11. P. Roblin, S. Samman, and S. Bibyk, "Simulation of Hot-Electron Trapping and Aging in nMOSFETs," *IEEE Trans. Electron Dev.*, **ED-35**, December 1988, p. 2229.

12. R.F. Pierret, *Modular Series on Solid-State Devices; Vol. IV - Field Effect Devices.*, R.F. Pierret and G.W. Neudeck, Eds., Addison-Wesley Publ. Co., Reading MA, 1983.

13. B.E. Deal, "Standardized Terminology for Oxide Charges Associated with Thermally Oxidized Silicon," *IEEE Trans. Electron Dev.*, **ED-27**, p. 606, (1980).

PROBLEMS

1. Explain why it is useful to study the behavior of the MOS-C prior to studying the behavior of the MOSFET.

2. Sketch the energy-band diagrams of a MOS-C with a heavily-doped *n*-polysilicon gate when (a) $V_{GB} = 0$, and (b) in the flat band condition when the Si substrate: (1) 1-Ω-cm *n*-type, and (2) 1-Ω-cm, *p*-type.

3. For an ideal MOS-C with $t_{OX} = 30$ nm and $N_A = 5x10^{15}$ cm^{-3}, find the applied voltage and electric field at the interface needed to: (a) make the silicon surface intrinsic; and (b) to bring about strong inversion.

4. Using Eq. 3-13c prove that for $n_{surf} = 10N_A$, φ_{surf} is only 58 mV greater than $2\varphi_B$.

5. Calculate the flatband voltage and the threshold voltage (a) in 1 Ω-cm *p*-type silicon and (b) in 1 Ω-cm *p*-type silicon. The MOS-Cs for each case have : (i) aluminum gate for which $q\varphi_m = 4.3$ eV, (ii) $t_{OX} = 100$ nm; and (iii) the oxide is free of charge except for a surface density of $Q_f = 5x10^{10}$ cm^{-2}. The electron affinity of silicon is 4.05 eV.

6. Use the "threshold" statement in SUPREM III to extract and plot V_T vs. V_{SB} for increments of 0.25V between 0 and 5V for the device specified in Prob. 4. Compare the zero bias ($V_{BS} = 0V$) value with the two corresponding values of V_T found in Prob. 4.

7. Consider Poisson's equation in the substrate of a MOS-C. Using the condition that there is charge neutrality in the substrate, show that the impurity doping terms $N_D - N_A$ can be written as n_i (exp (u_B) - exp ($-u_B$)] = 2 n_i sinh u_B , where u represents the potential normalized to (kT/q) and u_B is the substrate potential ($u_B = - q\varphi_B/kT$) in a *p*-type substrate.

8. Using the form derived in Problem 6, show that the Poisson equation can be written in terms of u as:

$$d^2u/dx^2 = (1/L_{Di})^2 [\sinh(u) - \sinh(u_B)]$$

where L_{Di} is the intrinsic Debye length

$$L_{Di} = (\varepsilon_{Si} k T/2q^2 n_i)^{1/2}$$

9. Calculate the maximum depletion layer depth in a MOS-C with a *p*-type substrate doped uniformly at $2x10^{15}$ cm^{-3} if T = 77K and 300K.

10. Consider the establishment of a p-type inversion layer in a MOS-C. Show that for a nondegenerate n-type semiconductor, the surface charge density Q_C needed to make $p_{surf} = n_o$ (the equilibrium electron concentration in the bulk) is given by

$$Q_C = \sqrt{2\varepsilon_{si}\, q\, N_D\, (\varphi_B)}.$$

CHAPTER 4

MOS TRANSISTOR DEVICE PHYSICS: PART 2

LONG-CHANNEL MOSFETS

A description of the physics of the MOS capacitor (MOS-C, chap. 3) provides a solid foundation for analyzing the behavior of the MOSFET. Hence, in this chapter the device physics of long-channel MOSFETs is undertaken. In chap. 5 we examine the behavior of short-channel (i.e., submicron) MOSFETs.

Long-channel MOSFETs are discussed first for the following reasons:

1. The derivation of accurate long-channel-MOSFET drain-current models can be achieved by solving both the *Poisson equation* and the *drift-diffusion current density equations* in one-dimensional form, whereas accurate modeling of short-channel MOSFET behavior requires the solution of these equations in two (or even three) dimensions. Thus, it is logical to undertake the simpler analysis first.

2. The long-channel drain-current MOSFET models derived from the solution of the 1-D device equations can be expressed in analytical form. Hence, these models are widely used in circuit-design applications. In addition, quantitative insight about the device physics of MOSFET operation can also be obtained by examining these models. On the other hand, physically based analytical models of short-channel MOSFET behavior do not arise from the solution of the two-dimensional device equations, and thus it is harder to gain quantitative insight into MOSFET operation from them.

3. The analytical expressions for I_D from two of the four long-channel MOSFET models derived in this chapter are the I_D expressions used in SPICE for simulating MOSFET behavior in IC circuit design applications.

4. As MOSFETs are scaled down, they eventually become prone to 'short-channel' effects. In most cases, however, these effects are considered detrimental. That is, it is the task of the device designer to suppress them when scaling-down MOSFET structures. The goal is to produce 'short-channel' MOSFETs with 'long-channel' behavior. Thus, it is important to define long-channel behavior and to describe the physics that produce such characteristics in MOSFETs. This understanding provides insight into the measures that must be employed to suppress unwanted short-channel effects.

Dell Ex. 1008C
Page 73

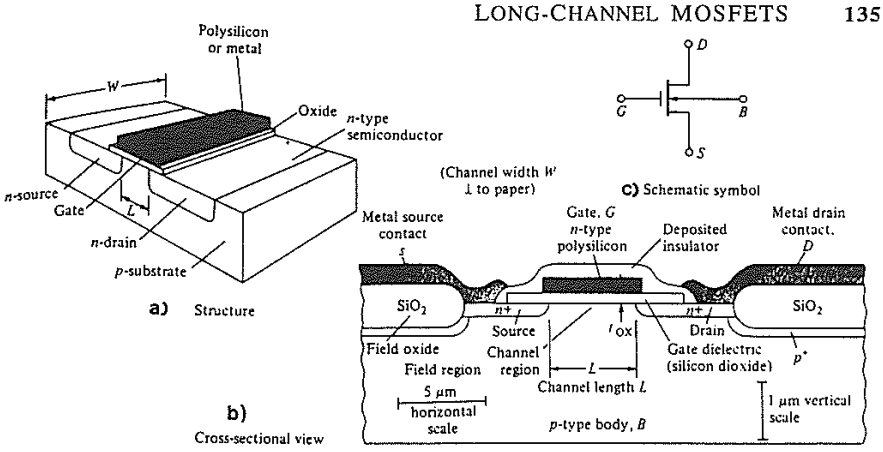


Fig. 4-1 (a) Structure of an MOS device. (b) Cross sectional view. (c) Schematic symbol.[1] From D. A. Hodges and H. G. Jackson, Analysis and Design of Digital Integrated Circuits, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission.

In this chapter the following topics dealing with MOSFETs are covered:

- **a.** Introduction to the terminology of MOSFETs and their behavior
- **b.** Effect of Applying a Bias to the Inversion Layer of the MOS-C
- **c.** Definition of the Long-Channel MOSFET and the Gradual-Channel Approximation
- **d.** Qualitative Description of Long-Channel MOSFETs
- **e.** Quantitative Models of dc Long-Channel MOSFET Behavior
  - **i.** Pao-Sah Model
  - **ii.** Charge-Sheet Model
  - **iii.** Variable-Depletion-Charge Model
  - **iv.** Square-Law Model
- **f.** Threshold-Voltage Control in Long-Channel MOSFETs
  - **i.** Body Effect
  - **ii.** Threshold-voltage adjustment by ion-implantation
- **g.** Subthreshold behavior of Long-Channel MOSFETs
  - **i.** Physics and modeling of subthreshold current flow
  - **ii.** Subthreshold swing
  - **iii.** Gate-induced leakage current (GIDL)

4.1 TERMINOLOGY OF MOSFET DEVICE STRUCTURES

Figure 4-1 shows the cross-section of the MOSFET, as introduced in chap. 3. The important difference between the MOS-C and the MOSFET is that in the latter a *pn*

junction exists at each side of the MOS-C channel region. Each of these junctions are also close enough to make contact with the inversion layer. By making external connections to each of these edge junctions, the MOSFET becomes a four-terminal device (with the terminals designated respectively as the *gate* G, *body* B, *source* S, and *drain* D). The MOS-C, on the other hand, remains a two terminal device (*gate* and *body*). The source and drain terminals of the MOSFET allow a bias V_{DS} to be applied across the ends of the channel region, giving rise to a *drain current* I_D.

As described in chap. 3, the inversion layer can be established by applying an appropriate bias to the gate. For the device shown in Fig. 4-1a, the induced charge in the channel inversion layer is *n*-type, and the device is known as an *n*-channel MOSFET (or simply an NMOS transistor). A *p*-channel MOSFET (or PMOS) is obtained by interchanging the *n* and *p* regions. Just as in the ideal MOS-C it is assumed that current in the gate oxide (*gate current,* I_G) is negligible. We also assume that the MOSFET is a symmetrical device, insofar as connecting the two junction terminals to a circuit (i.e., the drain current should remain the same if the two terminals are interchanged). Regardless of the way a MOSFET is connected, the terminal tied to the more positive voltage (in NMOS circuits) is defined to be the *drain*, and the other is the *source*.

