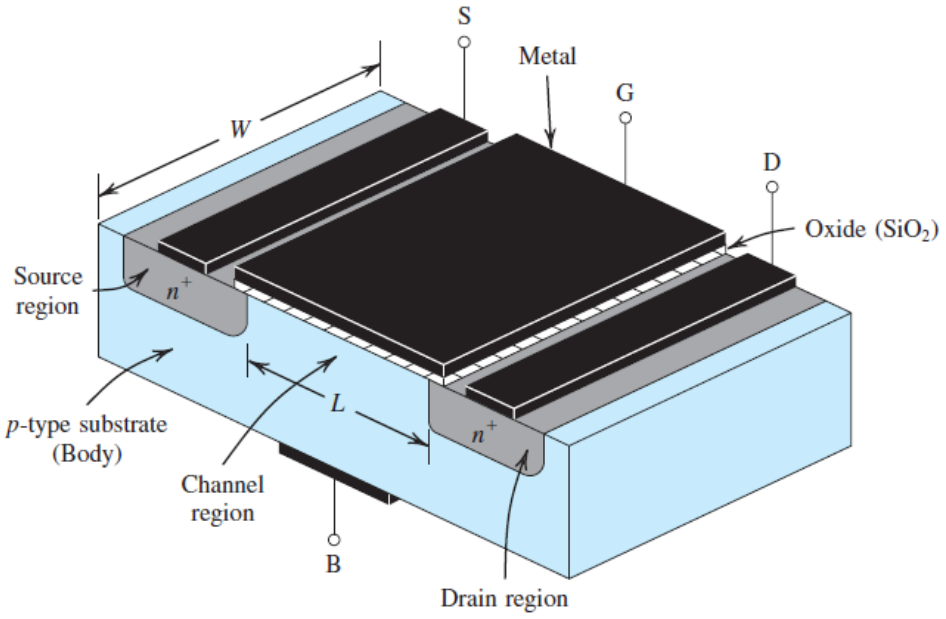
MOSFET

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 nm 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 **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).



(a)

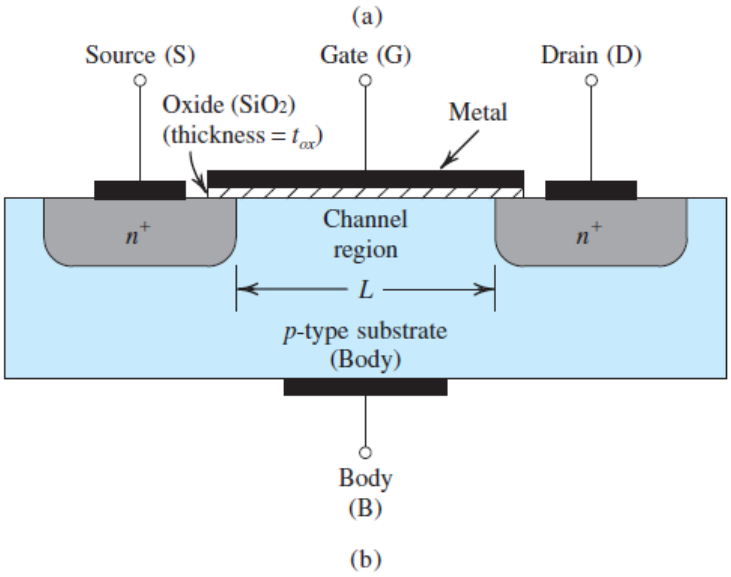


Figure 5.1 Physical structure of the enhancement-type NMOS transistor: (**a**) perspective view; (**b**) cross section. Typically $L = 0.03\ \mu m$ to $1\ \mu m$, $W = 0.05\ \mu m$ to $100\ \mu m$, and the thickness of the oxide layer (t_{ox}) is in the range of 1 to 10 nm.

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 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 $0.03\ \mu m$ to $1\ \mu m$, and W is in the range of $0.05\ \mu m$ to $100\ \mu m$. 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

