. Shown below are the actual I-V characteristics of an NFET with $W/L = 5\ \mu$m/40 nm for $V_{GS} = 0.3$ V $\cdots$ 0.8 V. Also plotted are the characteristics of a square-law device of the same dimensions. Despite our best efforts to match the latter device to the former, we still observe significant differences.



▶ **Example 2.7**

Keeping all other parameters constant, plot the I_D/V_{DS} characteristic of a MOSFET for $L = L_1$ and $L = 2L_1$.

Solution

Writing

$$I_D = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2(1 + \lambda V_{DS}) \tag{2.32}$$

and $\lambda \propto 1/L$, we note that if the length is doubled, the slope of I_D vs. V_{DS} is divided by *four* because $\partial I_D/\partial V_{DS} \propto \lambda/L \propto 1/L^2$ (Fig. 2.27). (This is true only if $V_{GS} - V_{TH}$ is constant.) For a given gate-source overdrive, a larger L gives a more ideal current source while degrading the current capability of the device. Thus, W may need to be increased proportionally. In fact, if we double W to restore I_D to its original value, the slope also doubles. In other words, for a required drain current and a given overdrive, doubling the length reduces the slope by a factor of 2.

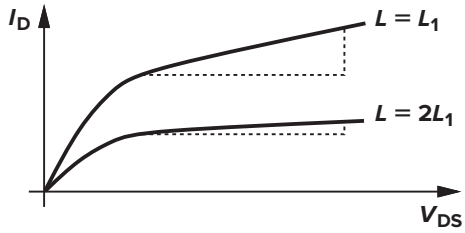


Figure 2.27 Effect of doubling channel length.

◀

The linear approximation $\Delta L/L \propto V_{DS}$ becomes less accurate in short-channel transistors, resulting in a *variable* slope in the saturated I_D/V_{DS} characteristics. We return to this issue in Chapter 17.

The dependence of I_D upon V_{DS} in saturation may suggest that the bias current of a MOSFET can be defined by the proper choice of the drain-source voltage, allowing freedom in the choice of $V_{GS} - V_{TH}$. However, since the dependence on V_{DS} is much weaker, the drain-source voltage is not used to set the current. That is, we always consider $V_{GS} - V_{TH}$ as the current-defining parameter. The effect of V_{DS} on I_D is usually considered an *error*, and it is studied in Chapter 5.

Subthreshold Conduction In our analysis of the MOSFET, we have assumed that the device turns off abruptly as V_{GS} drops below V_{TH}. In reality, for $V_{GS} \approx V_{TH}$, a "weak" inversion layer still exists and some current flows from D to S. Even for $V_{GS} < V_{TH}$, I_D is finite, but it exhibits an *exponential* dependence on V_{GS} [2, 3]. Called "subthreshold conduction," this effect can be formulated for V_{DS} greater than roughly 100 mV as

$$I_D = I_0 \exp\frac{V_{GS}}{\xi V_T} \tag{2.33}$$

where I_0 is proportional to W/L, $\xi > 1$ is a nonideality factor, and $V_T = kT/q$. We also say the device operates in "weak inversion." (Similarly, for $V_{GS} > V_{TH}$, we say the device operates in "strong inversion.") Except for ξ, (2.33) is similar to the exponential I_C/V_{BE} relationship of a bipolar transistor. The key point here is that as V_{GS} falls below V_{TH}, the drain current drops at a finite rate. With typical values of ξ, at room temperature V_{GS} must decrease by approximately 80 mV for I_D to decrease by one decade (Fig. 2.28). For example, if a threshold of 0.3 V is chosen in a process to allow low-voltage operation, then when V_{GS} is reduced to zero, the drain current decreases by only a factor of $10^{0.3\text{ V}/80\text{ mV}} = 10^{3.75} \approx 5.62 \times 10^3$. For example, if the transistor carries about 1 μA for $V_{GS} = V_{TH}$ and we have 100 million such devices, then

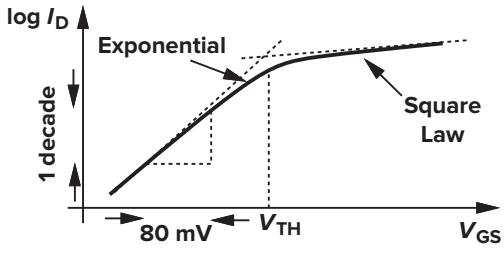


Figure 2.28 MOS subthreshold characteristics.

they draw 18 mA when they are nominally off. Especially problematic in large circuits such as memories, subthreshold conduction can result in significant power dissipation (or loss of analog information).

If a MOS device conducts for $V_{GS} < V_{TH}$, then how do we define the threshold voltage? Indeed, numerous definitions have been proposed. One possibility is to extrapolate, on a logarithmic vertical scale, the weak inversion and strong inversion characteristics and consider their intercept voltage as the threshold (Fig. 2.28).