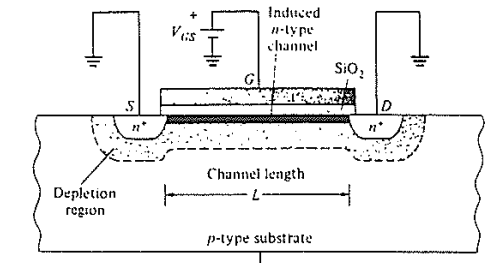
4.1.1 Basics of MOSFET Operation

In simplest terms, the operation of a MOS transistor involves the application of an input voltage to the gate electrode. This establishes an electric field perpendicular (or transverse) to the Si-SiO_2 interface in the channel region of the device (Fig. 4-2a). The conductance of the channel region (in which the device output current I_D flows) can be modulated by varying this electric field. Since an electric field is responsible for controlling the output current flow, such devices are termed *field-effect transistors* (FETs).

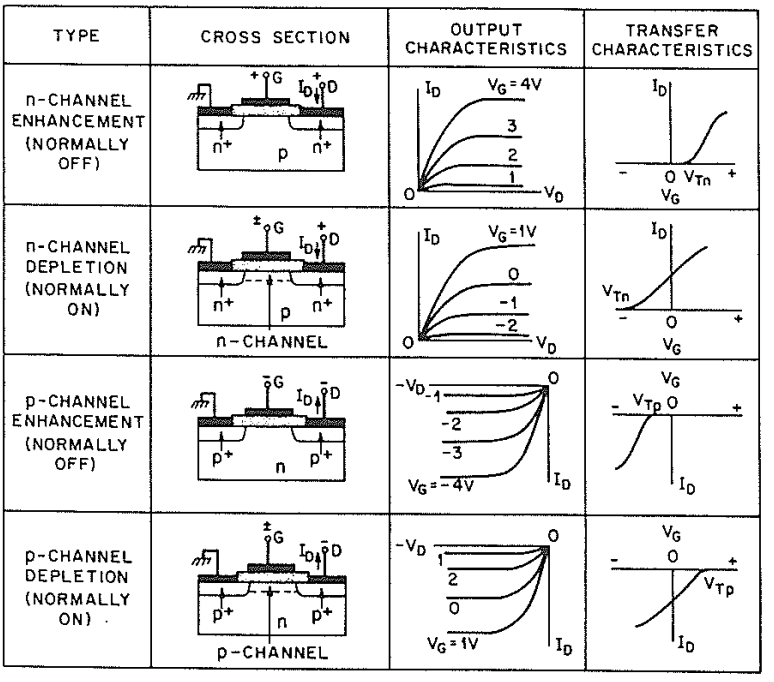
If no gate bias is applied, however, the circuit path between source and drain consists of two back-to-back *pn* junctions in series. In this case, if V_{DS} is applied (such that the drain and source *pn* junctions are reverse biased), I_D will consist of only the reverse-bias diode leakage current, which is normally negligible.

When positive bias is applied to an NMOS transistor gate, however, electrons will be attracted to the channel region and holes will be repelled (as is the case in the MOS-C). Once the positive gate voltage becomes strong enough to form an inversion layer, an *n*-type channel is formed that connects the source and drain regions (Fig. 4-2a). A drain current I_D can then flow if a voltage V_{DS} is applied between the source and drain terminals. In the simplest analysis we assume that the voltage-induced *n*-type channel does not form unless the voltage applied to the gate exceeds the threshold voltage V_T.

MOS devices such as those just described (in which no conducting channel exists when $V_{GS} = 0$), are referred to as *enhancement-mode* (or *normally OFF*) transistors (see Fig. 4-2b). With NMOS enhancement-mode transistors, a positive gate voltage, V_{GS}, greater than V_T must be applied to create the channel (or to turn them ON), while to turn ON PMOS enhancement-mode devices, a negative gate voltage (whose magnitude is $>V_T$) must be applied. Note that in NMOS transistors a positive voltage must also



a)



b)

Fig. 4-2 (a) Idealized NMOS cross section with positive V_{GS} applied showing depletion regions and induced channel. From D. A. Hodges and H. G. Jackson, Analysis and Design of Digital Integrated Circuits, Copyright, 1983 McGraw-Hill Book Co. Reprinted with permission. (b) Cross sections and output and transfer characteristics of four types of MOSFETs. From S.M.Sze, Semiconductor Devices - Physics and Technology, 1985, John Wiley & Sons. Reprinted with permission.

be applied to the drain to keep the drain-substrate reverse-biased, while in PMOS devices this voltage must be negative.

On the other hand, it is also possible to build MOS devices in which a conducting channel region exists when $V_{GS} = 0V$ (see also Fig. 4-2b), and such MOS devices are described as being *normally ON*. Since a bias voltage to the gate electrode is needed to deplete the channel region of majority carriers, and thus turn them OFF, such devices are also commonly called *depletion-mode MOSFETs*. NMOS depletion-mode devices require a negative gate voltage to be turned OFF, while corresponding PMOS devices require positive gate voltages.

4.2 BIASING THE INVERSION LAYER (BODY EFFECT)

Before undertaking a detailed discussion of MOSFET behavior, we first study the consequences of *applying a dc bias to the inversion layer of a MOS-structure*. This topic is actually an extension of the subject of MOS-C behavior, but also serves as an important prelude to the discussion of MOSFET operation. Consequently, when we discuss the operation of the MOSFET we will refer back to the results developed in this section.

We examine the effect of biasing the inversion layer from both the qualitative and quantitative perspectives. In the latter we will show how the expressions for the mobile charge concentrations in the inversion layer of the MOS structure are altered when a such bias is applied, as well as how the application of a bias to the inversion layer modifies the threshold voltage equation derived for the MOS-C.

Biasing the inversion layer is accomplished by the application of a dc voltage between the substrate and a diffused region adjacent to, and in contact with, the inversion layer. This implies that a modified MOS-C structure must be used. Specifically, a diffused region must be added to the edge of the MOS-C channel region (see Fig. 4-3a), and a terminal must be connected to this region (designated by the letter C). The result, as shown, is the *three-terminal MOS device* depicted in Fig. 4-3a.

If we choose the substrate (body or B) terminal as the reference, the voltage applied between terminal C and the substrate (body) is denoted by V_{CB}, and the voltage between the gate (terminal G) and substrate by V_{GB} (see Fig. 4-3a).* For simplicity we also assume an ideal *p*-type Si-substrate MOS structure (i.e., $\varphi_{ms} = 0$ and $Q_{tot} = 0$).

4.2.1 Qualitative Discussion of the Effect of Applying a Bias to the Inversion Layer of a MOS Structure

If a MOS-C with a *p*-type substrate is biased into inversion by a gate voltage V_{GB}, a *pn* junction exists between the surface and the bulk of the silicon (as a result of the region near the silicon surface having been inverted to *n*-type). If there is an adjacent *n-doped* region in contact with the *induced* *n*-type inversion layer, it is possible to apply a bias to the induced junction by applying a bias V_{CB} between terminal C and B (the substrate)

* As elsewhere, we use the convention that the second letter of the voltage subscript designates the reference terminal.

as shown in Fig. 4-3a. (Note that here we have a MOS structure in which the induced junction is contacted only at one end of the channel.*) We also assume that the polarity of V_{CB} is such that the induced *pn* junction is either unbiased or under reverse bias (i.e., V_{CB} in Fig. 4-3a is ≥0V). We will refer to V_{CB} as the *channel voltage.*

Before presenting the equations relevant to this structure, let us first consider the qualitative effect of applying various V_{CB} bias values.

4.2.1.1 Case When Both V_{GB} and V_{CB} = 0V. In Fig. 4-3b we show the two-dimensional potential φ versus position diagram for the three-terminal MOS structure when both V_{GB} and V_{CB} = 0V. We note that the channel region is in a flat-band condition and that a built-in voltage φ_{bi} exists across the n^+p (doped) junction. We also recall that even if $V_{GB} \neq 0V$, as long as V_{CB} remains at 0V, the structure is still in equilibrium, because I_G is assumed to be zero and there is no current in the n^+p junction if V_{CB} = 0V. Hence, this implies that when V_{CB} = 0V the channel region of this structure exhibits the same behavior as that of the MOS-C described in chap. 3.

4.2.1.2 Case When $V_{GB} \geq V_T$ and V_{CB} = 0V. If we now consider the case where V_{CB} = 0V, but $V_{GB} \geq V_T$, we encounter a situation as shown in Fig. 4-3c. That is, the channel region is now inverted, and a *field-induced junction* now exists between the *n*-type inversion layer and the underlying *p*-type substrate. Since a bias is not applied across the field-induced junction, it is in equilibrium and is again characterized by the same Fermi level as in the diffused junction. The surface depletion region in the field-induced junction under these conditions is also described by the equilibrium theory presented in chap. 3.

Note, however, that the potential difference between the inversion layer and the n^+ region is now much smaller than in the case when V_{GB} = 0V. Specifically, this potential difference is now ~(φ_{bi} - 2φ_B). This implies that the inversion layer is in "communication" with the n^+ region adjacent to it. The term "communication" means that once the inversion layer is formed and a bias V_{CB} is applied between terminals C and B, this bias will appear across the entire length of the field-induced junction (i.e., no longer just across the doped n^+p junction at the edge of the channel). Thus, *if an inversion layer exists in the channel, then V_{CB} does have an influence on φ_{surf}.*

For example, if V_{CB} is suddenly allowed to increase from 0V to some non-zero value (while V_{GB} is held fixed), electrons will be observed to flow out of the inversion layer through the C terminal. Why this happens, and what the effects are on the behavior of the MOS structure are described next.

4.2.1.3 Case When V_{GB} = 0V, But V_{CB}>0V. If we next apply a non-zero V_{CB} bias while keeping V_{GB} = 0V, equilibrium will cease to exist in the three-terminal structure. This is because V_{CB} causes a small reverse-bias leakage current to flow in the n^+p junction. (Although this reverse-bias leakage current is extremely small, its

* However, in a MOSFET, contact to the inversion layer can be made at both ends of the induced junction since a diffused region exists at each end of the channel. In the MOSFET these two regions are called the source and drain, and the terminals are designated by the abbreviations S and D.