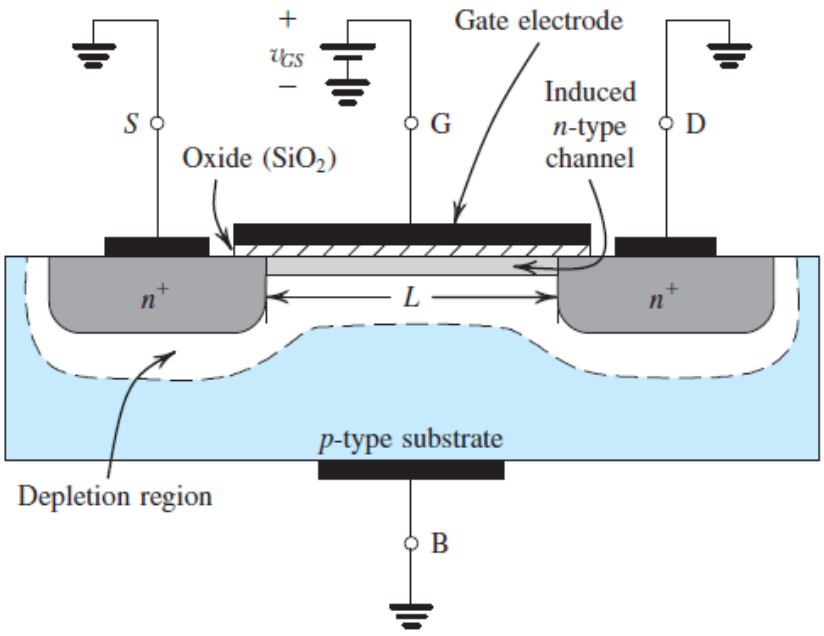


Figure 5.2 The enhancement-type NMOS transistor with a positive voltage applied to the gate. An *n* channel is induced at the top of the substrate beneath the gate.

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 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},

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

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.

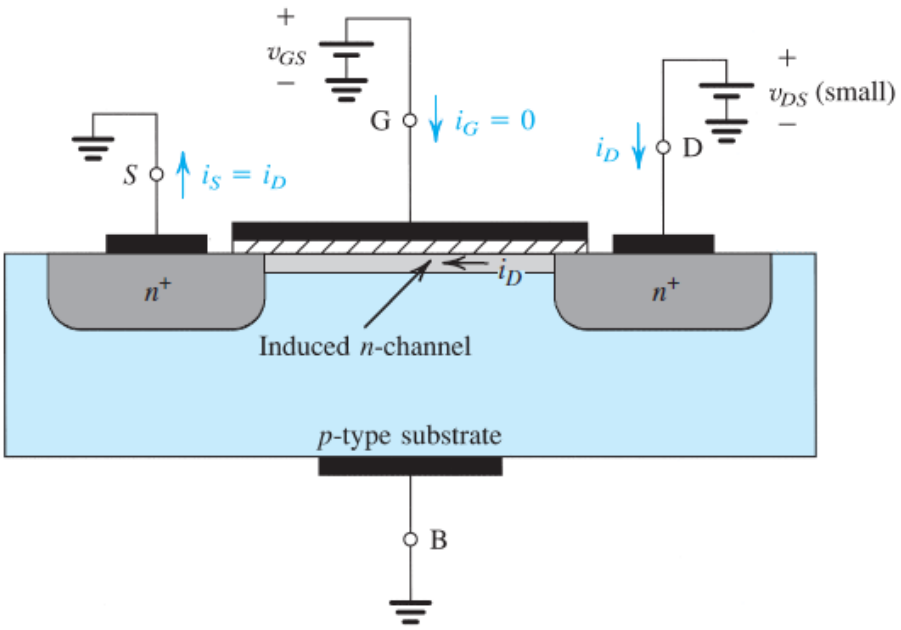


Figure 5.3 An NMOS transistor with $v_{GS} > V_t$ and with a small v_{DS} applied. The device acts as a resistance whose value is determined by v_{GS}. Specifically, the channel conductance is proportional to $v_{GS} - V_t$, and thus i_D is proportional to $(v_{GS} - V_t)v_{DS}$. Note that the depletion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

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.

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 $\left(V_{OV} - \frac{1}{2}v_{DS}\right)$. Thus, the relationship between i_D and v_{DS} can be found by replacing V_{OV} in Eq. (5.7) by $\left(V_{OV} - \frac{1}{2}v_{DS}\right)$,

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

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 $\left(V_{OV} - \frac{1}{2}v_{DS}\right)$, which is indeed the factor that appears in Eq. (5.14). Finally, we note that Eq. (5.14) is frequently written in the

