CHAPTER 5

MOS Field-Effect Transistors (MOSFETs)

IN THIS CHAPTER YOU WILL LEARN

1. The physical structure of the MOS transistor and how it works.
2. How the voltage between two terminals of the transistor controls the current that flows through the third terminal, and the equations that describe these current–voltage characteristics.
3. How to analyze and design circuits that contain MOS transistors, resistors, and dc sources.
4. How the transistor can be used to make an amplifier, and how it can be used as a switch in digital circuits.
5. How to obtain linear amplification from the fundamentally nonlinear MOS transistor.
6. The three basic ways for connecting a MOSFET to construct amplifiers with different properties.
7. Practical circuits for MOS–transistor amplifiers that can be constructed using discrete components.

Introduction

Having studied the junction diode, which is the most basic two-terminal semiconductor device, we now turn our attention to three-terminal semiconductor devices. Three-terminal devices are far more useful than two-terminal ones because they can be used in a multitude of applications, ranging from signal amplification to digital logic and memory. The basic principle involved is the use of the voltage between two terminals to control the current flowing in the third terminal. In this way a three-terminal device can be used to realize a controlled source, which as we have learned in Chapter 1 is the basis for amplifier design. Also, in the extreme, the control signal can be used to cause the current in the third terminal to change from zero to a large value, thus allowing the device to act as a switch. As we shall see in Chapter 13, the switch is the basis for the realization of the logic inverter, the basic element of digital circuits.

There are two major types of three-terminal semiconductor devices: the metal-oxide-semiconductor field-effect transistor (MOSFET), which is studied in this chapter, and the

bipolar junction transistor (BJT), which we shall study in Chapter 6. Although each of the two transistor types offers unique features and areas of application, the MOSFET has become by far the most widely used electronic device, especially in the design of integrated circuits (ICs), which are entire circuits fabricated on a single silicon chip.

Compared to BJTs, MOSFETs can be made quite small (i.e., requiring a small area on the silicon IC chip), and their manufacturing process is relatively simple (see Appendix A). Also, their operation requires comparatively little power. Furthermore, circuit designers have found ingenious ways to implement digital and analog functions utilizing MOSFETs almost exclusively (i.e., with very few or no resistors). All of these properties have made it possible to pack large numbers of MOSFETs (as many as 2 billion!) on a single IC chip to implement very sophisticated, very-large-scale-integrated (VLSI) digital circuits such as those for memory and microprocessors. Analog circuits such as amplifiers and filters can also be implemented in MOS technology, albeit in smaller, less-dense chips. Also, both analog and digital functions are increasingly being implemented on the same IC chip, in what is known as mixed-signal design.

The objective of this chapter is to develop in the reader a high degree of familiarity with the MOSFET: its physical structure and operation, terminal characteristics, circuit models, and basic circuit applications. Although discrete MOS transistors exist, and the material studied in this chapter will enable the reader to design discrete MOS circuits, our study of the MOSFET is strongly influenced by the fact that most of its applications are in integrated-circuit design. The design of IC analog and digital MOS circuits occupies a large proportion of the remainder of this book.

5.1 Device Structure and Physical Operation

The enhancement-type MOSFET is the most widely used field-effect transistor. Except for the last section, this chapter is devoted to the study of the enhancement-type MOSFET. We begin in this section by learning about its structure and physical operation. This will lead to the current–voltage characteristics of the device, studied in the next section.

5.1.1 Device Structure

Figure 5.1, shows the physical structure of the *n*-channel enhancement-type MOSFET. The meaning of the names "enhancement" and "*n*-channel" will become apparent shortly. The transistor is fabricated on a *p*-type substrate, which is a single-crystal silicon wafer that provides physical support for the device (and for the entire circuit in the case of an integrated circuit). Two heavily doped *n*-type regions, indicated in the figure as the n^+ **source**[1] and the n^+ **drain** regions, are created in the substrate. A thin layer of silicon dioxide (SiO_2) of thickness t_{ox} (typically 1 to 10 nm),[2] which is an excellent electrical insulator, is grown on the surface of the substrate, covering the area between the source and drain regions. Metal is deposited on top of the oxide layer to form the **gate electrode** of the device. Metal contacts are also made to the source region, the drain region, and the substrate, also known as the

[1]The notation n^+ indicates heavily doped *n*-type silicon. Conversely, n^- is used to denote lightly doped *n*-type silicon. Similar notation applies for *p*-type silicon.

[2]A nanometer (nm) is 10^{-9} m or 0.001 μm. A micrometer (μm), or micron, is 10^{-6} m. Sometimes the oxide thickness is expressed in angstroms. An angstrom (Å) is 10^{-1} nm, or 10^{-10} m.

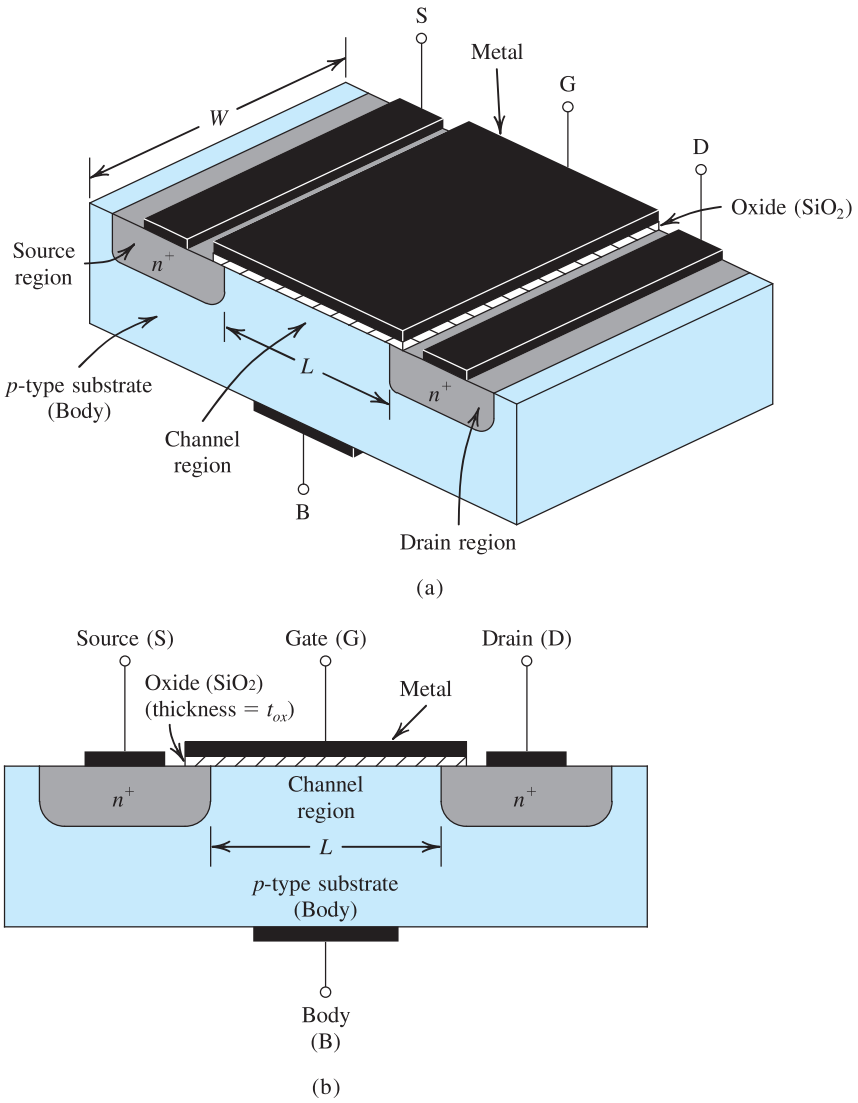


Figure 5.1 Physical structure of the enhancement-type NMOS transistor: **(a)** perspective view; **(b)** cross section. Typically $L = 0.03$ μm to 1 μm, $W = 0.1$ μm to 100 μm, and the thickness of the oxide layer (t_{ox}) is in the range of 1 to 10 nm.

body.[3] Thus four terminals are brought out: the gate terminal (G), the source terminal (S), the drain terminal (D), and the substrate or body terminal (B).

At this point it should be clear that the name of the device (metal-oxide-semiconductor FET) is derived from its physical structure. The name, however, has become a general one and

[3]In Fig. 5.1, the contact to the body is shown on the bottom of the device. This will prove helpful in Section 5.9 in explaining a phenomenon known as the "body effect." It is important to note, however, that in actual ICs, contact to the body is made at a location on the top of the device.

is used also for FETs that do not use metal for the gate electrode. In fact, most modern MOSFETs are fabricated using a process known as silicon-gate technology, in which a certain type of silicon, called polysilicon, is used to form the gate electrode (see Appendix A). Our description of MOSFET operation and characteristics applies irrespective of the type of gate electrode.

Another name for the MOSFET is the **insulated-gate FET** or **IGFET**. This name also arises from the physical structure of the device, emphasizing the fact that the gate electrode is electrically insulated from the device body (by the oxide layer). It is this insulation that causes the current in the gate terminal to be extremely small (of the order of 10^{-15} A).

Observe that the substrate forms *pn* junctions with the source and drain regions. In normal operation these *pn* junctions are kept reverse-biased at all times. Since, as we shall see shortly, the drain will always be at a positive voltage relative to the source, the two *pn* junctions can be effectively cut off by simply connecting the substrate terminal to the source terminal. We shall assume this to be the case in the following description of MOSFET operation. Thus, here, the substrate will be considered as having no effect on device operation, and the MOSFET will be treated as a three-terminal device, with the terminals being the gate (G), the source (S), and the drain (D). It will be shown that a voltage applied to the gate controls current flow between source and drain. This current will flow in the longitudinal direction from drain to source in the region labeled "channel region." Note that this region has a length L and a width W, two important parameters of the MOSFET. Typically, L is in the range of 0.03 μm to 1 μm, and W is in the range of 0.1 μm to 100 μm. Finally, note that the MOSFET is a symmetrical device; thus its source and drain can be interchanged with no change in device characteristics.

5.1.2 Operation with Zero Gate Voltage

With zero voltage applied to the gate, two back-to-back diodes exist in series between drain and source. One diode is formed by the *pn* junction between the n^+ drain region and the *p*-type substrate, and the other diode is formed by the *pn* junction between the *p*-type substrate and the n^+ source region. These back-to-back diodes prevent current conduction from drain to source when a voltage v_{DS} is applied. In fact, the path between drain and source has a very high resistance (of the order of 10^{12} Ω).