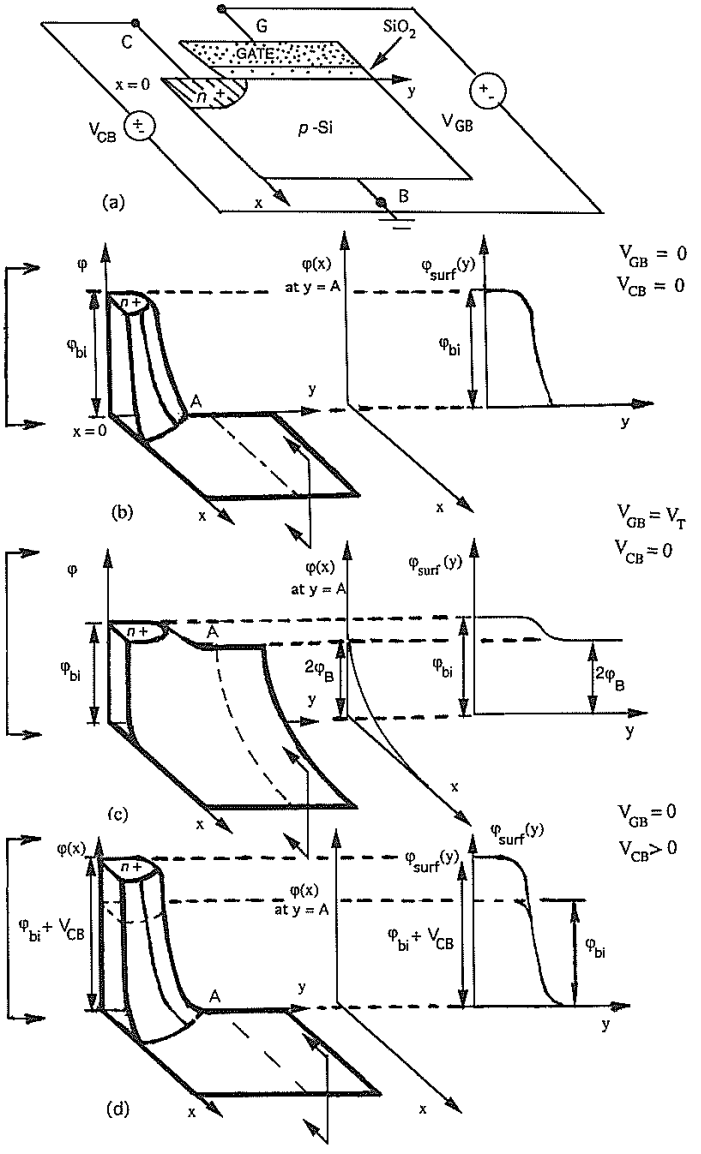


Fig. 4-3 (a) Applying a bias V_{CB} to an induced channel beneath the gate oxide by applying this bias V_{CB} between terminal C and the substrate of a three-terminal MOS structure. Two-dimensional potential φ versus position diagrams in the three terminal MOS structure when: (b) both V_{GB} and $V_{CB} = 0$; (c) $V_{GB}>V_T$ and $V_{CB} = 0V$; (d) $V_{GB} = 0V$ and $V_{CB}>0V$.

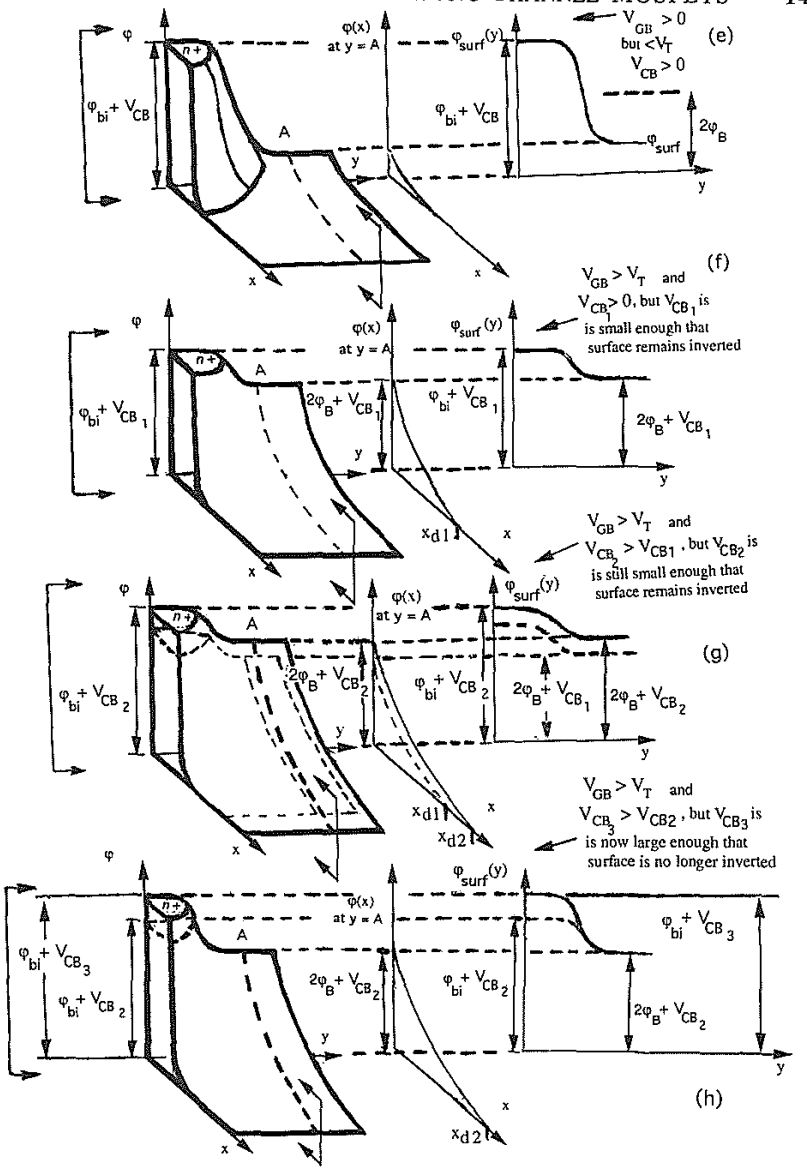


Fig. 4-3 (cont.) (e) $V_{GB}>0$, but is $<V_T$, and $V_{CB} >0$; (f) $V_{GB}>V_T$ and $V_{CB}>0$, but $V_{CB} = V_{CB1}$, which we assume is small enough so that inversion still exists. (g) V_{GB} is the same as in part (f), but $V_{CB} = V_{CB2}$, where $V_{CB2}>V_{CB1}$, but V_{CB2} is still small enough to keep surface inverted. (g) $V_{GB} = V_T$, $V_{CB} = V_{CB3}$ (large enough to suppress the channel inversion layer).

existence implies that the MOS structure is no longer in equilibrium because some of the carrier energy now depends on an externally applied force, namely the extra electric field created in the depletion region by V_{CB}.) However, if V_{GB} remains fixed at 0V, V_{CB} is dropped entirely across the n^+p junction. Hence, the surface potential along the channel is not altered from when V_{CB} = 0V. Instead, only the potential difference between the n^+ and p (substrate) regions is increased by V_{CB} (as shown in Fig. 4-3d), leaving the channel region to the right of the n^+p junction in a flat-band state. In addition, Fig. 4-3d indicates that *if no inversion layer is present, changing V_{CB} has negligible impact on the surface potential φ_{surf} along the channel.* This last statement also applies to the case illustrated in Fig. 4-3e (i.e., V_{GB}>0V, but is less than V_T, and V_{CB} is increased from 0V to some value greater than 0V). Here we observe that there is still no impact on φ_{surf} as long as no inversion layer is present in the channel.

4.2.1.4 Case When $V_{GB} \geq V_T$ and V_{CB}>0V. Let us now consider the case when $V_{GB}>V_T$ and a reverse-bias voltage V_{CB} is applied to the induced *pn* junction (such that the surface remains inverted even after V_{CB} is applied). Since the surface is inverted, the potential in the inversion region is very close to the potential value in the diffused region, as shown in Fig. 4-3f (and as exists in the case shown in Fig. 4-3c). That is, a point at the surface is at about the same potential (with respect to the substrate deep in the bulk) as a point in the n^+ region. Therefore, the surface is about as "attractive" for electrons as is the n^+ region. Consequently, electrons have no reason to prefer only the n^+ region as opposed to the "surface" at the right. As a result, the n^+ region can "communicate" with the inversion layer (implying that a bias applied to the n^+ region is also applied to the inversion layer). In addition, just as in the case of a MOS-C, the value of the surface potential φ_{surf} remains clamped at the value it has at the onset of strong inversion (i.e., for the case of Fig. 4-3f, at a value of $\varphi_{surf} \sim 2\varphi_B + V_{CB1}$). That is, as V_{GB} is increased beyond V_T, φ_{surf} remains unchanged, but the mobile charge concentration in the inversion layer continues to increase.

We also recall from *pn* junction theory that if a reverse-bias voltage is applied to a previously unbiased *pn* junction, the junction depletion-region width will increase. Thus, the depletion region on the *p*-side of the induced junction will also increase if a reverse-bias is applied between terminals C and B. However, if V_{GB} is kept fixed (such that $V_{GS} \geq V_T$) as the reverse-bias voltage V_{CB} is increased (see Fig. 4-3g), the charge on the gate terminal will remain the same as when V_{CB} = 0V. Therefore, the *total* charge in the semiconductor must also remain unchanged to maintain balance with the unchanged value of charge on the gate. However, since the application of V_{CB} widens the depletion region (thereby uncovering more charge in this region of the induced junction), fewer electrons are needed in the inversion layer to balance the unchanged gate charge. Consequently, electrons must exit the inversion layer, resulting in a lower level of inversion. (Looking at it in another way, the increased positive potential at terminal C will attract electrons from the inversion layer, which will start flowing toward the n^+ region, and from it into the top terminal of the voltage source.) However, if a sufficiently large value of V_{CB} is applied (while holding V_{GB} fixed), this effect can entirely extinguish the charge in the inversion layer. In other words, a gate voltage

capable of causing inversion in the three-terminal MOS structure when $V_{CB} = 0V$ may no longer be able to sustain inversion if a non-zero value of V_{CB} is applied.