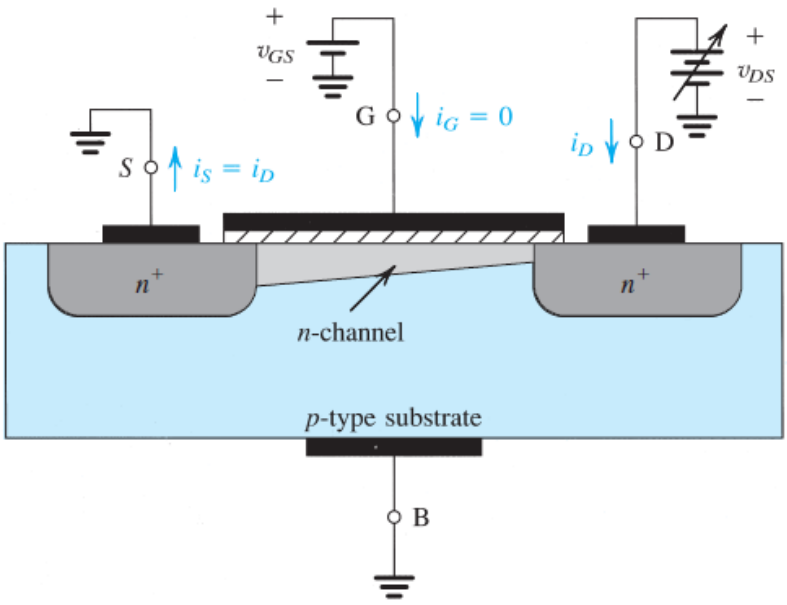


Figure 5.5 Operation of the enhancement NMOS transistor as v_{DS} is increased. The induced channel acquires a tapered shape, and its resistance increases as v_{DS} is increased. Here, v_{GS} is kept constant at a value $> V_t$; $v_{GS} = V_t + V_{OV}$.

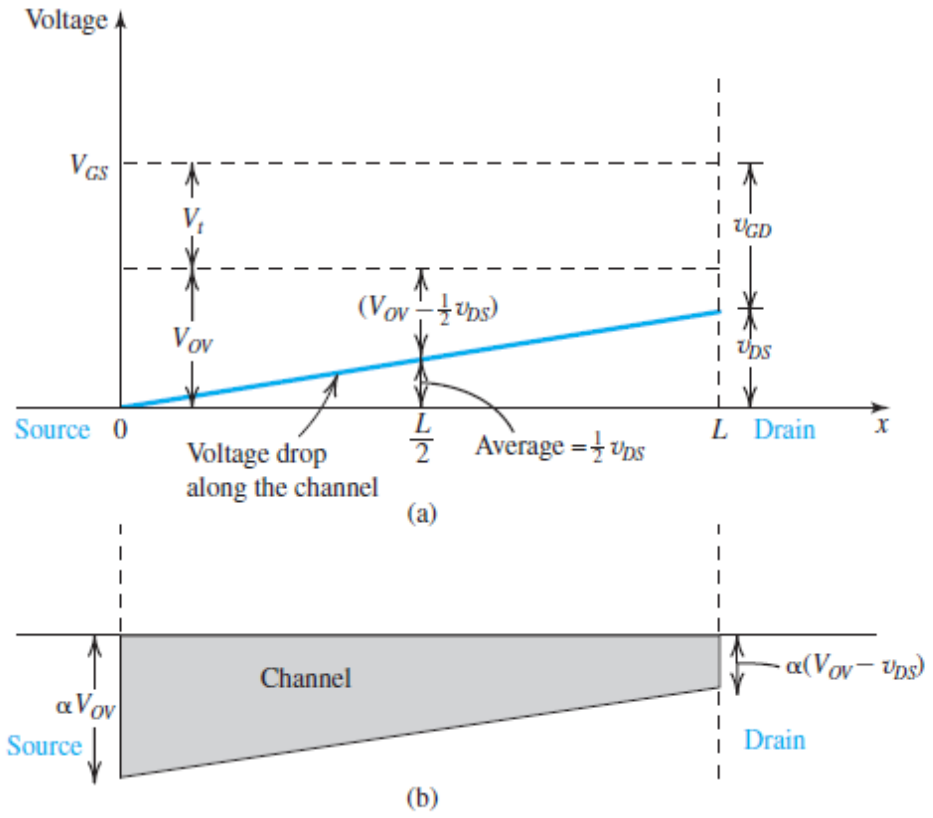


Figure 5.6 (a) For a MOSFET with $v_{GS} = V_t + V_{OV}$, application of v_{DS} causes the voltage drop along the channel to vary linearly, with an average value of $\frac{1}{2}v_{DS}$ at the midpoint. Since $v_{GD} > V_t$, the channel still exists at the drain end. (b) The channel shape corresponding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