We now reexamine Eq. (2.19) for the transconductance of a MOS device operating in the subthreshold region. Is it possible to achieve an arbitrarily high transconductance by increasing W while maintaining I_D constant? Is it possible to obtain a *higher* transconductance than that of a bipolar transistor (I_C/V_T) biased at the same current? Equation (2.19) was derived from the square-law characteristic $I_D = (1/2)\mu_n C_{ox}(W/L)(V_{GS} - V_{TH})^2$. However, if W increases while I_D remains constant, then $V_{GS} \to V_{TH}$ and the device enters the subthreshold region. As a result, the transconductance is calculated from (2.33) to be $g_m = I_D/(\xi V_T)$, revealing that MOSFETs are still inferior to bipolar transistors in this respect.

At what overdrive voltage can we say the transistor goes from strong inversion to weak inversion? While somewhat arbitrary, this transition point can be defined as the overdrive voltage, $(V_{GS} - V_{TH})_1$, at which the corresponding transconductances would become equal for the same drain current:

$$\frac{I_D}{\xi V_T} = \frac{2I_D}{(V_{GS} - V_{TH})_1} \tag{2.34}$$

and hence

$$(V_{GS} - V_{TH})_1 = 2\xi V_T \tag{2.35}$$

For $\xi \approx 1.5$, this amounts to about 80 mV.

The exponential dependence of I_D upon V_{GS} in subthreshold operation may suggest the use of MOS devices in this regime so as to achieve a higher gain. However, since such conditions are met only by a large device width or low drain current, the speed of subthreshold circuits is severely limited.

▶ Example 2.8

Examine the behavior of a MOSFET as the drain "current density," I_D/W, varies.

Solution

For a given drain current and device width, how do we determine the region of operation? We must consider the equations for both strong and weak inversion:

$$I_D = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2 \tag{2.36}$$

$$I_D = \alpha\frac{W}{L}\exp\frac{V_{GS}}{\xi V_T} \tag{2.37}$$

where channel-length modulation is neglected and I_0 in Eq. (2.33) has been expressed as a proportionality factor, α, multiplied by W/L. What happens if the device is in strong inversion and we continue to reduce I_D while W/L is constant? Can V_{GS} simply approach V_{TH} to yield an arbitrarily small value for $(V_{GS} - V_{TH})^2$? Why does the square-law equation not hold as V_{GS} approaches V_{TH}?

To answer these questions, we return to the plot of Fig. 2.28 and observe that only currents beyond a certain level can be supported in strong inversion. In other words, for a given current and W/L, we must obtain V_{GS} from both the square-law and exponential equations and select the lower value:

$$V_{GS} = \sqrt{\frac{2I_D}{\mu_n C_{ox} W/L}} + V_{TH} \tag{2.38}$$

$$V_{GS} = \xi V_T \ln \frac{I_D}{\alpha W/L} \tag{2.39}$$

If I_D remains constant and W increases, V_{GS} falls and the device goes from strong inversion to weak inversion. ◀

Voltage Limitations A MOSFET experiences various undesirable effects if its terminal voltage differences exceed certain limits (if the device is "stressed"). At high gate-source voltages, the gate oxide breaks down irreversibly, damaging the transistor. In short-channel devices, an excessively large drain-source voltage widens the depletion region around the drain so much that it touches that around the source, creating a very large drain current. (This effect is called "punchthrough.") Even without breakdown, MOSFETs' characteristics can change permanently if the terminal voltage differences exceed a specified value. Such effects are described in Chapter 17.

2.4 ■ MOS Device Models

2.4.1 MOS Device Layout

For the developments in subsequent sections, it is beneficial to have some understanding of the layout of a MOSFET. We describe only a simple view here, deferring the fabrication details and structural subtleties to Chapters 18 and 19.

The layout of a MOSFET is determined by both the electrical properties required of the device in the circuit and the "design rules" imposed by the technology. For example, W/L is chosen to set the transconductance or other circuit parameters while the minimum L is dictated by the process. In addition to the gate, the source and drain areas must be defined properly as well.

Shown in Fig. 2.29 are the "bird's-eye view" and the top view of a MOSFET. The gate polysilicon and the source and drain terminals must be tied to metal (aluminum) wires that serve as interconnects with low resistance and capacitance. To accomplish this, one or more "contact windows" must be opened in each region, filled with metal, and connected to the upper metal wires. Note that the gate poly extends beyond the channel area by some amount to ensure reliable definition of the "edge" of the transistor.

The source and drain junctions play an important role in the performance. To minimize the capacitance of S and D, the total area of each junction must be minimized. We see from Fig. 2.29 that one dimension of the junctions is equal to W. The other dimension must be large enough to accommodate the contact windows and is specified by the technology design rules.[7]

[7] This dimension is typically three to four times the minimum allowable channel length.