As also shown in Figs. 4-3f and g, as long as the surface remains in inversion as V_{CB} is increased, the depletion region under the field induced junction increases (from x_{d1} in Fig. 4-3f to x_{d2} in Fig. 4-3g). This means that the bands are bent more, and thus φ_{surf} also increases. This supports the statement that as V_{CB} increases, the n^+ region communicates this change in V_{CB} to the field-induced junction. However, once a sufficiently large number of electrons depart from the inversion layer so that it essentially disappears, further increases in V_{CB} no longer influence the channel region, and the depletion region no longer widens (as shown in Fig. 4-3h), and remains fixed at a value dependent on V_{CB} (e.g., at x_{d2}, see Figs. 4-3g and h) as expressed by Eq. 4-8.

If a non-zero V_{CB} is applied and it is desired to re-establish inversion, φ_{surf} must be increased to a value larger than that needed when $V_{CB} = 0$. Specifically, at the onset of inversion, it is necessary to increase V_{GB} so that $q\varphi_{surf} \sim q(V_{CB} + 2\varphi_B)$. Note that the effect just described can also be exploited as a technique for raising the threshold voltage of a MOSFET, as will be discussed in section 4.6.2.

4.2.1.5 Summary of the Behavior of the Three-Terminal MOS Structure and the Significance of Such Behavior on the Operation of the MOSFET. In summary, it is possible to establish "communication" with the channel region (i.e., an external bias voltage can be applied to the channel region of a MOS structure) if an inversion layer has been induced there. On the other hand, if the channel region is not inverted, the application of a bias voltage to a diffused region adjacent to the channel region will have no effect on the surface potential in the channel. Such "communication" (or lack of) has the following impact on the behavior of a MOSFET:

1. If the gate voltage applied to a MOSFET is less than the threshold voltage, the device is said to be operating in the subthreshold regime. From our discussion, this means that if a bias is applied to the drain (V_{DS}) of a MOSFET being operated in subthreshold, this voltage will be dropped entirely across the diffused n^+p junction (i.e., the drain-substrate junction), and it will have no impact on φ_{surf}. Hence, φ_{surf} will be determined by the gate potential, which is constant everywhere on the gate. Consequently, the potential drop along the y direction at the surface of the channel will be essentially zero. We will use this information when we discuss subthreshold current flow in MOSFETs in section 4.7.

2. If the channel is inverted, and V_{CB} is increased (and the channel remains in inversion for the range of applied V_{CB} values), the depletion region of the field-induced channel increases. If V_{GB} is held fixed, the number of electrons in the inversion layer will decrease as V_{CB} is increased. At some value of V_{CB} the inversion layer charge will entirely disappear. These effects occur when a drain bias V_{DS} is applied to a MOSFET. That is, for a fixed gate voltage (greater than the threshold voltage value), as V_{DS} is increased, the depletion region of

will correspondingly decrease, making it less inverted. Thus, in a MOSFET operating in inversion and with $V_{DS}>0$, the depletion region near the drain will be wider than at the source (i.e., assuming the source is the common terminal).

3. "Communication" with the channel also permits the diffused region to become an additional source (or sink) of mobile carriers to the channel (i.e., in addition to the carriers thermally generated in the depletion regions). As noted, two such diffused regions exist in a MOSFET, one which supplies mobile carriers (the source), and the other which removes them (the drain). Specifically, if a drain bias is applied to a MOSFET in inversion, the channel carriers will be attracted to and move into the drain. These are replenished by carriers entering the channel from the source. The resultant flow of charge in the channel constitutes a current in the drain terminal I_D. Since the channel is not a perfect conductor, this current also gives rise to a potential drop along the direction of the channel.

4.2.1.6 Body Effect. Until now we have chosen the substrate to be the reference terminal in the three-terminal MOS structure. However, when MOSFETs are used in circuits, the source terminal S (corresponding here to terminal C), is normally chosen to be the reference. Hence, *body effect* is the term used to describe the increase in the threshold voltage of a MOSFET (with reference to the source) upon the application of a reverse bias to the inversion layer. If, as shown in Fig. 4-4b, we choose terminal C to be the reference terminal (instead of the body terminal), the bias voltages are now applied between the gate and C (V_{GC}), and between the body and C (V_{BC}). In this case, for a given value of V_{BC}, a value of V_{GC} greater than that needed when $V_{BC} = 0V$ must be applied to the MOS structure to establish strong inversion. In section 4.2.2.2 we will derive an expression that determines the increase in V_T as a function of V_{BC}.

4.2.2 Quantitative Description of Effects of Biasing the Inversion Layer of MOS Structures

For the three-terminal MOS structure introduced above, it is possible to develop the quantitative relationships in a general manner for all cases of inversion. However, these can also be simplified somewhat for the more limited (but still most widely used form) of the strong inversion case. Note that in deriving these expressions we also include V_{FB}, which accounts for the work-function difference and oxide charge that can exist in 'real' MOS structures.

4.2.2.1 General Inversion Case. In the general inversion case (which covers weak as well as strong inversion), five equations characterize the phenomena in the channel region (i.e., the Si region in the vicinity of the Si-SiO_2 interface to the right of the n^+ junction in Fig. 4-3a) of three-terminal MOS structure (as they do for the MOS-C). If we invoke the charge-sheet and depletion approximations, four of these equations are identical to those for the expressions derived for the charge-sheet model of the MOS-C, and are repeated here for the readers convenience:

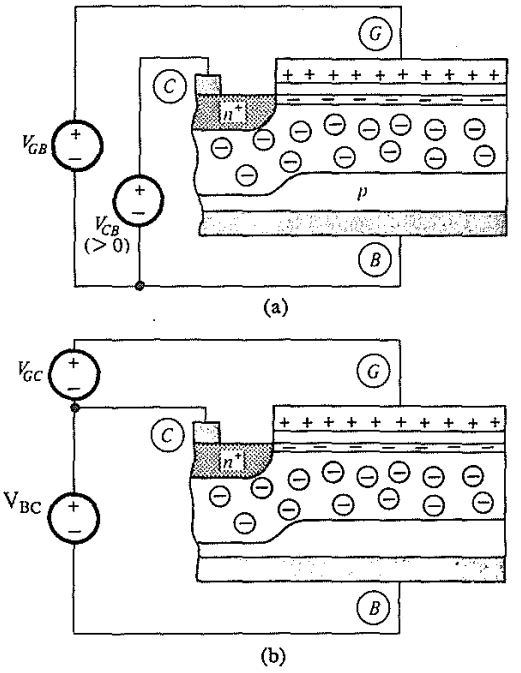


Fig. 4-4 Bias notation when: (a) terminal B is the reference terminal; (b) terminal C as the reference terminal instead of terminal B.[5] Reprinted with permission of McGraw-Hill Co.

$$V_{GB} = \varphi_{surf} + V_{FB} + \varphi_{ox} \qquad (3\text{-}10) \qquad (4\text{-}1a)$$

$$Q_C = Q_I + Q_B \qquad (3\text{-}20) \qquad (4\text{-}1b)$$

$$Q_C = -C_{ox}\varphi_{ox} \qquad (3\text{-}19) \qquad (4\text{-}1c)$$

$$Q_B = -\gamma C_{ox}\sqrt{\varphi_{surf}} \qquad (3\text{-}29) \qquad (4\text{-}1d1)$$

$$= [2qN_A\kappa_{Si}\varepsilon_o(\varphi_{surf})]^{1/2} \qquad (4\text{-}1d2)$$

The fifth equation is a modified form of Eq. 3-53, in which we replace $2\varphi_B$ by $(2\varphi_B + V_{CB})$ to get

$$Q_I = \sqrt{2q\,N_A\kappa_{Si}\varepsilon_o}\left(\sqrt{\varphi_{surf} + \varphi_t\, e^{(\varphi_{surf} - 2\varphi_B - V_{CB})/\varphi_t}} - \sqrt{\varphi_{surf}}\right) \qquad (4\text{-}2)$$

Then, combining Eqs. 4-1a, 4-1c, and 4-1b we get

$$V_{GB} = \varphi_{surf} + V_{FB} - [Q_I(\varphi_{surf}) + Q_B(\varphi_{surf})]/C_{ox} \qquad (4\text{-}3)$$

Next, using Eqs. 4-1d1 and 4-2 in Eq. 4-3, V_{GB} can be written as

$$V_{GB} = \varphi_{surf} + V_{FB} + \gamma\sqrt{\varphi_{surf} + \varphi_t e^{(\varphi_{surf} - 2\varphi_B - V_{CB})/\varphi_t}} \qquad (4 - 4)$$

Now, if V_{GB} and V_{CB} are given, Eq. 4-4 can be solved iteratively for φ_{surf}. This value can then be inserted into Eq. 4-2 to calculate Q_I, or into Eq. 4-1d2 to calculate Q_B.