5.1.3 Creating a Channel for Current Flow

Consider next the situation depicted in Fig. 5.2. Here we have grounded the source and the drain and applied a positive voltage to the gate. Since the source is grounded, the gate voltage appears in effect between gate and source and thus is denoted v_{GS}. The positive voltage on the gate causes, in the first instance, the free holes (which are positively charged) to be repelled from the region of the substrate under the gate (the channel region). These holes are pushed downward into the substrate, leaving behind a carrier-depletion region. The depletion region is populated by the bound negative charge associated with the acceptor atoms. These charges are "uncovered" because the neutralizing holes have been pushed downward into the substrate.

As well, the positive gate voltage attracts electrons from the n^+ source and drain regions (where they are in abundance) into the channel region. When a sufficient number of electrons accumulate near the surface of the substrate under the gate, an *n* region is in effect created, connecting the source and drain regions, as indicated in Fig. 5.2. Now if a voltage is applied between drain and source, current flows through this induced *n* region, carried by the mobile electrons. The *induced n* region thus forms a **channel** for current flow from drain to source and *is aptly called so*. Correspondingly, the MOSFET of Fig. 5.2 is called an ***n*-channel MOSFET** or, alternatively, an **NMOS transistor**. Note that an *n*-channel MOSFET is

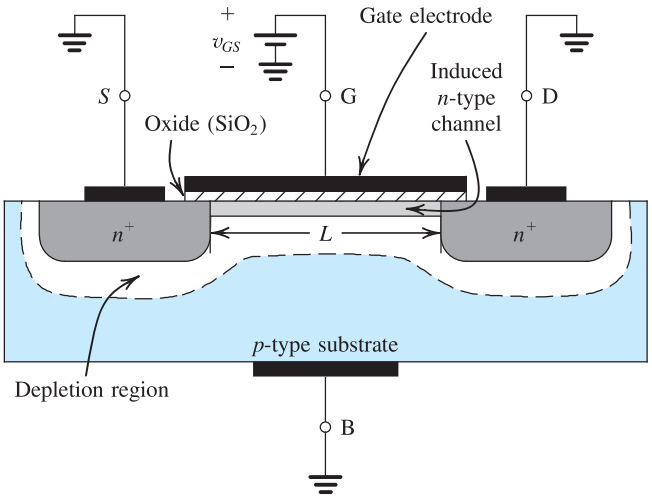


Figure 5.2 The enhancement-type NMOS transistor with a positive voltage applied to the gate. An *n* channel is induced at the top of the substrate beneath the gate.

formed in a *p*-type substrate: The channel is created by *inverting* the substrate surface from *p* type to *n* type. Hence the induced channel is also called an **inversion layer**.

The value of v_{GS} at which a sufficient number of mobile electrons accumulate in the channel region to form a conducting channel is called the **threshold voltage** and is denoted V_t.[4] Obviously, V_t for an *n*-channel FET is positive. The value of V_t is controlled during device fabrication and typically lies in the range of 0.3 V to 1.0 V.

The gate and the channel region of the MOSFET form a parallel-plate capacitor, with the oxide layer acting as the capacitor dielectric. The positive gate voltage causes positive charge to accumulate on the top plate of the capacitor (the gate electrode). The corresponding negative charge on the bottom plate is formed by the electrons in the induced channel. An electric field thus develops in the vertical direction. It is this field that controls the amount of charge in the channel, and thus it determines the channel conductivity and, in turn, the current that will flow through the channel when a voltage v_{DS} is applied. This is the origin of the name "field-effect transistor" (FET).

The voltage across this parallel-plate capacitor, that is, the voltage across the oxide, must exceed V_t for a channel to form. When $v_{DS} = 0$, as in Fig. 5.2, the voltage at every point along the channel is zero, and the voltage across the oxide (i.e., between the gate and the points along the channel) is uniform and equal to v_{GS}. The excess of v_{GS} over V_t is termed the **effective voltage** or the **overdrive voltage** and is the quantity that determines the charge in the channel. In this book, we shall denote $(v_{GS} - V_t)$ by v_{OV},

$$v_{GS} - V_t \equiv v_{OV} \tag{5.1}$$

We can express the magnitude of the electron charge in the channel by

$$|Q| = C_{ox}(WL)v_{OV} \tag{5.2}$$

[4]Some texts use V_T to denote the threshold voltage. We use V_t to avoid confusion with the thermal voltage V_T.

where C_{ox}, called the **oxide capacitance**, is the capacitance of the parallel-plate capacitor per unit gate area (in units of F/m^2), W is the width of the channel, and L is the length of the channel. The oxide capacitance C_{ox} is given by

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \tag{5.3}$$

where ε_{ox} is the permittivity of the silicon dioxide,

$$\varepsilon_{ox} = 3.9\varepsilon_0 = 3.9 \times 8.854 \times 10^{-12} = 3.45 \times 10^{-11} \text{ F/m}$$

The oxide thickness t_{ox} is determined by the process technology used to fabricate the MOSFET. As an example, for a process with $t_{ox} = 4$ nm,

$$C_{ox} = \frac{3.45 \times 10^{-11}}{4 \times 10^{-9}} = 8.6 \times 10^{-3} \text{ F/m}^2$$

It is much more convenient to express C_{ox} per micron squared. For our example, this yields 8.6 fF/μm^2, where fF denotes femtofarad (10^{-15} F). For a MOSFET fabricated in this technology with a channel length $L = 0.18$ μm and a channel width $W = 0.72$ μm, the total capacitance between gate and channel is

$$C = C_{ox}WL = 8.6 \times 0.18 \times 0.72 = 1.1 \text{ fF}$$

Finally, note from Eq. (5.2) that as v_{OV} is increased, the magnitude of the channel charge increases proportionately. Sometimes this is depicted as an increase in the depth of the channel; that is, the larger the overdrive voltage, the deeper the channel.

5.1.4 Applying a Small v_{DS}

Having induced a channel, we now apply a positive voltage v_{DS} between drain and source, as shown in Fig. 5.3. We first consider the case where v_{DS} is small (i.e., 50 mV or so). The voltage v_{DS} causes a current i_D to flow through the induced n channel. Current is carried by free electrons traveling from source to drain (hence the names source and drain). By convention, the direction of current flow is opposite to that of the flow of negative charge. Thus the current in the channel, i_D, will be from drain to source, as indicated in Fig. 5.3.

We now wish to calculate the value of i_D. Toward that end, we first note that because v_{DS} is small, we can continue to assume that the voltage between the gate and various points along the channel remains approximately constant and equal to the value at the source end, v_{GS}. Thus, the effective voltage between the gate and the various points along the channel remains equal to v_{OV}, and the channel charge Q is still given by Eq. (5.2). Of particular interest in calculating the current i_D is the charge per unit channel length, which can be found from Eq. (5.2) as

$$\frac{|Q|}{\text{unit channel length}} = C_{ox}Wv_{OV} \tag{5.4}$$

The voltage v_{DS} establishes an electric field E across the length of the channel,

$$|E| = \frac{v_{DS}}{L} \tag{5.5}$$

This electric field in turn causes the channel electrons to drift toward the drain with a velocity given by

$$\text{Electron drift velocity} = \mu_n |E| = \mu_n \frac{v_{DS}}{L} \tag{5.6}$$

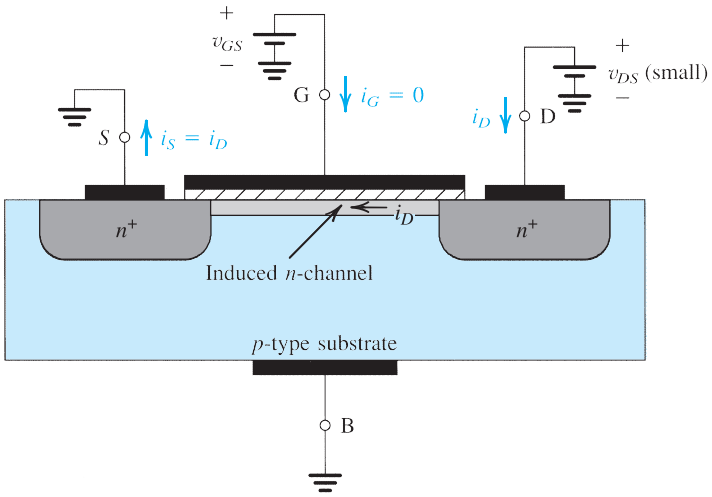


Figure 5.3 An NMOS transistor with $v_{GS} > V_t$ and with a small v_{DS} applied. The device acts as a resistance whose value is determined by v_{GS}. Specifically, the channel conductance is proportional to $v_{GS} - V_t$, and thus i_D is proportional to $(v_{GS} - V_t)v_{DS}$. Note that the depletion region is not shown (for simplicity).

where μ_n is the mobility of the electrons at the surface of the channel. It is a physical parameter whose value depends on the fabrication process technology. The value of i_D can now be found by multiplying the charge per unit channel length (Eq. 5.4) by the electron drift velocity (Eq. 5.6),

$$i_D = \left[(\mu_n C_{ox})\left(\frac{W}{L}\right)v_{OV}\right]v_{DS} \tag{5.7}$$

Thus, for small v_{DS}, the channel behaves as a linear resistance whose value is controlled by the overdrive voltage v_{OV}, which in turn is determined by v_{GS}:

$$i_D = \left[(\mu_n C_{ox})\left(\frac{W}{L}\right)(v_{GS} - V_t)\right]v_{DS} \tag{5.8}$$

The conductance g_{DS} of the channel can be found from Eq. (5.7) or (5.8) as

$$g_{DS} = (\mu_n C_{ox})\left(\frac{W}{L}\right)v_{OV} \tag{5.9}$$

or

$$g_{DS} = (\mu_n C_{ox})\left(\frac{W}{L}\right)(v_{GS} - V_t) \tag{5.10}$$

Observe that the conductance is determined by the product of three factors: $(\mu_n C_{ox})$, (W/L), and v_{OV} (or equivalently, $v_{GS} - V_t$). To gain insight into MOSFET operation, we consider each of the three factors in turn.

The first factor, $(\mu_n C_{ox})$, is determined by the process technology used to fabricate the MOSFET. It is the product of the electron mobility, μ_n, and the oxide capacitance, C_{ox}. It makes physical sense for the channel conductance to be proportional to each of μ_n and C_{ox}

(why?) and hence to their product, which is termed the **process transconductance parameter**[5] and given the symbol k_n' where the subscript n denotes n channel,

$$k_n' = \mu_n C_{ox} \tag{5.11}$$

It can be shown that with μ_n having the dimensions of meters squared per volt-second ($m^2/V{\cdot}s$) and C_{ox} having the dimensions of farads per meter squared (F/m^2), the dimensions of k_n' are amperes per volt squared (A/V^2).