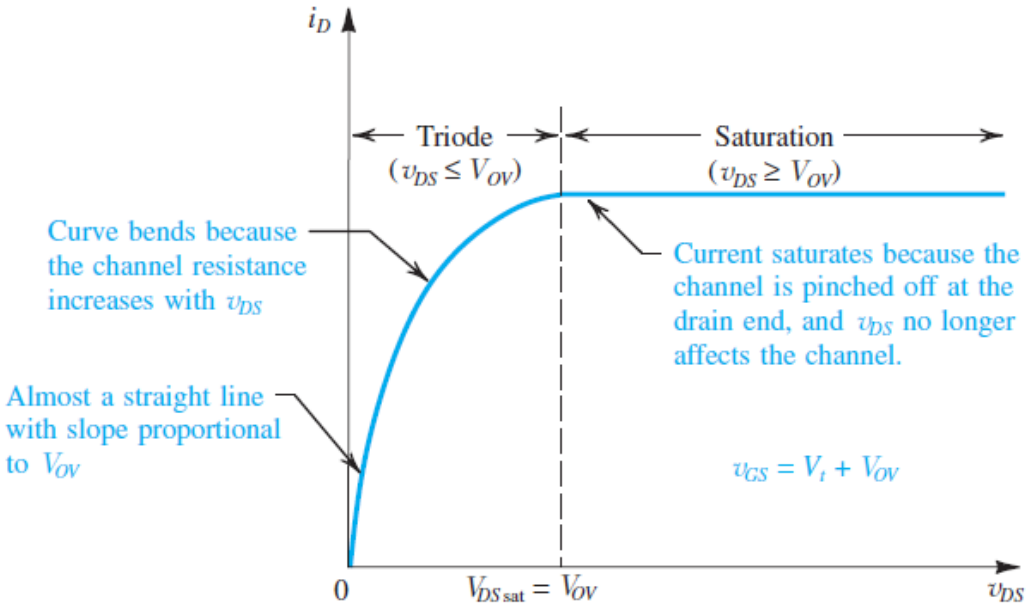


Figure 5.7 The drain current i_D versus the drain-to-source voltage v_{DS} for an enhancement-type NMOS transistor operated with $v_{GS} = V_t + V_{OV}$.

5.1.6 Operation for $v_{DS} \geq V_{OV}$: Channel Pinch-Off and Current Saturation

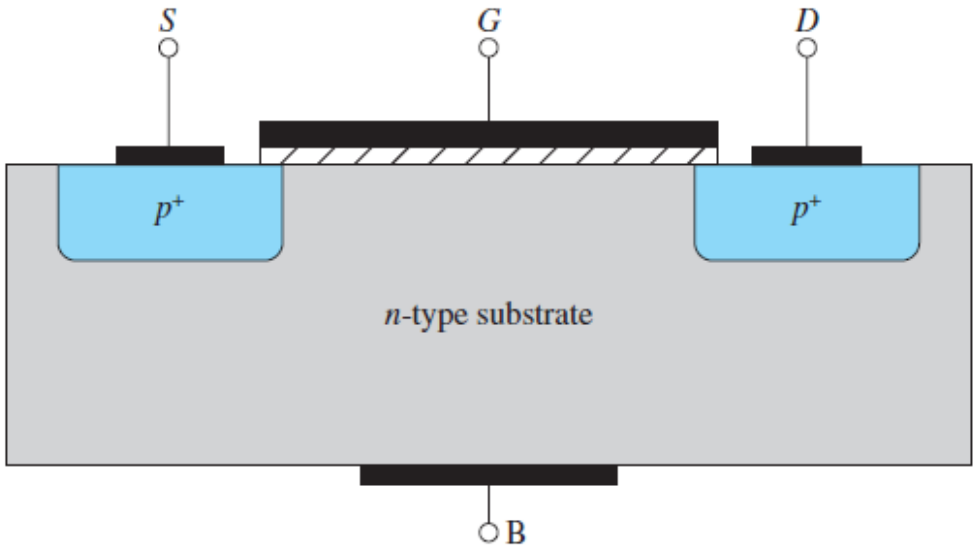
The above description of operation assumed that even though the channel became tapered, it still had a finite (nonzero) depth at the drain end. This in turn is achieved by keeping v_{DS} sufficiently small that the voltage between the gate and the drain, v_{GD}, exceeds V_t. This is indeed the situation 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),

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

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 transistors are said to be *complementary* devices.



(a)