An expression for Q_I in terms of φ_{surf} and V_{GB} can also be derived from the above equations. That is, by inserting Eqs. 4-1c and 4-1d1 into Eq. 4-1b above, we get

$$- C_{ox}\varphi_{ox} = Q_I + \gamma C_{ox}\sqrt{\varphi_{surf}} \qquad (4 - 5)$$

Then, solving for φ_{ox}, and substituting the result into Eq. 4-1a, we obtain

$$Q_I = - C_{ox} (V_{GB} - V_{FB} - \varphi_{surf} - \gamma \sqrt{\varphi_{surf}}) \qquad (4 - 6)$$

4.2.2.2 Strong Inversion Case. For the three-terminal structure biased to strong inversion, the above general analysis can be simplified. That is, with $V_{CB} = 0V$, φ_{surf} of the three-terminal MOS structure at the onset of inversion is $2\varphi_B$. With a reverse bias applied to the induced junction, the surface potential at the onset of strong inversion becomes

$$\varphi_{surf}(inv) \approx 2\varphi_B + V_{CB} \qquad (4 - 7)$$

Since the applied V_{CB} extends the range of gate voltages under which the surface region remains in depletion, d_{max} is now larger. The expression for d_{max} for the case of an applied V_{CB} is (assuming the depletion approximation):

$$d_{max} = [2\kappa_{si}\varepsilon_o(2\varphi_B + V_{CB})/qN_A]^{1/2} \qquad (4 - 8)$$

and the depletion layer charge (per unit area), Q_{Bmax}, when $d = d_{max}$ is expressed as

$$Q_{Bmax} = [2q\kappa_{si}\varepsilon_o N_A (2\varphi_B + V_{CB})]^{1/2} \qquad (4 - 9)$$

The expression for the inversion layer charge per unit area Q_I in strong inversion in terms of V_{GB} and V_{CB} is

$$Q_I = - C_{ox} [V_{GB} - V_{FB} - 2\varphi_B - V_{CB} - \gamma (2\varphi_B + V_{CB})^{1/2}] \qquad (4 - 10)$$

Assuming again that Q_I is negligible at the onset of inversion, the total charge per unit area in the semiconductor is given by Q_B. Hence, the gate-substrate bias at the onset of strong inversion is

$$V_{GB} = V_{FB} + \varphi_{surf} + \varphi_{ox} = V_{FB} + (2\varphi_B + V_{CB}) + \gamma (2\varphi_B + V_{CB})^{1/2} \qquad (4 - 11)$$

Noting that $V_{GC} = V_{GB} - V_{CB}$, the threshold voltage, referenced to terminal C, is given by the expression

$$V_T\,(\text{with ref. to term C}) = V_{T(ud,lc)} = V_{FB} + 2\varphi_B + \gamma\,(2\varphi_B + V_{CB})^{1/2} \quad (4\text{ - }12)$$

Note that the term γ, which was defined in Eq. 3-60a as $[\sqrt{2qN_A\kappa_{si}\varepsilon_o}]/C_{ox} = [\sqrt{2qN_A\kappa_{si}\varepsilon_o}]t_{ox}/\varepsilon_{ox}$ is a factor which determines how much V_T will increase when V_{CB} is increased. Hence, γ has been given the name *body effect coefficient.* By inspecting the expression for γ we note that it predicts the body effect is stronger for heavier substrate dopings and/or thicker gate oxides. Conversely, the body effect can be made smaller by using lower doping concentrations and/or thinner gate oxides. Figure 4-5 shows the threshold voltage shift versus substrate bias for various γ values.[26]

4.3 DEFINITION OF TERM: LONG-CHANNEL MOSFET

It is appropriate at this point to provide a detailed definition of the term *long-channel MOSFET.* From a qualitative perspective, long-channel MOSFETs are assumed to have channel regions sufficiently long and wide so that "edge" effects along the four sides of the channel can be neglected. Insofar as the channel length is concerned, we assume the channel length L is much longer than the sum of the source and drain

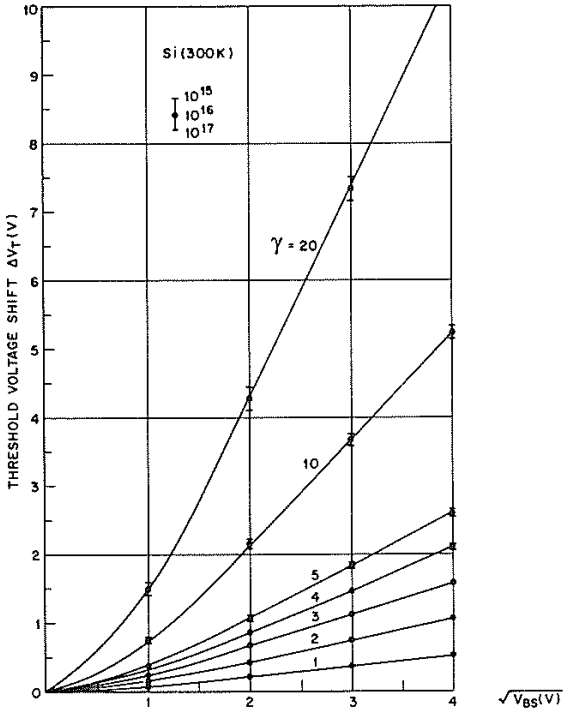


Fig. 4-5 Threshold voltage shift versus substrate reverse bias for various γ values.[2] Reprinted with permission of John Wiley & Sons.

depletion-region widths, or $L >> (d_{source} + d_{drain})$. The channel width, on the other hand is assumed to be much greater than the gate depletion-region depth, d (or Z>>d).

Even when the above criteria are well-satisfied, however, it is apparent that the electric fields and potential in the MOSFET still vary in at least two dimensions.* It seems that to accurately model device characteristics, the semiconductor device equations in their two-, or even three-dimensional forms must be solved. But in modeling long-channel MOSFET behavior this complexity can in fact be circumvented. That is, in such structures the solutions of the semiconductor device equations in their one-dimensional form yield drain-current device characteristics that are quite close to those obtained from measurements of actual long-channel MOSFETs.

Why is it possible to treat long-channel MOSFET behavior as a one-dimensional problem and still obtain relationships that give accurate estimates of I_D as a function of the applied the terminal voltages? First, consider the case of the MOSFET in which no drain bias is applied, and V_{GB} is the only external bias. We know that V_{GB} produces an electric field perpendicular to the device surface. Now, if $L >> (d_{source} + d_{drain})$ the junctions extend over such a small fraction of L that the $\mathcal{E}$-fields of these junctions have negligible impact on the $\mathcal{E}$-field due to the gate. Likewise, insofar as the channel width is concerned, if Z >>d the extension of the gate depletion region beyond the sides of the channel along which there is no source or drain hardly alters the total ionic charge uncovered by the gate voltage. Thus, the fringing field effect in the z direction can also be ignored. If both of these conditions are met, we can approximate the $\mathcal{E}$-field in a long-channel MOSFET when V_{DS} = 0V as being perpendicular to the surface everywhere along the channel. In this case, the one-dimensional analysis we applied to the MOS-C can also be used to determine Q_I, $\mathcal{E}$, φ and V_T in a long-channel MOSFET.

Next, if a drain bias V_{DS} is simultaneously applied, we assume that in long-channel MOSFETs the *gradual-channel approximation* is valid. This approximation assumes that the two aspects of drain current flow can be treated as strictly one-dimensional, independent problems. The two aspects are: (1) that the gate field acts approximately vertically to induce the channel (i.e., the mobile charge concentration that participates in I_D), and; (2) that the drain-to-source voltage acts approximately horizontally to move these mobile charges, and thus to drive the current I_D in the channel. The assumption that these two one-dimensional effects are independent of one another permits the results of the MOS-C analysis (shown to be valid in the case of a MOSFET with no drain bias), to also be applied when deriving the I_D-V_{DS} relationship of the MOSFET. Since the gradual-channel approximation is such a key simplification invoked when developing the models of long-channel MOSFET behavior, we need to examine it more rigorously.

* This is the case whether or not a drain current flows in the MOSFET. That is, if a gate bias is applied, but I_D = 0, there will still be electric fields arising from the source and drain junctions at the edges of the channel. If a drain current flows, it is obvious that there must be a component of the electric field along the y-direction. In fact, each of these cases should be examined separately when considering our definition in detail.

4.3.1 The Gradual Channel Approximation (GCA)

If a drain bias V_{DS} is applied to a MOSFET in inversion, drain current I_D flows. Owing to the channel's finite resistance and the passage of current, a potential variation occurs along the channel length (i.e., if the source is grounded, the channel potential near the source is close to ground, whereas near the drain it is close to V_{DS}). This causes the inversion-layer (mobile) charge density to decrease with position toward the drain because the gate is assumed to be an equipotential region (i.e., the voltage along the gate is invariant with position y along the channel), making the gate-to-channel voltage decrease with position toward the drain. Since I_D must remain constant at all points in the channel under dc biasing conditions, the electric field parallel to the flow of I_D ($\mathcal{E}_y$) must increase toward the drain to offset the decrease in mobile charge density. Thus, a gradient in the $\mathcal{E}_y$ value must exist in the channel (i.e., $\partial\mathcal{E}_y/\partial y \neq 0$). The *gradual-channel approximation* (GCA), however, assumes that this gradient $\partial\mathcal{E}_y/\partial y$ is much smaller than the $\mathcal{E}$-field. gradient in the x-direction, $\partial\mathcal{E}_x/\partial x$ (i.e., $\partial\mathcal{E}_x/\partial x >> \partial\mathcal{E}_y/\partial y$).

The two-dimensional form of the Poisson equation with $\partial^2\varphi/\partial x^2$ and $\partial^2\varphi/\partial y^2$ replaced respectively by $-\partial\mathcal{E}(x,y)/\partial x$ and $-\partial\mathcal{E}(x,y)/\partial y$) can be written as:

$$\frac{\partial\mathcal{E}(x,y)}{\partial x} + \frac{\partial\mathcal{E}(x,y)}{\partial y} = \frac{\rho(x,y)}{\kappa_{si}\,\varepsilon_o} \qquad (4 - 13)$$

If the GCA approximation is valid, the $\partial\mathcal{E}(x,y)/\partial y$ term in Eq. 4-13 can be neglected. Then, Eq. 4-13 can be approximated as a one dimensional differential equation with the form:

$$\frac{d\mathcal{E}_x(x)}{dx} \approx \frac{\rho(x,y)}{\kappa_{si}\,\varepsilon_o} \qquad (4 - 13a)$$

The GCA is generally valid when the channel of the MOSFET is long enough (e.g., L ~2 μm or longer), in which case the electric field strength along the y-direction $\mathcal{E}_y$ changes gradually from a small value near the source to values of the order of V_{DS}/L near the drain. Thus, neglecting the $\partial\mathcal{E}(x,y)/\partial y$ term in such devices is appropriate, and the one-dimensional form of the Poisson equation is sufficiently accurate for calculating Q_I used in the expressions for I_D. However, at large drain biases, the GCA always becomes invalid (even in long-channel MOSFETs), at least in a certain section of the channel near the drain where the gradient of the longitudinal field $\mathcal{E}_y$ becomes comparable to or larger than the gradient of the transverse field $\mathcal{E}_x$. In such cases, as will be described later, the drain current then becomes nearly independent of drain bias, and the MOSFET is said to be in *saturation*. This implies that under conditions of bias that cause saturation, the I_D behavior of long-channel MOSFETs can no longer be adequately modeled by invoking the GCA along the entire length of the channel. Instead, modifications to the long-channel models must be made (as described in chap. 5).