The second factor in the expression for the conductance g_{DS} in Eqs. (5.9) and (5.10) is the transistor **aspect ratio** (W/L). That the channel conductance is proportional to the channel width W and inversely proportional to the channel length L should make perfect physical sense. The (W/L) ratio is obviously a dimensionless quantity that is determined by the device designer. Indeed, the values of W and L can be selected by the device designer to give the device the i–v characteristics desired. For a given fabrication process, however, there is a minimum channel length, L_{min}. In fact, the minimum channel length that is possible with a given fabrication process is used to characterize the process and is being continually reduced as technology advances. For instance, in 2009 the state-of-the-art in commercially available MOS technology was a 45-nm process, meaning that for this process the minimum channel length possible was 45 nm. Finally, we should note that the oxide thickness t_{ox} scales down with L_{min}. Thus, for a 0.13-μm technology, t_{ox} is 2.7 nm, but for the modern 45-nm technology t_{ox} is about 1.4 nm.

The product of the process transconductance parameter k_n' and the transistor aspect ratio (W/L) is the **MOSFET transconductance parameter** k_n,

$$k_n = k_n'(W/L) \tag{5.12a}$$

or

$$k_n = (\mu_n C_{ox})(W/L) \tag{5.12b}$$

The MOSFET parameter k_n has the dimensions of A/V^2.

The third term in the expression of the channel conductance g_{DS} is the overdrive voltage v_{OV}. This is hardly surprising since v_{OV} directly determines the magnitude of electron charge in the channel. As will be seen, v_{OV} is a very important circuit-design parameter. In this book, we will use v_{OV} and $v_{GS} - V_t$ interchangeably.

We conclude this subsection by noting that with v_{DS} kept small, the MOSFET behaves as a linear resistance r_{DS} whose value is controlled by the gate voltage v_{GS},

$$r_{DS} = \frac{1}{g_{DS}}$$

$$r_{DS} = \frac{1}{(\mu_n C_{ox})(W/L)v_{OV}} \tag{5.13a}$$

$$r_{DS} = \frac{1}{(\mu_n C_{ox})(W/L)(v_{GS} - V_t)} \tag{5.13b}$$

The operation of the MOSFET as a voltage-controlled resistance is further illustrated in Fig. 5.4, which is a sketch of i_D versus v_{DS} for various values of v_{GS}. Observe that the

[5]This name arises from the fact that $(\mu_n C_{ox})$ determines the transconductance of the MOSFET, as will be seen shortly.

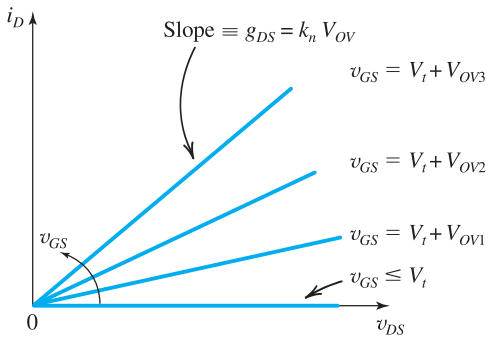


Figure 5.4 The i_D–v_{DS} characteristics of the MOSFET in Fig. 5.3 when the voltage applied between drain and source, v_{DS}, is kept small. The device operates as a linear resistance whose value is controlled by v_{GS}.

resistance is infinite for $v_{GS} \le V_t$ and decreases as v_{GS} is increased above V_t. It is interesting to note that although v_{GS} is used as the parameter for the set of graphs in Fig. 5.4, the graphs in fact depend only on v_{OV} (and, of course, k_n).

The description above indicates that for the MOSFET to conduct, a channel has to be induced. Then, increasing v_{GS} above the threshold voltage V_t enhances the channel, hence the names **enhancement-mode operation** and **enhancement-type MOSFET**. Finally, we note that the current that leaves the source terminal (i_S) is equal to the current that enters the drain terminal (i_D), and the gate current $i_G = 0$.

EXERCISE

5.1 A 0.18-μm fabrication process is specified to have $t_{ox} = 4$ nm, $\mu_n = 450$ cm^2/V·s, and $V_t = 0.5$ V. Find the value of the process transconductance parameter k'_n. For a MOSFET with minimum length fabricated in this process, find the required value of W so that the device exhibits a channel resistance r_{DS} of 1 kΩ at $v_{GS} = 1$V.
Ans. 388 μA/V^2; 0.93 μm

5.1.5 Operation as v_{DS} Is Increased

We next consider the situation as v_{DS} is increased. For this purpose, let v_{GS} be held constant at a value greater than V_t; that is, let the MOSFET be operated at a constant overdrive voltage V_{OV}. Refer to Fig. 5.5, and note that v_{DS} appears as a voltage drop across the length of the channel. That is, as we travel along the channel from source to drain, the voltage (measured relative to the source) increases from zero to v_{DS}. Thus the voltage between the gate and points along the channel decreases from $v_{GS} = V_t + V_{OV}$ at the source end to $v_{GD} = v_{GS} - v_{DS} = V_t + V_{OV} - v_{DS}$ at the drain end. Since the channel depth depends on this voltage, and specifically on the amount by which this voltage exceeds V_t, we find that the channel is no longer of uniform depth; rather, the channel will take the tapered shape shown in Fig. 5.5, being deepest at the source end (where the depth is proportional to V_{OV}) and shallowest at the drain end[6] (where the depth is proportional to $V_{OV} - v_{DS}$). This point is further illustrated in Fig. 5.6.

[6]For simplicity, we do not show in Fig. 5.5 the depletion region. Physically speaking, it is the widening of the depletion region as a result of the increased v_{DS} that makes the channel shallower near the drain.

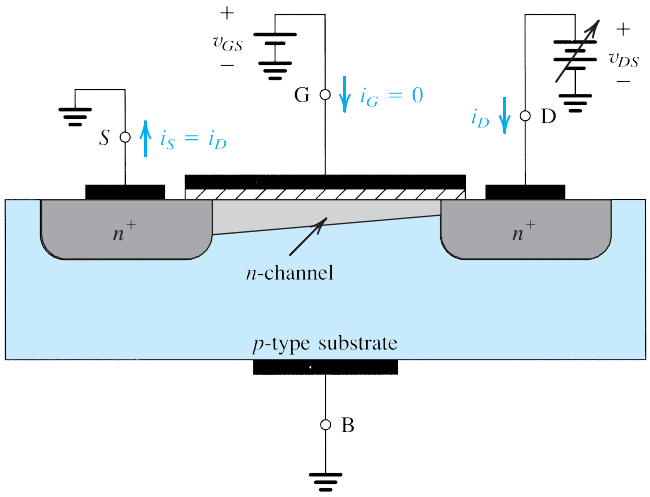


Figure 5.5 Operation of the enhancement NMOS transistor as v_{DS} is increased. The induced channel acquires a tapered shape, and its resistance increases as v_{DS} is increased. Here, v_{GS} is kept constant at a value $> V_t$; $v_{GS} = V_t + V_{OV}$.

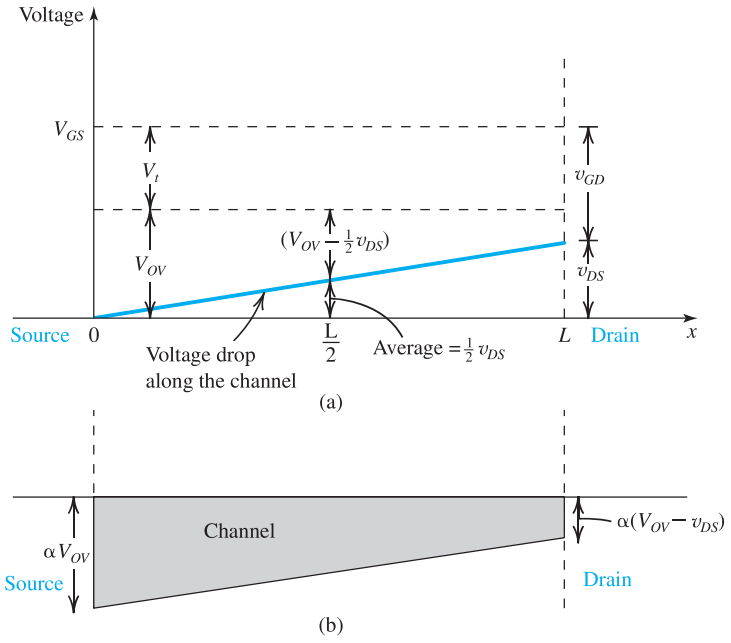


Figure 5.6 (**a**) For a MOSFET with $v_{GS} = V_t + V_{OV}$, application of v_{DS} causes the voltage drop along the channel to vary linearly, with an average value of $\frac{1}{2}v_{DS}$ at the midpoint. Since $v_{GD} > V_t$, the channel still exists at the drain end. (**b**) The channel shape corresponding to the situation in (a). While the depth of the channel at the source end is still proportional to V_{OV}, that at the drain end is proporational to $(V_{OV} - v_{DS})$.

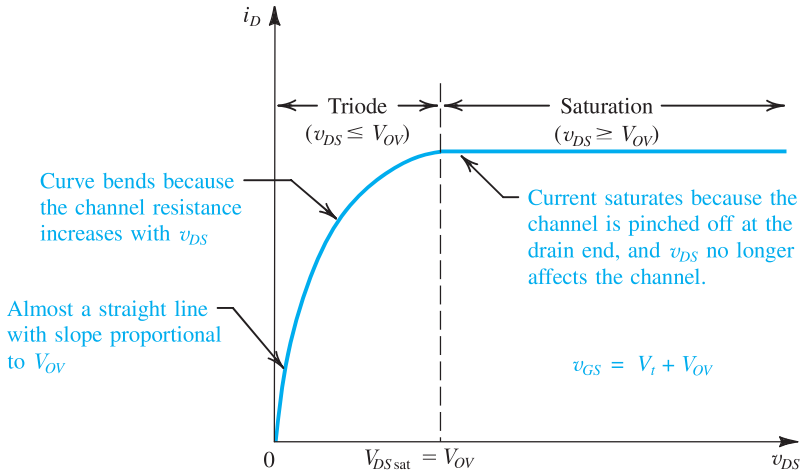


Figure 5.7 The drain current i_D versus the drain-to-source voltage v_{DS} for an enhancement-type NMOS transistor operated with $v_{GS} = V_t + V_{OV}$.