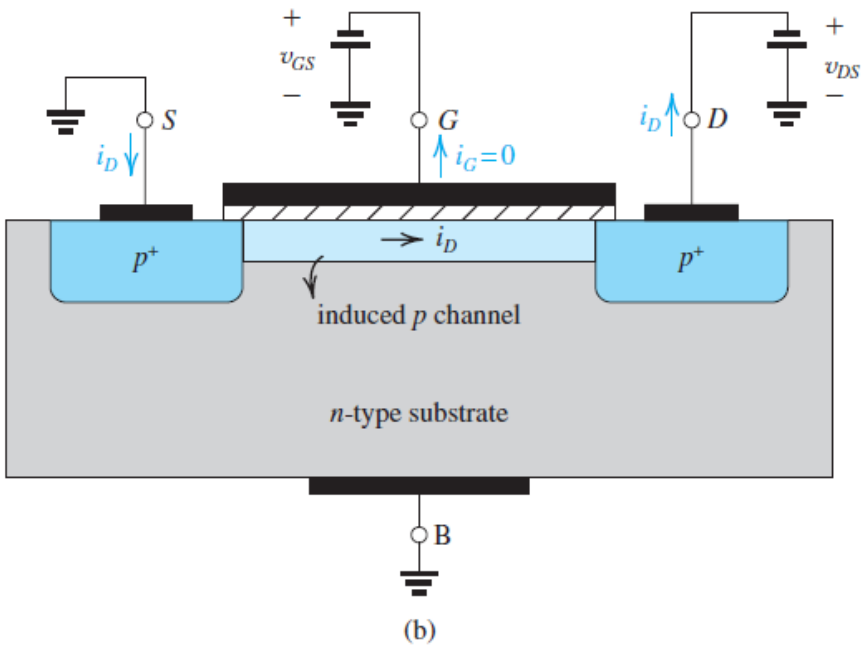


Figure 5.9 (a) Physical structure of the PMOS transistor. Note that it is similar to the NMOS transistor shown in Fig. 5.1(b) except that all semiconductor regions are reversed in polarity. (b) A negative voltage v_{GS} of magnitude greater than $|V_{tp}|$ induces a p channel, and a negative v_{DS} causes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

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 magnitude 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,

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

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

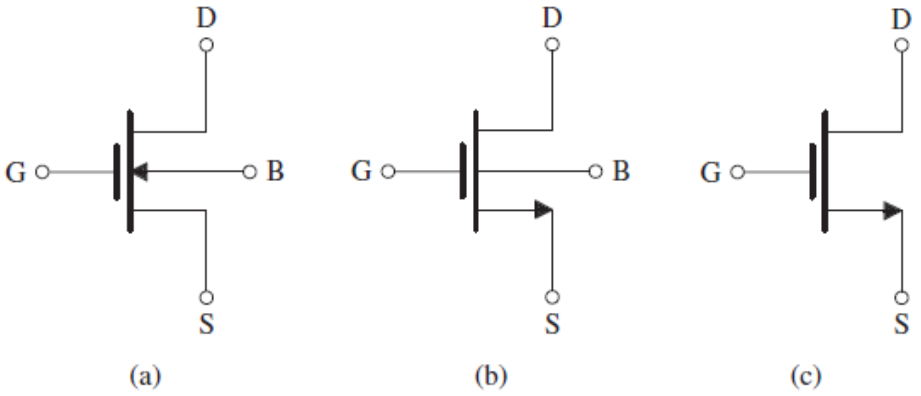


Figure 5.11 (a) Circuit symbol for the *n*-channel enhancement-type MOSFET. (b) Modified circuit symbol with an arrowhead on the source terminal to distinguish it from the drain and to indicate device polarity (i.e., *n* channel). (c) Simplified circuit symbol to be used when the source is connected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

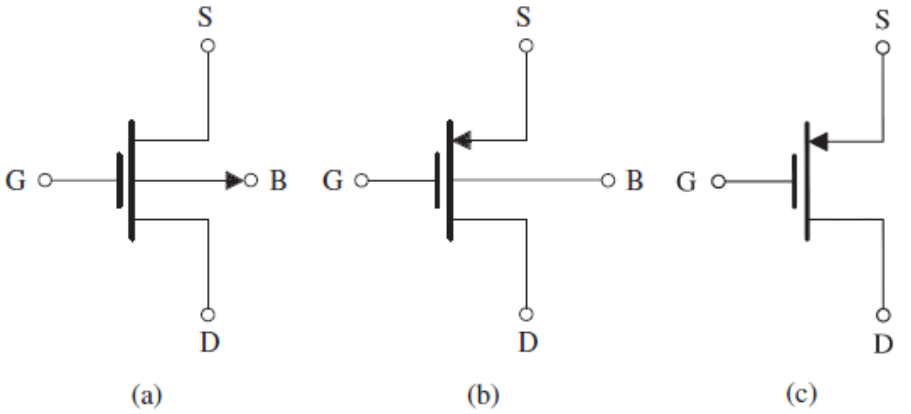
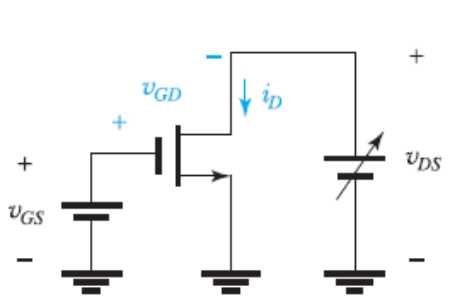
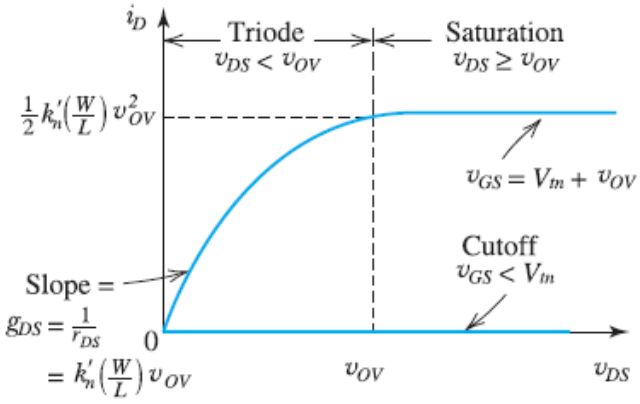


Figure 5.19 (a) Circuit symbol for the *p*-channel enhancement-type MOSFET. (b) Modified symbol with an arrowhead on the source lead. (c) Simplified circuit symbol for the case where the source is connected to the body.