The importance of being able to invoke the gradual channel approximation should be emphasized, as this is the basis of the derivation of the long-channel MOSFET I_D-V_{DS} models. That is, by invoking the GCA, the problem of having to treat the semiconductor equations in their two-dimensional form to calculate I_D in a MOSFET is

avoided. Instead, the task is broken down into two one-dimensional analyses, each independent of the other. As a result, I_D is determined by the superposition of two one-dimensional solutions. Namely, by assuming that the *gradient* of the "horizontal" field in the channel is negligible, the one-dimensional *Poisson equation* can be solved to get the total charge in the silicon at each point y along the channel (i.e., the sum of the inversion layer charge Q_I [see Fig. 3-8] and the ionic charge Q_B in the depletion region). This means that the same formulas derived for the MOS-C are also applicable for calculating the local values of $Q_I(y)$ and $Q_B(y)$ in the long-channel MOSFET (provided, of course, that the effect of the channel voltage caused by the drain-source bias is also included in the MOS-C formulas). Then we note that within the conducting channel of a MOSFET the current flows almost exclusively in the y-direction. Thus, we can ignore the fact that there must be a slight x-component in the "horizontal" $\mathcal{E}$-field. This allows us to solve the one-dimensional *drift-diffusion current-density equation* in the y direction for I_D as a function of the terminal voltages [using the local $Q(y)$ values obtained from the solution of the Poisson equation in this current-density equation].

Note, however, that if the GCA is not valid, but it is still invoked in order to enable one-dimensional analysis to be used for calculating Q_I, it is unlikely that accurate values of I_D will be yielded This is the situation that arises in so-called short-channel devices (as described in chap. 5). In such cases, to obtain good agreement between calculated and measured values of I_D, the two-dimensional forms of the semiconductor equations must be solved (usually numerically).

4.3.1.1 Verifying the Validity of the GCA as a Function of the Geometrical Factors of the MOSFET Device Structure and Bias Conditions. The validity of the GCA for a particular MOSFET structure can be checked by making rough estimates of $\partial\mathcal{E}_x/\partial x$ and $\partial\mathcal{E}_y/\partial y$ values under given bias conditions. Since we are most interested in I_D when the MOSFET is in strong inversion, we will establish expressions that allow the GCA to be tested if the MOSFET is being operated in strong inversion.*

We begin this derivation by recalling that under strong inversion the inversion layer charge density Q_I is much larger than the ionic charge density Q_B in the depletion region (see section 3.3.3), and that the inversion layer is very thin (~30-300Å). Thus, in strong inversion $\partial\mathcal{E}_x/\partial x$ in the silicon is greatest in the inversion layer region (see Fig. 4-6). Now $\partial\mathcal{E}_x/\partial x$ can be estimated by approximating the vertical electric field value at the bottom edge of the inversion layer as zero. Using this assumption implies that $\partial\mathcal{E}_x/\partial x \sim [\mathcal{E}_{si}(x=0) - 0]/t_{inv} = [\mathcal{E}_{si}(x=0)]/t_{inv}$, where t_{inv} is the effective inversion layer

* Note that in weak inversion the surface potential $\varphi_{surf}(y)$ along the channel in long-channel MOSFETs is almost constant (see section 4.7). Thus, $\partial\mathcal{E}_y/\partial y$ is very small, implying that $\partial\mathcal{E}_x/\partial x >> \partial\mathcal{E}_y/\partial y$, because $\mathcal{E}_x$ drops from its value at $x = 0$ to zero at the bottom edge of the channel depletion region, which is a relatively small distance. Thus, in long-channel MOSFETs the GCA is also valid in weak inversion. However, in short-channel MOSFETs, this will no longer necessarily be the case, as we shall see in chap. 5.

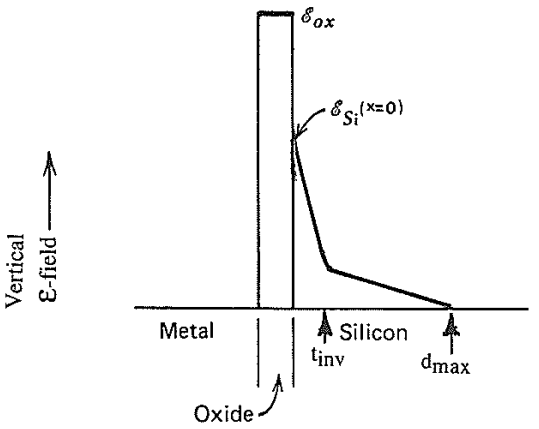


Fig. 4-6 The vertical electric field in a MOS-C (or MOSFET channel) when a large enough gate bias to cause inversion is applied. Note that since $Q_I > Q_B$ in inversion, that the gradient of the $\mathcal{E}$-field over the inversion layer thickness t_{inv} is larger than that over the depletion region width d_{max}, and we can therefore approximate the $\mathcal{E}$-field value at the bottom of the inversion layer thickness as being equal to zero for the purposes of estimating t_{inv}.

thickness. From Eq. 3-15, $\mathcal{E}_{si}(x=0) = \varepsilon_{ox}\mathcal{E}_{ox}/\kappa_{si}\varepsilon_o$, and from Eq. 3-14, $\mathcal{E}_{ox}$ is $\sim(V_{GS} - V_T)/t_{ox}$. Thus,

$$\frac{\partial \mathcal{E}_x}{\partial x} \approx \left(\frac{\varepsilon_{ox}}{\kappa_{si}\varepsilon_o t_{inv}}\right) \frac{(V_{GS} - V_T)}{t_{ox}} \qquad (4 - 14a)$$

Now, according to Eq. 3-12a, the concentration of electrons at some point in the semiconductor at equilibrium is exponentially dependent on the potential, with the exponential constant being $1/\varphi_t = kT/q$. This implies that the majority of the charges is contained within a distance from the surface over which φ drops by kT/q (i.e., since Eq. 3-12a predicts that if the electron concentration at the surface [x = 0] is n_{surf}, then at the point where the potential drops below φ_{surf} by an amount kT/q, n_{surf} will have dropped to a value 1/e of its value at x = 0, or to $0.37n_{surf}$). Thus, we assume that t_{inv} is roughly equal to that distance below x = 0. If we also assume that the vertical electric field $\mathcal{E}_x$ in the inversion layer is roughly the potential difference across the inversion layer $\Delta\varphi_{inv}$ divided by its thickness t_{inv} ($\mathcal{E}_x \sim \Delta\varphi_{inv}/t_{inv}$) we get $t_{inv} \sim (kT/q)/\mathcal{E}_{si}(x=0)$. Thus, substituting this expression for t_{inv} into Eq. 4-14a we finally obtain the following expression for roughly estimating $\partial\mathcal{E}_x/\partial x$:

$$\frac{\partial \mathcal{E}_x}{\partial x} \approx \left(\frac{\varepsilon_{ox}}{\kappa_{si}\varepsilon_o}\right)^2 \frac{q\,(V_{GS} - V_T)^2}{kT\, t_{ox}^2} \qquad (4 - 14b)$$

If we also assume that the value of $\mathcal{E}_y$ varies gradually from a small value near the source to values of the order of V_{DS}/L near the drain, $\partial\mathcal{E}_y/\partial y$ can be roughly estimated as $\partial\mathcal{E}_y/\partial y \sim V_{DS}/L^2$. Thus, if we assume that the GCA is valid when the ratio $(\partial\mathcal{E}_x/\partial x)/\partial\mathcal{E}_y/\partial y$ is large, then it is also valid if the following inequality is large

$$\frac{\partial\mathcal{E}_x/\partial x}{\partial\mathcal{E}_y/\partial y} = \left(\frac{\varepsilon_{ox}\, L}{\kappa_{si}\varepsilon_o}\right)^2 \frac{q\,(V_{GS} - V_T)^2}{kT\, t_{ox}^2\, V_{DS}} >> 1 \qquad (4 - 15)$$

For a MOSFET at 300K with $L = 1.0\ \mu$m, $t_{ox} = 300$Å, $V_{GS} - V_T = 0.5$V, and $V_{DS} = 0.5$V, the left-hand side of the inequality (4-15) is ~2300, indicating that the GCA is a very good approximation for such a MOSFET. This also implies that the GCA can be valid even in submicron MOSFETs, provided that $V_{GS} - V_T$ is not too small.

However, as we will see in chap. 5, $\mathcal{E}_y$ is quite non-uniform along the length of the channel at high values of V_{DS}, becoming much stronger near the drain. Furthermore, from Eq. 4-10 we note that the inversion layer charge density can be approximated as $Q_I \sim [V_{GS} - V_T - V(y)]$, if we replace V_{GB} with V_{GS}, and V_{CB} with $V(y)$ [where $V(y)$ is the channel voltage induced by V_{DS}]. However, this approximation is only valid when Q_I is positive, i.e., when $V_{GS} - V_T > V(y)$. Thus, if V_{DS} exceeds $V_{GS} - V_T$ (i.e., the condition which occurs in the region of the channel near the drain when the MOSFET is operated in saturation), the GCA becomes invalid in this region of the channel. [Because, when $V(y)$ is greater than $V_{GS} - V_T$, this implies that $\mathcal{E}_y > \mathcal{E}_x$, and that $\partial\mathcal{E}_y/\partial y$ is therefore at least comparable to, if not greater than, $\partial\mathcal{E}_x/\partial x$].