As v_{DS} is increased, the channel becomes more tapered and its resistance increases correspondingly. Thus, the i_D–v_{DS} curve does not continue as a straight line but bends as shown in Fig. 5.7. The equation describing this portion of the $i_D - v_{DS}$ curve can be easily derived by utilizing the information in Fig. 5.6. Specifically, note that the charge in the tapered channel is proportional to the channel cross-sectional area shown in Fig. 5.6(b). This area in turn can be easily seen as proportional to $\frac{1}{2}[V_{OV} + (V_{OV} - v_{DS})]$ or $(V_{OV} - \frac{1}{2}v_{DS})$. Thus, the relationship between i_D and v_{DS} can be found by replacing V_{OV} in Eq. (5.7) by $(V_{OV} - \frac{1}{2}v_{DS})$,

$$i_D = k'_n\left(\frac{W}{L}\right)\left(V_{OV} - \frac{1}{2}v_{DS}\right)v_{DS} \tag{5.14}$$

This relationship describes the semiparabolic portion of the i_D–v_{DS} curve in Fig. 5.7. It applies to the entire segment down to $v_{DS} = 0$. Specifically, note that as v_{DS} is reduced, we can neglect $\frac{1}{2}v_{DS}$ relative to V_{OV} in the factor in parentheses, and the expression reduces to that in Eq. (5.7). The latter of course is an approximation and applies only for small v_{DS} (i.e., near the origin).

There is another useful interpretation of the expression in Eq. (5.14). From Fig. 5.6(a) we see that the average voltage along the channel is $\frac{1}{2}v_{DS}$. Thus, the average voltage that gives rise to channel charge and hence to i_D is no longer V_{OV} but $(V_{OV} - \frac{1}{2}v_{DS})$, which is indeed the factor that appears in Eq. (5.14). Finally, we note that Eq. (5.14) is frequently written in the alternate form

$$i_D = k'_n\left(\frac{W}{L}\right)\left(V_{OV}\ v_{DS} - \frac{1}{2}v_{DS}^2\right) \tag{5.15}$$

Furthermore, for an arbitrary value of V_{OV}, we can replace V_{OV} by $(v_{GS} - V_t)$ and rewrite Eq. (5.15) as

$$i_D = k'_n\left(\frac{W}{L}\right)\left[(v_{GS} - V_t)v_{DS} - \frac{1}{2}v_{DS}^2\right] \tag{5.16}$$

5.1.6 Operation for $v_{DS} \geqslant V_{OV}$

The above description of operation assumed that even though the channel became tapered, it still had a finite (nonzero) depth at the drain end. This in turn is achieved by keeping v_{DS} sufficiently small that the voltage between the gate and the drain, v_{GD}, exceeds V_t. This is indeed the situation shown in Fig. 5.6(a). Note that for this situation to obtain, v_{DS} must not exceed V_{OV}, for as $v_{DS} = V_{OV}$, $v_{GD} = V_t$, and the channel depth at the drain end reduces to zero.

Figure 5.8 shows v_{DS} reaching V_{OV} and v_{GD} correspondingly reaching V_t. The zero depth of the channel at the drain end gives rise to the term **channel pinch-off**. Increasing v_{DS} beyond this value (i.e., $v_{DS} > V_{OV}$) has no effect on the channel shape and charge, and the current through the channel remains constant at the value reached for $v_{DS} = V_{OV}$. The drain current thus **saturates** at the value found by substituting $v_{DS} = V_{OV}$ in Eq. (5.14),

$$i_D = \frac{1}{2}k'_n\left(\frac{W}{L}\right)V_{OV}^2 \tag{5.17}$$

The MOSFET is then said to have entered the **saturation region** (or, equivalently, the saturation mode of operation). The voltage v_{DS} at which saturation occurs is denoted V_{DSsat},

$$V_{DSsat} = V_{OV} = V_{GS} - V_t \tag{5.18}$$

It should be noted that channel pinch-off does *not* mean channel blockage: Current continues to flow through the pinched-off channel, and the electrons that reach the drain end of the

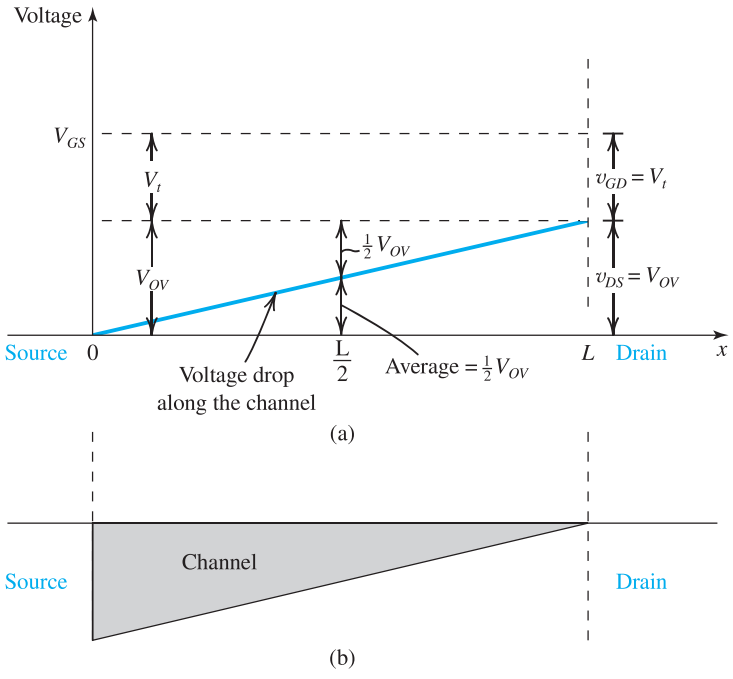


Figure 5.8 Operation of MOSFET with $v_{GS} = V_t + V_{OV}$, as v_{DS} is increased to V_{OV}. At the drain end, v_{GD} decreases to V_t and the channel depth at the drain end reduces to zero (pinch off). At this point, the MOSFET enters the saturation mode of operation. Further increasing v_{DS} (beyond $V_{Dsat} = V_{OV}$) has no effect on the channel shape and i_D remains constant.

channel are accelerated through the depletion region that exists there (not shown in Fig. 5.5) and into the drain terminal. Any increase in v_{DS} above V_{DSsat} appears as a voltage drop across the depletion region. Thus, both the current through the channel and the voltage drop across it remain constant in saturation.

The saturation portion of the $i_D - v_{DS}$ curve is, as expected, a horizontal straight line, as indicated in Fig. 5.7. Also indicated in Fig. 5.7 is the name of the region of operation obtained with a continuous (non-pinched-off) channel, the **triode region**. This name is a carryover from the days of vacuum-tube devices, whose operation a FET resembles.

Finally, we note that the $i_D - v_{DS}$ relationship in saturation can be generalized by replacing the constant overdrive voltage V_{OV} by a variable one, v_{OV}:

$$i_D = \frac{1}{2}k'_n\left(\frac{W}{L}\right)v_{OV}^2 \tag{5.19}$$

Also, v_{OV} can be replaced by $(v_{GS} - V_t)$ to obtain the alternate expression for saturation-mode i_D,

$$i_D = \frac{1}{2}k'_n\left(\frac{W}{L}\right)(v_{GS} - V_t)^2 \tag{5.20}$$

Example 5.1

Consider a process technology for which $L_{min} = 0.4\ \mu m$, $t_{ox} = 8$ nm, $\mu_n = 450\ cm^2/V \cdot s$, and $V_t = 0.7$ V.

(a) Find C_{ox} and k'_n.

(b) For a MOSFET with $W/L = 8\ \mu m/0.8\ \mu m$, calculate the values of V_{OV}, V_{GS}, and V_{DSmin} needed to operate the transistor in the saturation region with a dc current $I_D = 100\ \mu A$.

(c) For the device in (b), find the values of V_{OV} and V_{GS} required to cause the device to operate as a 1000-Ω resistor for very small v_{DS}.

Solution

(a)
$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} = \frac{3.45 \times 10^{-11}}{8 \times 10^{-9}} = 4.32 \times 10^{-3}\ F/m^2$$
$$= 4.32\ fF/\mu m^2$$
$$k'_n = \mu_n C_{ox} = 450\ (cm^2/V \cdot s) \times 4.32\ (fF/\mu m^2)$$
$$= 450 \times 10^8\ (\mu m^2/V \cdot s) \times 4.32 \times 10^{-15} (F/\mu m^2)$$
$$= 194 \times 10^{-6}\ (F/V \cdot s)$$
$$= 194\ \mu A/V^2$$

(b) For operation in the saturation region,

$$i_D = \frac{1}{2}k'_n\frac{W}{L}v_{OV}^2$$

Thus,

$$100 = \frac{1}{2} \times 194 \times \frac{8}{0.8}V_{OV}^2$$

which results in

$$V_{OV} = 0.32\ V$$

Example 5.1 *continued*

Thus,

$$V_{GS} = V_t + V_{OV} = 1.02 \text{ V}$$

and

$$V_{DS\min} = V_{OV} = 0.32 \text{ V}$$

(c) For the MOSFET in the triode region with v_{DS} very small,

$$r_{DS} = \frac{1}{k'_n \frac{W}{L} V_{OV}}$$

Thus

$$1000 = \frac{1}{194 \times 10^{-6} \times 10 \times V_{OV}}$$

which yields

$$V_{OV} = 0.52 \text{ V}$$

Thus,

$$V_{GS} = 1.22 \text{ V}$$

EXERCISES

5.2 For a 0.8-μm process technology for which $t_{ox} = 15$ nm and $\mu_n = 550$ cm^2/V·s, find C_{ox}, k'_n, and the overdrive voltage V_{OV} required to operate a transistor having $W/L = 20$ in saturation with $I_D = 0.2$ mA. What is the minimum value of V_{DS} needed?

Ans. 2.3 fF/μm^2; 127 μA/V^2; 0.40 V; 0.40 V

D5.3 A circuit designer intending to operate a MOSFET in saturation is considering the effect of changing the device dimensions and operating voltages on the drain current I_D. Specifically, by what factor does I_D change in each of the following cases?

(a) The channel length is doubled.
(b) The channel width is doubled.
(c) The overdrive voltage is doubled.
(d) The drain-to-source voltage is doubled.
(e) Changes (a), (b), (c), and (d) are made simultaneously.

Which of these cases might cause the MOSFET to leave the saturation region?

Ans. 0.5; 2; 4; no change; 4; case (c) if v_{DS} is smaller than 2 V_{OV}

5.1.7 The *p*-Channel MOSFET

Figure 5.9(a) shows a cross-sectional view of a *p*-channel enhancement-type MOSFET. The structure is similar to that of the NMOS device except that here the substrate is *n* type and the source and the drain regions are p^+ type; that is, all semiconductor regions are reversed in polarity relative to their counterparts in the NMOS case. The PMOS and NMOS transistor are said to be *complementary* devices.