Table 5.1 Regions of Operation of the Enhancement NMOS Transistor





- $v_{GS} < V_{tn}$: no channel; transistor in cutoff; $i_D = 0$
- $v_{GS} = V_{tn} + v_{OV}$: 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:

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

Saturation Region

Pinched-off channel, obtained by:

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

or equivalently:

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

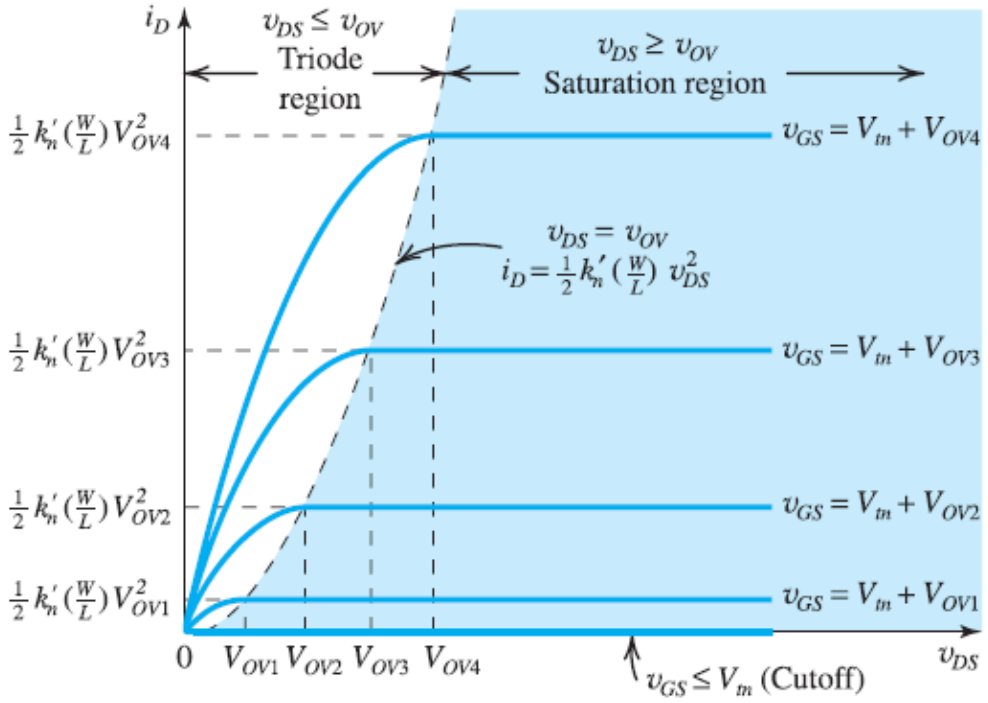


Figure 5.13 The $i_D - v_{DS}$ characteristics for an enhancement-type NMOS transistor.

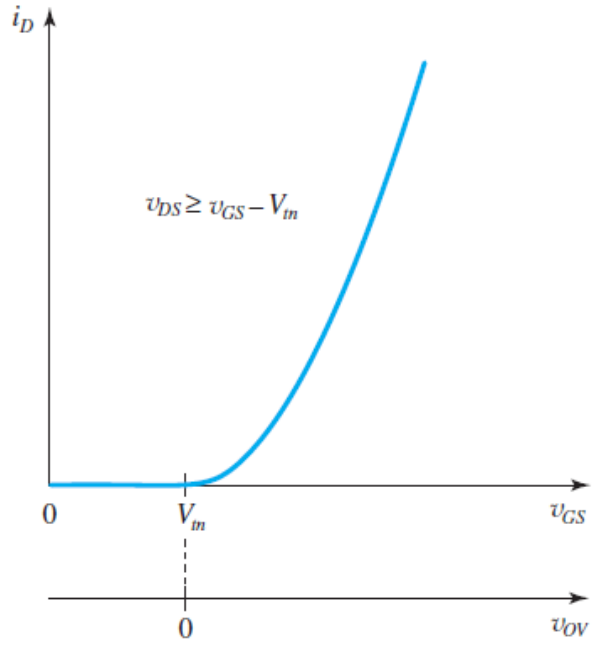


Figure 5.14 The i_D–v_{GS} characteristic of an NMOS transistor operating in the saturation region. The i_D–v_{OV} characteristic can be obtained by simply relabeling the horizontal axis, that is, shifting the origin to the point $v_{GS} = V_{tn}$.