In summary, the GCA appears to be valid for MOSFETs operated in strong inversion even if the channel length is shorter than 1 μm - provided that V_{DS} is small enough that the MOSFET is not being operated in saturation (i.e., $V_{DS} \leq V_{DSsat}$, where V_{DSsat} is the voltage at the onset of saturation). On the other hand, if the MOSFET is operated in saturation, use of the GCA to calculate I_D will not yield accurate results, and alternative approximations (as described in chap. 5) must be employed to get good agreement with experimental I_D-V_{DS} characteristics.

4.3.2 Circuit Characteristics of Long-Channel MOSFETs

From the perspective of the MOSFET as a circuit component, long-channel behavior has been specified in a number of ways. A list of device parameters from a circuit perspective deemed to be characteristic of long-channel MOSFETs include the following:

1. The threshold voltage V_T is independent of channel length L.
2. The threshold voltage V_T is independent of drain bias voltage, V_{DS}.
3. The drain current in saturation I_{Dsat} is independent of V_{DS}.
4. The drain current has a linear dependence on $1/L$ (i.e., one quantitative criterion is that there should be no more than a 10% departure from linearity as the channel length is decreased).

5. The subthreshold current I_{Dst} is independent of drain bias.
6. The subthreshold swing S_t is independent of gate length.

As the behavior of the MOSFET is described later in this chapter, the terms associated with these device characteristics will be defined and discussed in more detail.

4.4 QUALITATIVE DESCRIPTION OF MOSFET OPERATION

The operation of the MOSFET is first described from a qualitative perspective. Specifically, our goal is to qualitatively explain the so-called *static characteristics* of the MOSFET; i.e., the drain current characteristics of the MOSFET as a function of the applied dc terminal voltages V_{GS} and V_{DS}. This behavior is normally depicted using an I_D-V_{DS} plot, with V_{GS} as the parameter.

For simplicity we consider an ideal (i.e., $V_{FB} = 0$) long-channel NMOSFET (*p*-type substrate) as our example device structure, with the source being chosen as the reference terminal (Fig. 4-4b). We also assume that the source-substrate and drain-substrate junctions are always kept under reverse bias and that the resulting reverse-bias leakage currents are negligible (this is valid as long as the bias voltage is kept smaller than the junction breakdown voltage), and that $V_{SB} = 0V$. We also use energy-band versus position and potential versus position diagrams to help depict the qualitative MOSFET behavior. Note that in the case of the MOS-C one-dimensional diagrams were adequate. Here, however, two-dimensional energy-band and potential diagrams must be used. Axes and terminal voltage designations are the same as in Fig. 4-2a.

4.4.1 MOSFET Behavior in Response to Gate Bias Alone

We begin by describing how a MOSFET behaves in response to a gate bias V_{GS} alone. In this case, the region under the oxide responds in the same way as does an ideal MOS-C. However, because the source and drain regions are present in a MOSFET the overall device behavior differs from that of a MOS-C with a V_{GB} applied.

To analyze the MOSFET behavior, let us assume that all three bias voltages are initially set to zero (i.e., $V_{GS} = V_{DS} = V_{BS} = 0V$). In this case a flat band condition in exists in the gated region of the MOSFET (as is also observed when $V_{GB} = 0V$ in an ideal MOS-C). The two-dimensional energy-band diagram corresponding to this condition is shown in Fig. 4-7b. We observe that E_c in the *x*-direction at the cut A-A is indeed flat. (The potential versus position diagram in the MOSFET for this bias condition is also shown in Fig. 4-8b). Assuming that $\varphi(x = \infty) = 0$, in this case $\varphi_{surf} = 0$ as well.)

At the surface ($x = 0$), however, a potential energy barrier of value qV_{bi} exists between the *p* region and each of the n^+ regions (where V_{bi} is the built-in potential of the *doped* pn^+ junction). This is shown in the E_c versus *y* direction plot for $x = 0$ in Fig. 4-7b. For this case (and for cases of when a reverse bias exists across the doped junctions) only the very small leakage current of a reverse-biased *pn* junction flows.

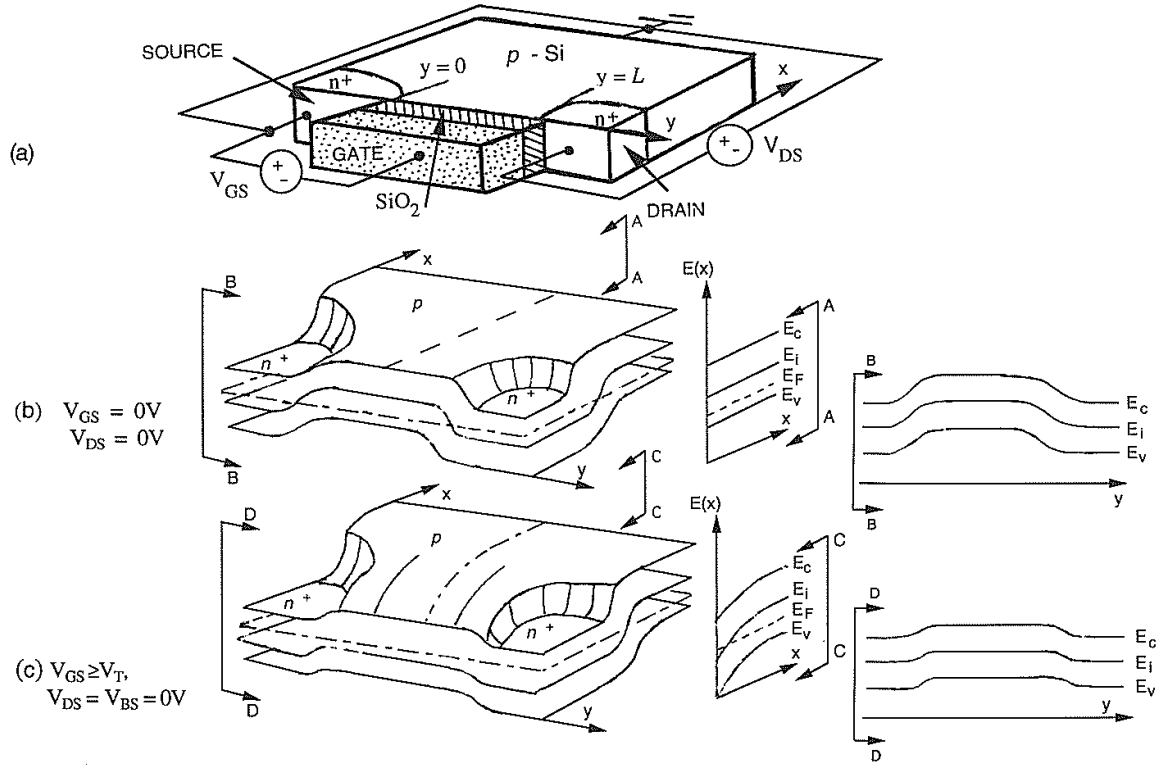


Fig. 4-7 (a) Cross section of an NMOS device. (b & c) 2-D perspective plots of the energy bands versus position of an ideal long-channel NMOS device when (b) $V_{GS} = 0V$ (flat-band condition), and $V_{DS} = 0V$; and (c) $V_{GS} > V_T$, $V_{DS} = V_{BS} = 0V$.

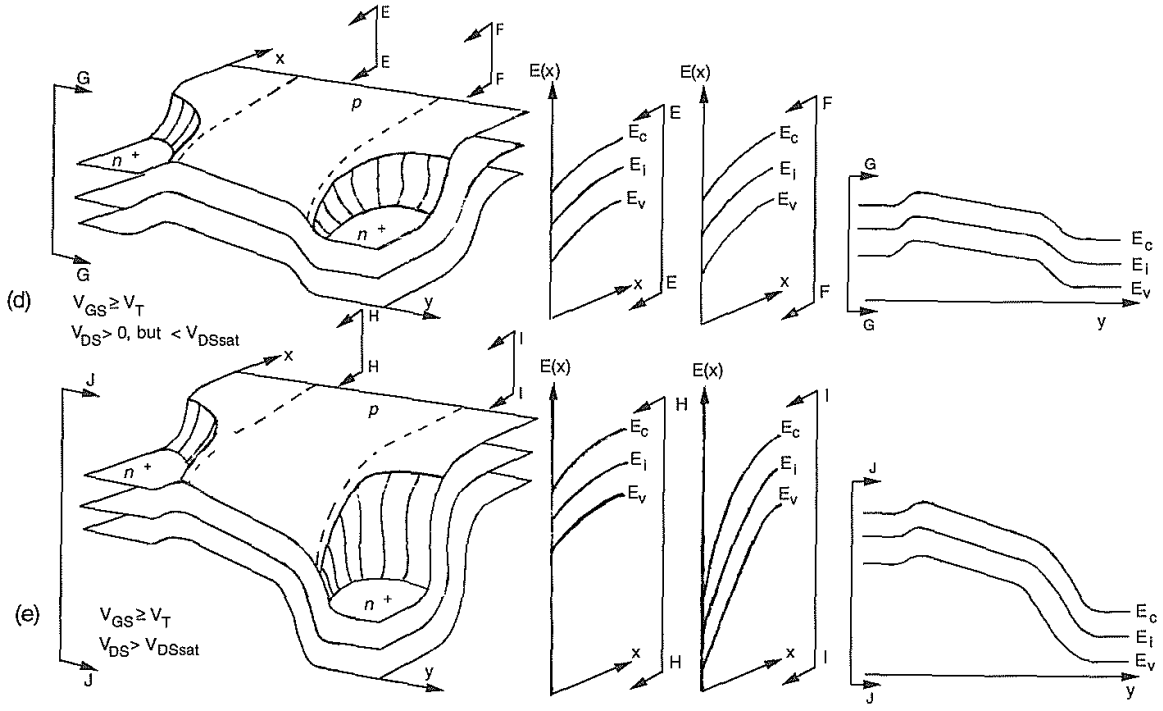


Fig. 4-8 (cont.) (d & e) 2-D perspective plots of the energy bands as a function of position an ideal long-channel NMOS device when: (d) $V_{GS} \geq V_T$ and V_{DS} is small (i.e., $< V_{DSsat}$); and (e) $V_{GS} \geq V_T$ and $V_{DS} > V_{DSsat}$. $V_{BS} = 0$ in both cases.