To induce a channel for current flow between source and drain, a negative voltage is applied to the gate, that is, between gate and source, as indicated in Fig. 5.9(b). By increasing the magni-

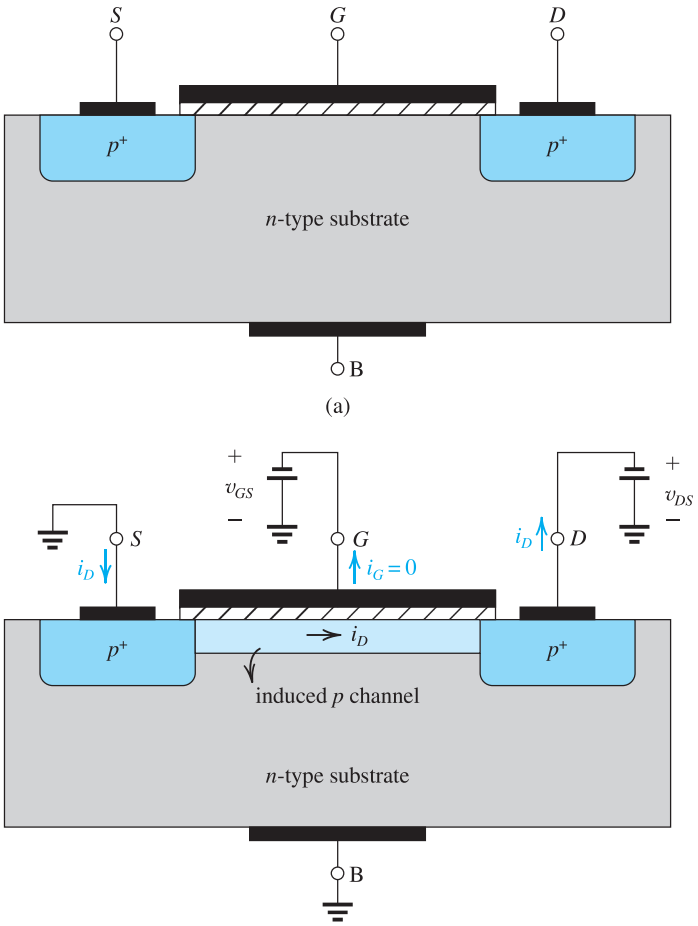


Figure 5.9 (**a**) Physical structure of the PMOS transistor. Note that it is similar to the NMOS transistor shown in Fig. 5.1(b) except that all semiconductor regions are reversed in polarity. (**b**) A negative voltage v_{GS} of magnitude greater than $|V_{tp}|$ induces a p channel, and a negative v_{DS} causes a current i_D to flow from source to drain.

tude of the negative v_{GS} beyond the magnitude of the threshold voltage V_{tp}, which by convention is negative, a p channel is established as shown in Fig. 5.9(b). This condition can be described as

$$v_{GS} \leq V_{tp}$$

or, to avoid dealing with negative signs,

$$|v_{GS}| \geq |V_{tp}|$$

Now, to cause a current i_D to flow in the p channel, a negative voltage v_{DS} is applied to the drain. The current i_D is carried by holes and flows through the channel from source to drain. As we have done for the NMOS transistor, we define the process transconductance parameter for the PMOS device as

$$k'_p = \mu_p C_{ox}$$

where μ_p is the mobility of the holes in the induced p channel. Typically, $\mu_p = 0.25\ \mu_n$ to $0.5\ \mu_n$ and is process technology dependent. The transistor transconductance parameter k_p is obtained by multiplying k'_p by the aspect ratio W/L,

$$k_p = k'_p(W/L)$$

The remainder of the description of the physical operation of the p-channel MOSFET follows that for the NMOS device, except of course for the sign reversals of all voltages. We will present the complete current–voltage characteristics of both NMOS and PMOS transistors in the next section.

PMOS technology originally dominated MOS integrated-circuit manufacturing, and the original microprocessors utilized PMOS transistors. As the technological difficulties of fabricating NMOS transistors were solved, NMOS completely supplanted PMOS. The main reason for this change is that electron mobility μ_n is higher by a factor of 2 to 4 than the hole mobility μ_p, resulting in NMOS transistors having greater gains and speeds of operation than PMOS devices. Subsequently, a technology was developed that permits the fabrication of both NMOS and PMOS transistors on the same chip. Appropriately called **complementary MOS**, or **CMOS**, this technology is currently the dominant electronics technology.

5.1.8 Complementary MOS or CMOS

As the name implies, complementary MOS technology employs MOS transistors of both polarities. Although CMOS circuits are somewhat more difficult to fabricate than NMOS, the availability of complementary devices makes possible many powerful circuit configurations. Indeed, at the present time CMOS is the most widely used of all the IC technologies. This statement applies to both analog and digital circuits. CMOS technology has virtually replaced designs based on NMOS transistors alone. Furthermore, by 2009 CMOS technology had taken over many applications that just a few years earlier were possible only with bipolar devices. Throughout this book, we will study many CMOS circuit techniques.

Figure 5.10 shows a cross section of a CMOS chip illustrating how the PMOS and NMOS transistors are fabricated. Observe that while the NMOS transistor is implemented directly in the p-type substrate, the PMOS transistor is fabricated in a specially created n region, known as an ***n* well**. The two devices are isolated from each other by a thick region of oxide that functions as an insulator. Not shown on the diagram are the connections made to the p-type body and to the n well. The latter connection serves as the body terminal for the PMOS transistor.

5.1.9 Operating the MOS Transistor in the Subthreshold Region

The above description of the n-channel MOSFET operation implies that for $v_{GS} < V_t$, no current flows and the device is cut off. This is not entirely true, for it has been found that for values of v_{GS} smaller than but close to V_t, a small drain current flows. In this **subthreshold region** of operation, the drain current is exponentially related to v_{GS}, much like the i_C–v_{BE} relationship of a BJT, as will be shown in the next chapter.

Although in most applications the MOS transistor is operated with $v_{GS} > V_t$, there are special, but a growing number of, applications that make use of subthreshold operation. In Chapter 13, we will briefly consider subthreshold operation.

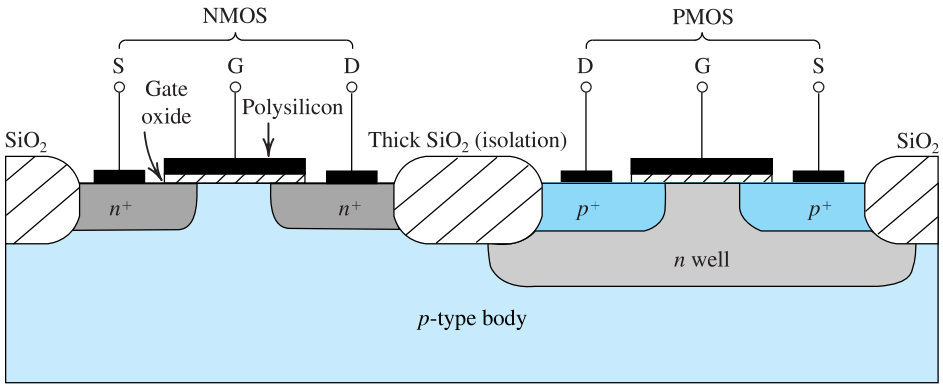


Figure 5.10 Cross-section of a CMOS integrated circuit. Note that the PMOS transistor is formed in a separate *n*-type region, known as an *n* well. Another arrangement is also possible in which an *n*-type body is used and the *n* device is formed in a *p* well. Not shown are the connections made to the *p*-type body and to the *n* well; the latter functions as the body terminal for the *p*-channel device.

5.2 Current–Voltage Characteristics

Building on the physical foundation established in the previous section for the operation of the enhancement MOS transistor, in this section we present its complete current–voltage characteristics. These characteristics can be measured at dc or at low frequencies and thus are called static characteristics. The dynamic effects that limit the operation of the MOSFET at high frequencies and high switching speeds will be discussed in Chapter 9.

5.2.1 Circuit Symbol

Figure 5.11(a) shows the circuit symbol for the *n*-channel enhancement-type MOSFET. Observe that the spacing between the two vertical lines that represent the gate and the channel indicates the fact that the gate electrode is insulated from the body of the device. The polarity of the *p*-type substrate (body) and the *n* channel is indicated by the arrowhead on the

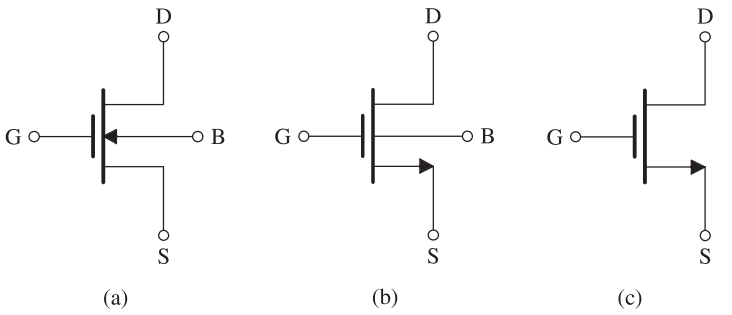


Figure 5.11 **(a)** Circuit symbol for the *n*-channel enhancement-type MOSFET. **(b)** Modified circuit symbol with an arrowhead on the source terminal to distinguish it from the drain and to indicate device polarity (i.e., *n* channel). **(c)** Simplified circuit symbol to be used when the source is connected to the body or when the effect of the body on device operation is unimportant.

line representing the body (B). This arrowhead also indicates the polarity of the transistor, namely, that it is an *n*-channel device.

Although the MOSFET is a symmetrical device, it is often useful in circuit design to designate one terminal as the source and the other as the drain (without having to write S and D beside the terminals). This objective is achieved in the modified circuit symbol shown in Fig. 5.11(b). Here an arrowhead is placed on the source terminal, thus distinguishing it from the drain terminal. The arrowhead points in the normal direction of current flow and thus indicates the polarity of the device (i.e., *n* channel). Observe that in the modified symbol, there is no need to show the arrowhead on the body line. Although the circuit symbol of Fig. 5.11(b) clearly distinguishes the source from the drain, in practice it is the polarity of the voltage impressed across the device that determines source and drain; *the drain is always positive relative to the source in an n-channel FET*.

In applications where the source is connected to the body of the device, a further simplification of the circuit symbol is possible, as indicated in Fig. 5.11(c). This symbol is also used in applications when the effect of the body on circuit operation is not important, as will be seen later.