4.4 THE MOSFET AS AN AMPLIFIER AND AS A SWITCH

In this section we begin our study of the use of MOSFETs in the design of amplifier circuits.[6] The basis for this important MOSFET application is that when operated in the saturation region, the MOSFET acts as a voltage-controlled current source: Changes in the gate-to-source voltage v_{GS} give rise to changes in the drain current i_D. Thus the saturated MOSFET can be used to implement a *transconductance amplifier* (see Section 1.5). However, since we are interested in linear amplification—that is, in amplifiers whose output signal (in this case, the drain current i_D) is linearly related to their input signal (in this case, the gate-to-source voltage v_{GS})—we will have to find a way around the highly nonlinear (square-law) relationship of i_D to v_{GS}.

The technique we will utilize to obtain linear amplification from a fundamentally nonlinear device is that of **dc biasing** the MOSFET to operate at a certain appropriate V_{GS} and a corresponding I_D and then superimposing the voltage signal to be amplified, v_{gs}, on the dc bias voltage V_{GS}. By keeping the signal v_{gs} "small," the resulting change in drain current, i_d, can be made proportional to v_{gs}. This technique was introduced in a general way in Section 1.4 and was applied in the case of the diode in Section 3.3.8. However, before considering the small-signal operation of the MOSFET amplifier, we will look at the "big picture": We will study the total or large-signal operation of a MOSFET amplifier. We will do this by deriving the voltage transfer characteristic of a commonly used MOSFET amplifier circuit. From the voltage transfer characteristic we will be able to clearly see the region over which the transistor can be biased to operate as a small-signal amplifier as well as those regions where it can be operated as a switch (i.e., being either fully "on" or fully "off"). MOS switches find application in both analog and digital circuits.

4.4.1 Large-Signal Operation—The Transfer Characteristic

Figure 4.26(a) shows the basic structure (skeleton) of the most commonly used MOSFET amplifier, the **common-source (CS)** circuit. The name common-source or **grounded-source**