5.2.2 The i_D–v_{DS} Characteristics

Table 5.1 provides a compilation of the conditions and the formulas for the operation of the NMOS transistor in each of the three possible regions: the cutoff region, the triode region, and the saturation region. The first two are useful if the MOSFET is to be utilized as a switch. On the other hand, if the MOSFET is to be used to design an amplifier, it must be operated in the saturation region. The rationale for these choices will be addressed in Section 5.4.

At the top of Table 5.1 we show a circuit consisting of an NMOS transistor and two dc supplies providing v_{GS} and v_{DS}. This conceptual circuit can be used to measure the i_D–v_{DS} characteristic curves of the NMOS transistor. Each curve is measured by setting v_{GS} to a desired constant value, varying v_{DS}, and measuring the corresponding i_D. Two of these characteristic curves are shown in the accompanying diagram: one for $v_{GS} < V_{tn}$ and the other for $v_{GS} = V_{tn} + v_{OV}$. (Note that we now use V_{tn} to denote the threshold voltage of the NMOS transistor, to distinguish it from that of the PMOS transistor, denoted V_{tp}.)

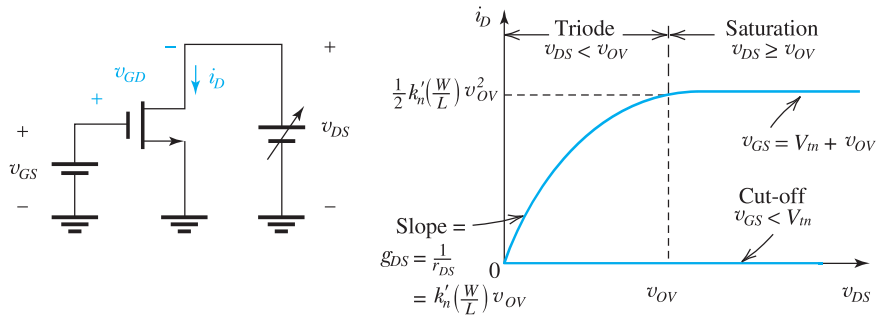
As Table 5.1 shows, the boundary between the triode region and the saturation region is determined by whether v_{DS} is less or greater than the overdrive voltage v_{OV} at which the transistor is operating. An equivalent way to check for the region of operation is to examine the relative values of the drain and gate voltages. To operate in the triode region, the gate voltage must exceed the drain voltage by at least V_{tn} volts, which ensures that the channel remains continuous (not pinched off). On the other hand, to operate in saturation, the channel must be pinched off at the drain end; pinch-off is achieved here by keeping v_D higher than $v_G - V_{tn}$, that is, not allowing v_D to fall below v_G by more than V_{tn} volts. The graphical construction of Fig. 5.12 should serve to remind the reader of these conditions.

A set of i_D–v_{DS} characteristics for the NMOS transistor is shown in Fig. 5.13. Observe that each graph is obtained by setting v_{GS} above V_{tn} by a specific value of overdrive voltage, denoted V_{OV1}, V_{OV2}, V_{OV3}, and V_{OV4}. This in turn is the value of v_{DS} at which the corresponding graph saturates, and the value of the resulting saturation current is directly determined by the value of v_{OV}, namely, $\frac{1}{2}k'_n V_{OV1}^2$, $\frac{1}{2}k'_n V_{OV2}^2$, . . . The reader is advised to commit to memory both the structure of these graphs and the coordinates of the saturation points.

Finally, observe that the boundary between the triode and the saturation regions, that is, the locus of the saturation points, is a parabolic curve described by

$$i_D = \frac{1}{2}k'_n \left(\frac{W}{L}\right) v_{DS}^2$$

Table 5.1 Regions of Operation of the Enhancement NMOS Transistor



- $v_{GS} < V_{tn}$: no channel; transistor in cut-off; $i_D = 0$
- $v_{GS} = V_{tn} + v_{OV}$: a channel is induced; transistor operates in the triode region or the saturation region depending on whether the channel is continuous or pinched-off at the drain end;

Triode Region

Continuous channel, obtained by:

$$v_{GD} > V_{tn}$$

or equivalently:

$$v_{DS} < v_{OV}$$

Then,

$$i_D = k_n'\left(\frac{W}{L}\right)\left[(v_{GS} - V_{tn})v_{DS} - \frac{1}{2}v_{DS}^2\right]$$

or equivalently,

$$i_D = k_n'\left(\frac{W}{L}\right)\left(v_{OV} - \frac{1}{2}v_{DS}\right)v_{DS}$$

Saturation Region

Pinched-off channel, obtained by:

$$v_{GD} \leq V_{tn}$$

or equivalently:

$$v_{DS} \geq v_{OV}$$

Then

$$i_D = \frac{1}{2}k_n'\left(\frac{W}{L}\right)(v_{GS} - V_{tn})^2$$

or equivalently,

$$i_D = \frac{1}{2}k_n'\left(\frac{W}{L}\right)v_{OV}^2$$

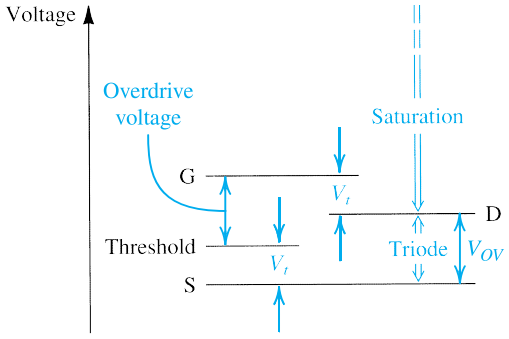


Figure 5.12 The relative levels of the terminal voltages of the enhancement NMOS transistor for operation in the triode region and in the saturation region.

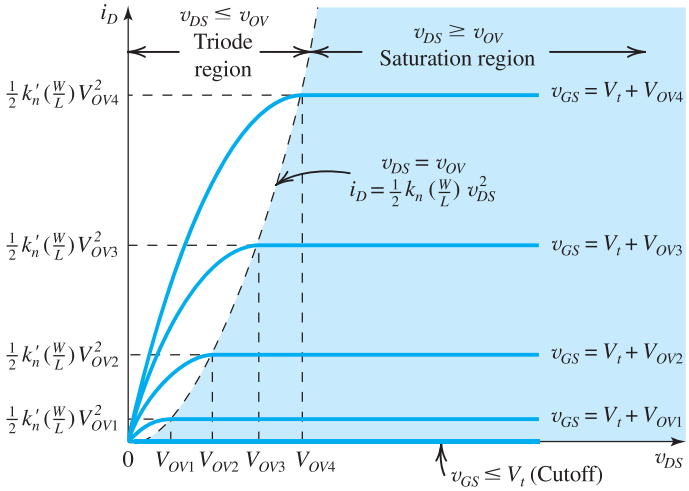


Figure 5.13 The i_D–v_{DS} characteristics for an enhancement-type NMOS transistor.

5.2.3 The i_D–v_{GS} Characteristic

When the MOSFET is used to design an amplifier, it is operated in the saturation region. As Fig. 5.13 indicates, in saturation the drain current is constant determined by v_{GS} (or v_{OV}) and is independent of v_{DS}. That is, the MOSFET operates as a constant-current source where the value of the current is determined by v_{GS}. In effect, then, the MOSFET operates as a voltage-controlled current source with the control relationship described by

$$i_D = \frac{1}{2}k_n'\left(\frac{W}{L}\right)(v_{GS} - V_{tn})^2 \tag{5.21}$$

or in terms of v_{OV},

$$i_D = \frac{1}{2}k_n'\left(\frac{W}{L}\right)v_{OV}^2 \tag{5.22}$$

This is the relationship that underlies the application of the MOSFET as an amplifier. That it is nonlinear should be of concern to those interested in designing linear amplifiers. Nevertheless, later in this chapter, we will see how one can obtain linear amplification from this nonlinear control or transfer characteristic.

Figure 5.14 shows the i_D–v_{GS} characteristic of an NMOS transistor operating in saturation. Note that if we are interested in a plot of i_D versus v_{OV}, we simply shift the origin to the point $v_{GS} = V_{tn}$.

The view of the MOSFET in the saturation region as a voltage-controlled current source is illustrated by the equivalent-circuit representation shown in Fig. 5.15. For reasons that will become apparent shortly, the circuit in Fig. 5.15 is known as a **large-signal equivalent circuit**. Note that the current source is ideal, with an infinite output resistance representing the independence, in saturation, of i_D from v_{DS}. This, of course, has been assumed in the idealized model of device operation utilized thus far. We are about to rectify an important shortcoming of this model. First, however, we present an example.

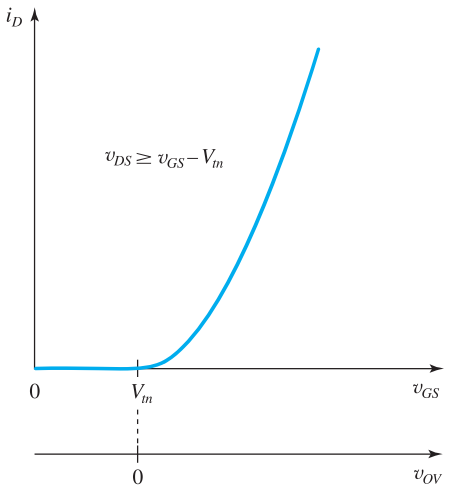


Figure 5.14 The i_D–v_{GS} characteristic of an NMOS transistor operating in the saturation region. The i_D–v_{OV} characteristic can be obtained by simply re-labelling the horizontal axis; that is, shifting the origin to the point $v_{GS} = V_{tn}$.

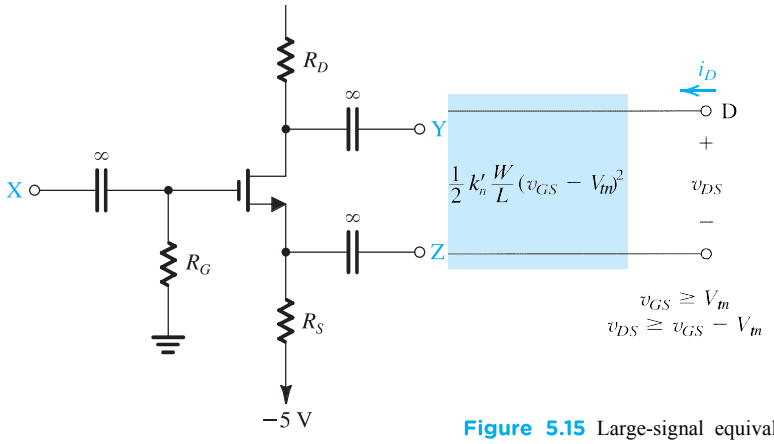


Figure 5.15 Large-signal equivalent-circuit model of an *n*-channel MOSFET operating in the saturation

Example 5.2

Consider an NMOS transistor fabricated in a 0.18-µm process with $L = 0.18$ µm and $W = 2$ µm. The process technology is specified to have $C_{ox} = 8.6$ fF/µm^2, $\mu_n = 450$ cm^2/V·s, and $V_{tn} = 0.5$ V.