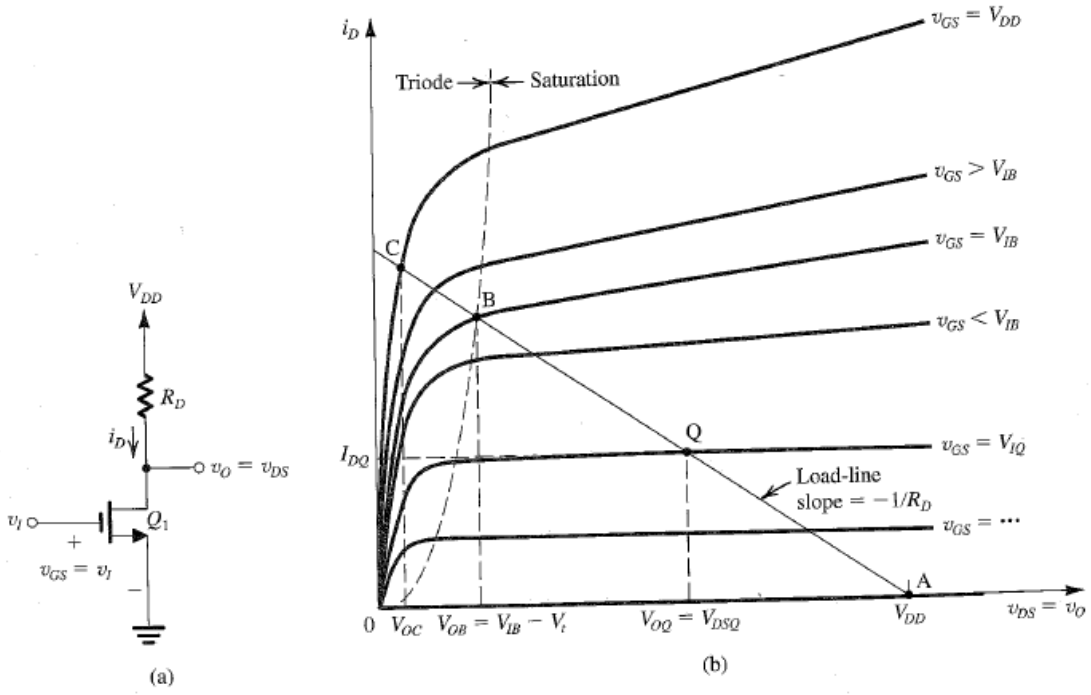
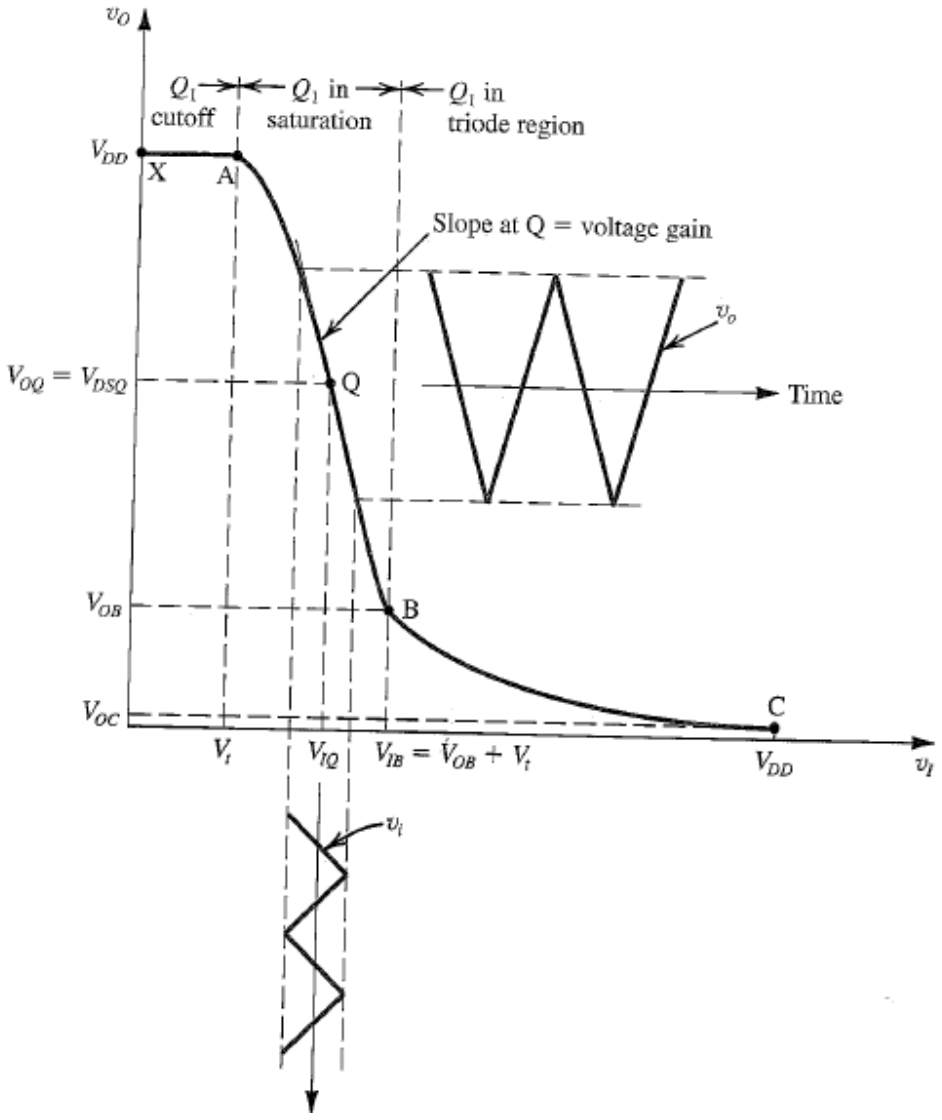


FIGURE 4.26 (a) Basic structure of the common-source amplifier. (b) Graphical construction to determine the transfer characteristic of the amplifier in (a).



circuit arises because when the circuit is viewed as a two-port network, the grounded source terminal is common to both the input port, between gate and source, and the output port, between drain and source. Note that although the basic control action of the MOSFET is that changes in v_{GS} (here, changes in v_I as $v_{GS} = v_I$) give rise to changes in i_D, we are using a resistor R_D to obtain an output voltage v_O,

$$v_O = v_{DS} = V_{DD} - R_D i_D \tag{4.35}$$

4.4.3 Operation as a Switch

When the MOSFET is used as a switch, it is operated at the extreme points of the transfer curve. Specifically, the device is turned off by keeping $v_I < V_t$ resulting in operation somewhere on the segment XA with $v_O = V_{DD}$. The switch is turned on by applying a voltage close to V_{DD}, resulting in operation close to point C with v_O very small (at C, $v_O = V_{OC}$). At this juncture we observe that the transfer curve of Fig. 4.26(c) is of the form presented in Section 1.7 for the digital logic inverter. Indeed, the common-source MOS circuit can be used as a logic inverter with the "low" voltage level close to 0 V and the "high" level close to V_{DD}. More elaborate MOS logic inverters are studied in Section 4.10.