(a) Find V_{GS} and V_{DS} that result in the MOSFET operating at the edge of saturation with $I_D = 100$ µA.
(b) If V_{GS} is kept constant, find V_{DS} that results in $I_D = 50$ µA.
(c) To investigate the use of the MOSFET as a linear amplifier, let it be operating in saturation with $V_{DS} = 0.3$ V. Find the change in i_D resulting from v_{GS} changing from 0.7 V by +0.01 V and by –0.01 V.

Solution

First we determine the process transconductance parameter k_n'

$$\begin{aligned} k_n' &= \mu_n C_{ox} \\ &= 450 \times 10^{-4} \times 8.6 \times 10^{-15} \times 10^{12} \text{ A/V}^2 \\ &= 387 \text{ μA/V}^2 \end{aligned}$$

and the transistor transconductance parameter k_n,

$$\begin{aligned} k_n &= k_n'\left(\frac{W}{L}\right) \\ &= 387\left(\frac{2}{0.18}\right) = 4.3 \text{ mA/V}^2 \end{aligned}$$

(a) With the transistor operating in saturation,

$$I_D = \frac{1}{2} k_n V_{OV}^2$$

Thus,

$$100 = \frac{1}{2} \times 4.3 \times 10^3 \times V_{OV}^2$$

which results in

$$V_{OV} = 0.22 \text{ V}$$

Thus,

$$V_{GS} = V_{tn} + V_{OV} = 0.5 + 0.22 = 0.72 \text{ V}$$

and since operation is at the edge of saturation,

$$V_{DS} = V_{OV} = 0.22 \text{ V}$$

(b) With V_{GS} kept constant at 0.72 V and I_D reduced from the value obtained at the edge of saturation, the MOSFET will now be operating in the triode region, thus

$$\begin{aligned} I_D &= k_n\left[V_{OV} V_{DS} - \frac{1}{2} V_{DS}^2\right] \\ 50 &= 4.3 \times 10^3\left[0.22 V_{DS} - \frac{1}{2} V_{DS}^2\right] \end{aligned}$$

which can be rearranged to the form

$$V_{DS}^2 - 0.44 V_{DS} + 0.023 = 0$$

This quadratic equation has two solutions

$$V_{DS} = 0.06 \text{ V} \quad \text{and} \quad V_{DS} = 0.39 \text{ V}$$

The second answer is greater than V_{OV} and thus is physically meaningless, since we know that the transistor is operating in the triode region. Thus we have

$$V_{DS} = 0.06 \text{ V}$$

Example 5.2 *continued*

(c) For $v_{GS} = 0.7$ V, $V_{OV} = 0.2$ V, and since $V_{DS} = 0.3$ V, the transistor is operating in saturation and

$$I_D = \frac{1}{2}k_n V_{OV}^2$$
$$= \frac{1}{2} \times 4300 \times 0.04$$
$$= 86\ \mu\text{A}$$

Now for $v_{GS} = 0.710$ V, $v_{OV} = 0.21$ V and

$$i_D = \frac{1}{2} \times 4300 \times 0.21^2 = 94.8\ \mu\text{A}$$

and for $v_{GS} = 0.690$ V, $v_{OV} = 0.19$ V, and

$$i_D = \frac{1}{2} \times 4300 \times 0.19^2 = 77.6\ \mu\text{A}$$

Thus, with $\Delta V_{GS} = +0.01$ V, $\Delta i_D = 8.8\ \mu$A; and for $\Delta V_{GS} = -0.01$ V, $\Delta i_D = -8.4\ \mu$A.

We conclude that the two changes are almost equal, an indication of almost-linear operation when the changes in v_{GS} are kept small. This is just a preview of the "small-signal operation" of the MOSFET studied in Sections 5.4 and 5.5.

EXERCISES

5.4 An NMOS transistor is operating at the edge of saturation with an overdrive voltage V_{OV} and a drain current I_D. If V_{OV} is doubled, and we must maintain operation at the edge of saturation, what should V_{DS} be changed to? What value of drain current results?
Ans. $2\ V_{OV}$; $4\ I_D$

5.5 An *n*-channel MOSFET operating with $V_{OV} = 0.5$ V exhibits a linear resistance $r_{DS} = 1$ kΩ when v_{DS} is very small. What is the value of the device transconductance parameter k_n? What is the value of the current I_D obtained when v_{DS} is increased to 0.5 V? and to 1 V?
Ans. 2 mA/V^2; 0.25 mA; 0.25 mA

5.2.4 Finite Output Resistance in Saturation

Equation (5.21) and the corresponding large-signal equivalent circuit in Fig. 5.15, as well as the graphs in Fig. 5.13, indicate that in saturation, i_D is independent of v_{DS}. Thus, a change Δv_{DS} in the drain-to-source voltage causes a zero change in i_D, which implies that the incremental resistance looking into the drain of a saturated MOSFET is infinite. This, however, is an idealization based on the premise that once the channel is pinched off at the drain end, further increases in v_{DS} have no effect on the channel's shape. But, in practice, increasing v_{DS} beyond v_{OV} does affect the channel somewhat. Specifically, as v_{DS} is increased, the channel pinch-off point is moved slightly away from the drain, toward the source. This is illustrated in Fig. 5.16, from which we note that the voltage across the channel remains constant at v_{OV}, and the additional voltage applied to the drain appears as a voltage drop across the narrow depletion region between the end of the channel and the drain region. This voltage accelerates the electrons that reach the drain end of the channel and sweeps them across the depletion region into the drain. Note, however, that (with depletion-layer widening) the channel

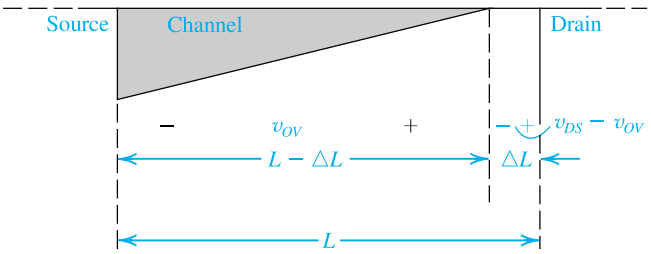


Figure 5.16 Increasing v_{DS} beyond v_{DSsat} causes the channel pinch-off point to move slightly away from the drain, thus reducing the effective channel length (by ΔL).

length is in effect reduced, from L to $L - \Delta L$, a phenomenon known as **channel-length modulation**. Now, since i_D is inversely proportional to the channel length (Eq. 5.21), i_D increases with v_{DS}.

This effect can be accounted for in the expression for i_D by including a factor $1 + \lambda(v_{DS} - v_{OV})$ or, for simplicity, $(1 + \lambda v_{DS})$,

$$i_D = \frac{1}{2}k'_n\left(\frac{W}{L}\right)(v_{GS} - V_{tn})^2 \ (1 + \lambda v_{DS}) \tag{5.23}$$

Here λ is a device parameter having the units of reciprocal volts (V^{-1}). The value of λ depends both on the process technology used to fabricate the device and on the channel length L that the circuit designer selects. Specifically, the value of λ is much larger for newer submicron technologies than for older technologies. This makes intuitive sense: Newer technologies have very short channels, and are thus much greatly impacted by the channel-length modulation effect. Also, for a given process technology, λ is inversely proportional to L.

A typical set of i_D–v_{DS} characteristics showing the effect of channel-length modulation is displayed in Fig. 5.17. The observed linear dependence of i_D on v_{DS} in the saturation region is represented in Eq. (5.23) by the factor $(1 + \lambda v_{DS})$. From Fig. 5.17 we observe that when the straight-line i_D–v_{DS} characteristics are extrapolated, they intercept the v_{DS} axis at the point, $v_{DS} = -V_A$, where V_A is a positive voltage. Equation (5.23), however, indicates that $i_D = 0$ at $v_{DS} = -1/\lambda$. It follows that

$$V_A = \frac{1}{\lambda}$$

and thus V_A is a device parameter with the dimensions of V. For a given process, V_A is proportional to the channel length L that the designer selects for a MOSFET. We can isolate the dependence of V_A on L by expressing it as

$$V_A = V'_A L$$

where V'_A is entirely process-technology dependent with the dimensions of volts per micron. Typically, V'_A falls in the range of 5 V/μm to 50 V/μm. The voltage V_A is usually referred to as the Early voltage, after J. M. Early, who discovered a similar phenomenon for the BJT (Chapter 6).

Equation (5.23) indicates that when channel-length modulation is taken into account, the saturation values of i_D depend on v_{DS}. Thus, for a given v_{GS}, a change Δv_{DS} yields a corresponding change Δi_D in the drain current i_D. It follows that the output resistance of the current source representing i_D in saturation is no longer infinite. Defining the output resistance r_o as[7]

[7]In this book we use r_o to denote the output resistance in saturation, and r_{DS} to denote the drain-to-source resistance in the triode region, for small v_{DS}.

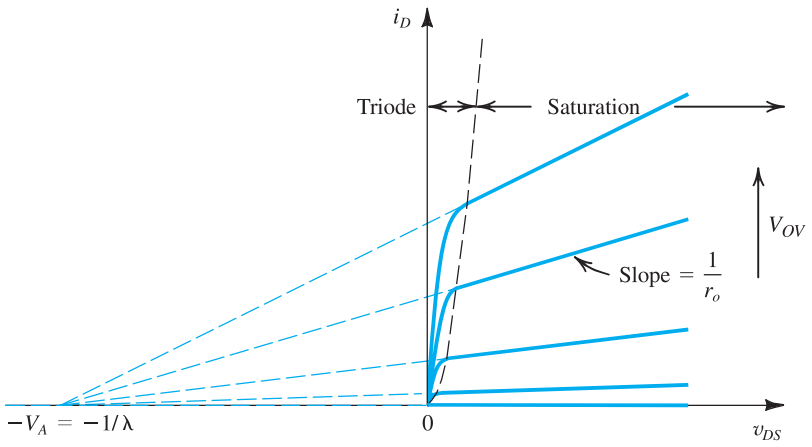


Figure 5.17 Effect of v_{DS} on i_D in the saturation region. The MOSFET parameter V_A depends on the process technology and, for a given process, is proportional to the channel length L.

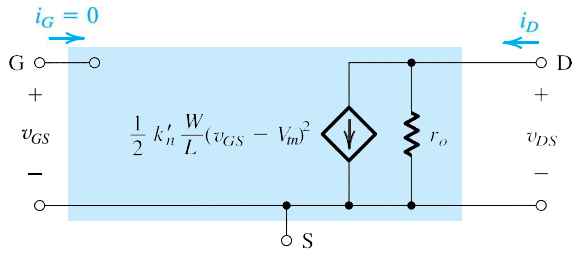


Figure 5.18 Large-signal equivalent circuit model of the *n*-channel MOSFET in saturation, incorporating the output resistance r_o. The output resistance models the linear dependence of i_D on v_{DS} and is given by Eq. (5.23).

$$r_o \equiv \left[\frac{\partial i_D}{\partial v_{DS}}\right]^{-1}_{v_{GS}\text{ constant}} \tag{5.24}$$

and using Eq. (5.23) results in

$$r_o = \left[\lambda \frac{k_n'}{2}\frac{W}{L}(V_{GS} - V_{tn})^2\right]^{-1} \tag{5.25}$$

which can be written as

$$r_o = \frac{1}{\lambda I_D} \tag{5.26}$$

or, equivalently,

$$r_o = \frac{V_A}{I_D} \tag{5.27}$$

where I_D is the drain current *without* channel-length modulation taken into account; that is,

$$I_D = \frac{1}{2} k_n' \frac{W}{L}(V_{GS} - V_{tn})^2 \tag{5.27'}$$

Thus the output resistance is inversely proportional to the drain current. Finally, we show in Fig. 5.18 the large-signal, equivalent-circuit model incorporating r_o.

EXERCISE

5.6 An NMOS transistor is fabricated in a 0.4-μm process having $\mu_n C_{ox} = 200\ \mu A/V^2$ and $V_A' = 50\ V/\mu m$ of channel length. If $L = 0.8\ \mu m$ and $W = 16\ \mu m$, find V_A and λ. Find the value of I_D that results when the device is operated with an overdrive voltage $V_{OV} = 0.5$ V and $V_{DS} = 1$ V. Also, find the value of r_o at this operating point. If V_{DS} is increased by 2 V, what is the corresponding change in I_D?
Ans. 40 V; 0.025 V^{-1}; 0.51 mA; 80 kΩ; 0.025 mA

5.2.5 Characteristics of the *p*-Channel MOSFET

The circuit symbol for the *p*-channel enhancement-type MOSFET is shown in Fig. 5.19(a). Figure 5.19(b) shows a modified circuit symbol in which an arrowhead pointing in the normal direction of current flow is included on the source terminal. For the case where the source is connected to the substrate, the simplified symbol of Fig. 5.19(c) is usually used.

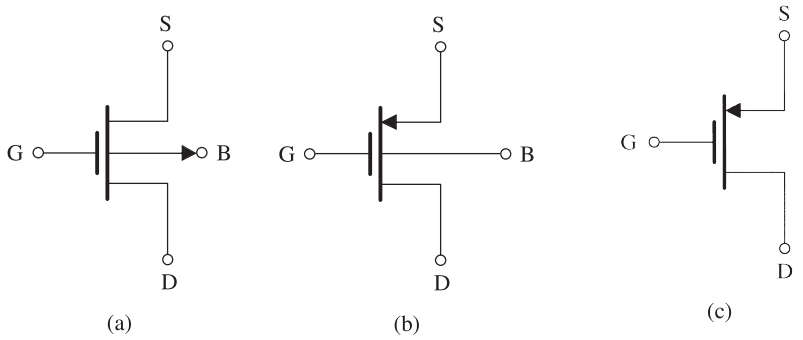


Figure 5.19 **(a)** Circuit symbol for the *p*-channel enhancement-type MOSFET. **(b)** Modified symbol with an arrowhead on the source lead. **(c)** Simplified circuit symbol for the case where the source is connected to the body.

The regions of operation of the PMOS transistor and the corresponding conditions and expression for i_D are shown in Table 5.2. Observe that the equations are written in a way that emphasizes physical intuition and avoids the confusion of negative signs. Thus while V_{tp} is by convention negative, we use $|V_{tp}|$, and the voltages v_{SG} and v_{SD} are positive. Also, in all of our circuit diagrams we will always draw *p*-channel devices with their sources on top so that current flows from top to bottom. Finally, we note that PMOS devices also suffer from the channel-length modulation effect. This can be taken into account by including a factor $(1 + |\lambda| v_{SD})$ in the saturation-region expression for i_D as follows

$$i_D = \frac{1}{2} k_p' \left(\frac{W}{L}\right) (v_{SG} - |V_{tp}|)^2 (1 + |\lambda| v_{SD}) \tag{5.28}$$

or equivalently

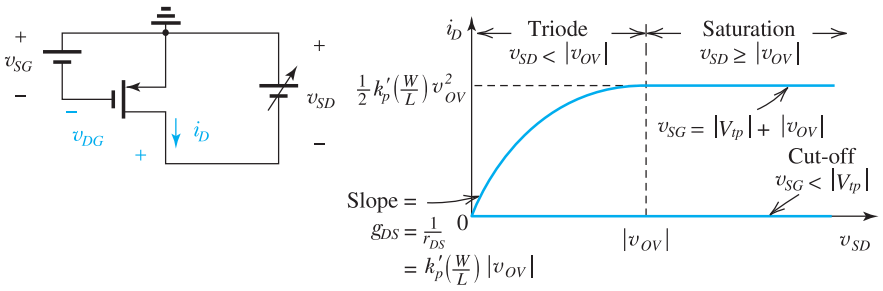
$$i_D = \frac{1}{2}k_p'\left(\frac{W}{L}\right)(v_{SG} - |V_{tp}|)^2\left(1 + \frac{v_{SD}}{|V_A|}\right) \tag{5.29}$$

where λ and V_A (the Early voltage for the PMOS transistor) are by convention negative quantities, hence we use $|\lambda|$ and $|V_A|$.

Finally, we should note that for a given CMOS fabrication process λ_n and $|\lambda_p|$ are generally not equal, and similarly for V_{An} and $|V_{Ap}|$.

To recap, to turn a PMOS transistor on, the gate voltage has to be made lower than that of the source by at least $|V_{tp}|$. To operate in the triode region, the drain voltage has to exceed that of the gate by at least $|V_{tp}|$; otherwise, the PMOS operates in saturation. Finally, Fig. 5.20 provides a pictorial representation of these operating conditions.

Table 5.2 Regions of Operation of the Enhancement PMOS Transistor



- $v_{SG} < |V_{tp}|$: no channel; transistor in cut-off; $i_D = 0$
- $v_{SG} = |V_{tp}| + |v_{OV}|$: a channel is induced; transistor operates in the triode region or in the saturation region depending on whether the channel is continuous or pinched-off at the drain end;

Triode Region

Continuous channel, obtained by:

$$v_{DG} > |V_{tp}|$$

or equivalently:

$$v_{SD} < |v_{OV}|$$

Then,

$$i_D = k_p'\left(\frac{W}{L}\right)\left[(v_{SG} - |V_{tp}|)v_{SD} - \frac{1}{2}v_{SD}^2\right]$$

or equivalently

$$i_D = k_p'\left(\frac{W}{L}\right)\left(|v_{OV}| - \frac{1}{2}v_{SD}\right)v_{SD}$$

Saturation Region

Pinched-off channel, obtained by:

$$v_{DG} \leq |V_{tp}|$$

or equivalently

$$v_{SD} \geq |v_{OV}|$$

Then

$$i_D = \frac{1}{2}k_p'\left(\frac{W}{L}\right)(v_{SG} - |V_{tp}|)^2$$

or equivalently

$$i_D = \frac{1}{2}k_p'\left(\frac{W}{L}\right)v_{OV}^2$$

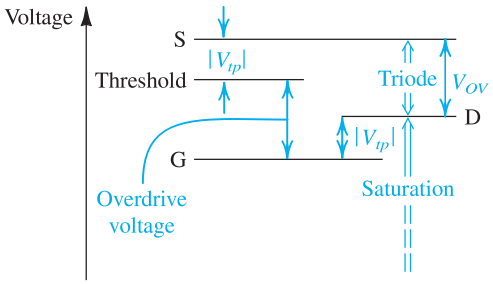


Figure 5.20 The relative levels of the terminal voltages of the enhancement-type PMOS transistor for operation in the triode region and in the saturation region.

EXERCISE

5.7 The PMOS transistor shown in Fig. E5.7 has $V_{tp} = -1$ V, $k_p' = 60\ \mu A/V^2$, and $W/L = 10$.
(a) Find the range of V_G for which the transistor conducts.
(b) In terms of V_G, find the range of V_D for which the transistor operates in the triode region.
(c) In terms of V_G, find the range of V_D for which the transistor operates in saturation.
(d) Neglecting channel-length modulation (i.e., assuming $\lambda = 0$), find the values of $|V_{OV}|$ and V_G and the corresponding range of V_D to operate the transistor in the saturation mode with $I_D = 75\ \mu A$.
(e) If $\lambda = -0.02\ V^{-1}$, find the value of r_o corresponding to the overdrive voltage determined in (d).
(f) For $\lambda = -0.02\ V^{-1}$ and for the value of V_{OV} determined in (d), find I_D at $V_D = +3$ V and at $V_D = 0$ V; hence, calculate the value of the apparent output resistance in saturation. Compare to the value found in (e).

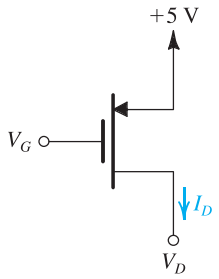


Figure E5.7

Ans. (a) $V_G \leq +4$ V; (b) $V_D \geq V_G + 1$; (c) $V_D \leq V_G + 1$; (d) 0.5 V, 3.5 V, ≤4.5 V; (e) 0.67 MΩ; (f) 78 μA, 82.5 μA, 0.67 MΩ (same)

5.3 MOSFET Circuits at DC

Having studied the current–voltage characteristics of MOSFETs, we now consider circuits in which only dc voltages and currents are of concern. Specifically, we shall present a series of design and analysis examples of MOSFET circuits at dc. The objective is to instill in the reader a familiarity with the device and the ability to perform MOSFET circuit analysis both rapidly and effectively.

In the following examples, to keep matters simple and thus focus attention on the essence of MOSFET circuit operation, we will generally neglect channel-length modulation; that is, we will assume $\lambda = 0$. We will find it convenient to work in terms of the overdrive voltage; $V_{OV} = V_{GS} - V_{tn}$ for NMOS and $|V_{OV}| = V_{SG} - |V_{tp}|$ for PMOS.