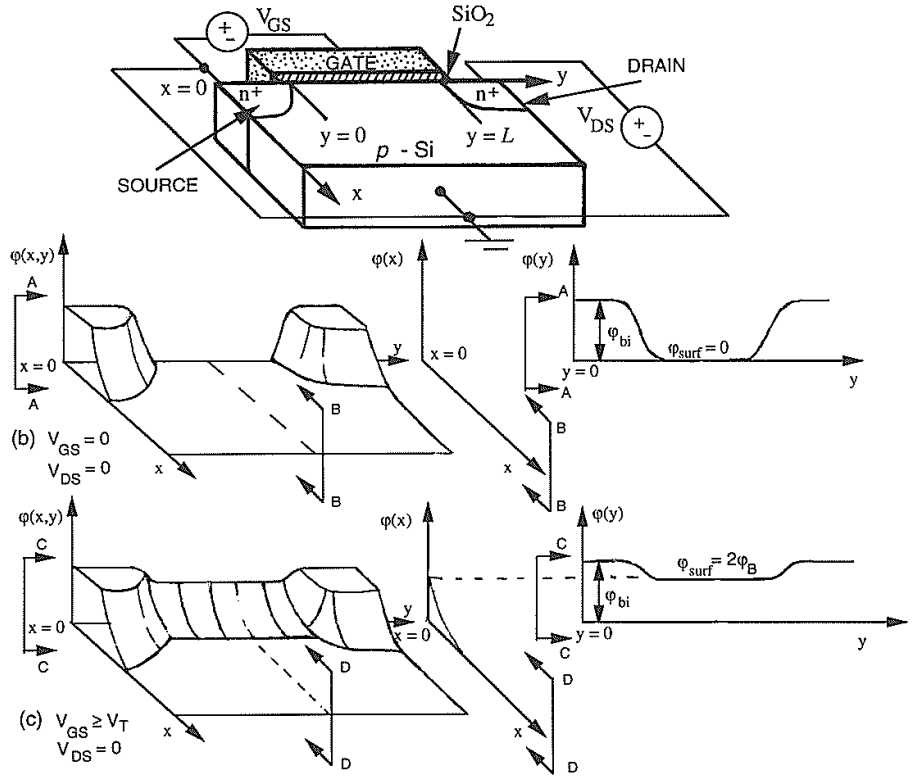


Fig. 4-8 (a) Cross section of an NMOS device. (b & c) 2-D perspective potential versus distance diagrams of an ideal NMOS device when: (b) $V_{GS} = 0V$ (flat-band condition), and $V_{DS} = 0V$; and (c) $V_{GS} \geq V_T$, $V_{DS} = V_{BS} = 0V$.

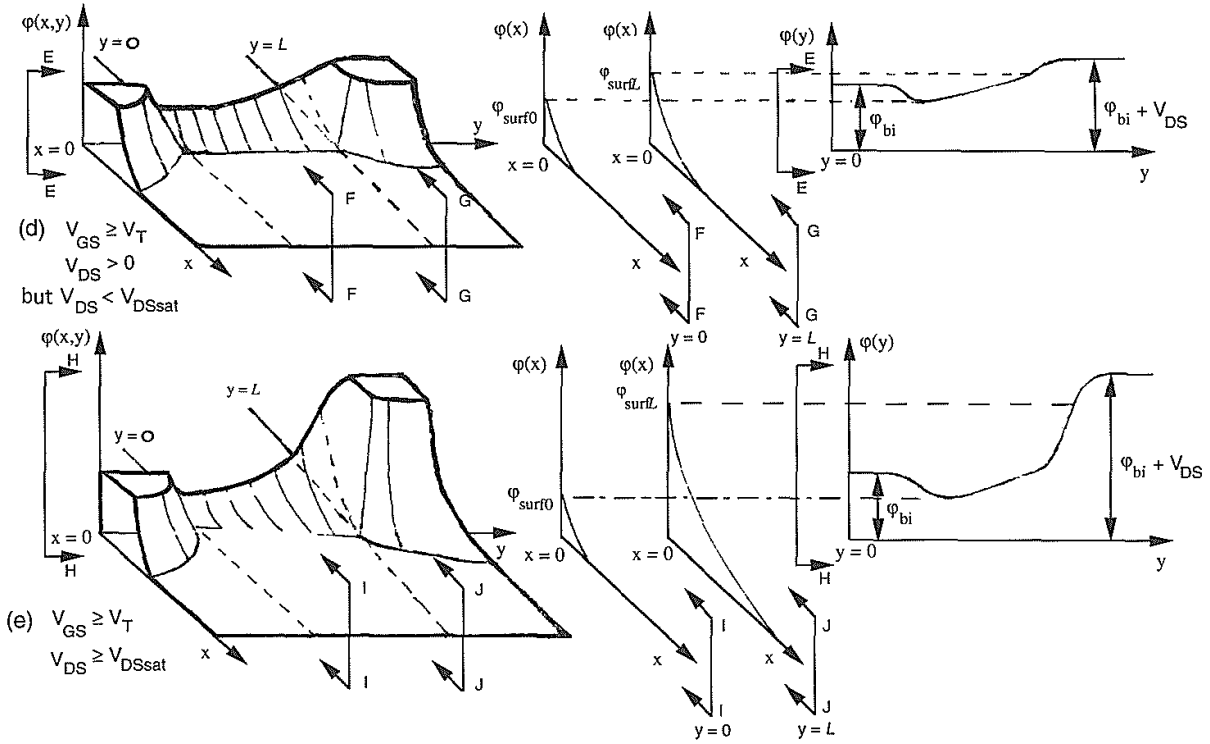


Fig. 4-8 (cont.) (d & e) 2-D perspective potential versus distance diagrams of an ideal long-channel NMOS device when (d) $V_{GS} \geq V_T$ and $V_{DS} > 0V$, but $< V_T$; and (e) $V_{GS} > V_T$, and $V_{DS} > V_{DSsat} \sim V_{GS} - V_T$.

Thus, the circuit path between the n^+ source and drain regions effectively appears as an open circuit. This means that even if a voltage is applied between the source and drain ($V_{DS} \neq 0V$), essentially no drain current will flow. Moreover, if we apply either a negative V_{GS} (or a positive V_{GS} such that $V_{GS} < V_T$,) the gated region contains either an excess of holes (*accumulation*, when $V_{GS} < 0V$) or a deficit of holes and very few electrons (*depletion*, when $V_{GS} < V_T$). In this case, the circuit path between source and drain still appears essentially as an open circuit. That is, no current flows in the drain terminal even if a drain voltage is applied. For all of these conditions, the MOSFET is said to be in an "OFF" state.

If a positive V_{GS} is applied such that $V_{GS} > V_T$, an inversion layer containing mobile electrons is formed adjacent to the Si surface throughout the entire gated region between the source and drain (here we still assume that $V_{DS} = V_{BS} = 0V$). Thus, the circuit path between source and drain now contains an induced "*n*-type" region (the inversion layer), as shown in Fig. 4-9. This means that a conducting path or *channel* exists between the source and drain. (However, if $V_{DS} = 0V$, no drain current will yet flow.) Now, if a drain voltage is applied (i.e., $V_{DS} > 0V$), drain current will flow. The device in this state is said to be "ON." Obviously, the larger the value of gate voltage (V_{GS}), the more electrons pile up at the Si surface, and the greater the conductance of this channel. This demonstrates that the input voltage to the MOSFET can induce a source-to-drain channel and will determine the maximum conductance of the channel.*

Figure 4-7c shows the energy band diagram of the MOSFET when $V_{GS} > V_T$ and $V_{DS} = V_{BS} = 0V$, and Fig 4-8c also shows the potential diagram for this case. Note that the bands along the *x*-direction in the gated region are no longer flat (see section B-B) and that $\varphi_{surf} > 2\varphi_B$ everywhere along the channel. The potential-energy barrier between the n^+ and *p* regions is also reduced (see the E_c versus *y* plot in Fig. 4-7c). Hence, only a very low barrier to current injection exists between the source and the channel. The barrier height is smallest at the surface with its value controlled by the bias on the gate. At strong inversion and beyond, the barrier height is assumed to be remain constant at $\sim q(V_{bi} - 2\varphi_B)$, because φ_{surf} is assumed to be clamped at $2\varphi_B$ in strong inversion. However, as V_{GS} is increased, the inversion charge density continues to increase, increasing the conductance of the channel

4.4.2 Qualitative MOSFET Behavior when both Gate and Drain Bias are Applied

We now consider the variation of I_D with V_{DS}. At first we assume that V_{GS} is kept fixed at a single value while V_{DS} is varied, and $V_{BS} = 0V$. Later we show how varying the value of V_{GS} impacts the I_D-V_{DS} characteristic. In this section we also assume that

* Note that when V_{GS} is positive but smaller than V_T, the gated region can be in state of weak inversion. In weak inversion only small, so-called *subthreshold currents* are able to flow between the drain and source. These subthreshold currents are not large enough to be useful as driving currents for most applications, but they are sufficiently large to cause troublesome leakage - which can complicate the design of some circuits. We will describe subthreshold currents in more detail in section 4.7 and in chap. 5, as part of discussion on the quantitative behavior of the long-channel MOSFET and the submicron MOSFET.

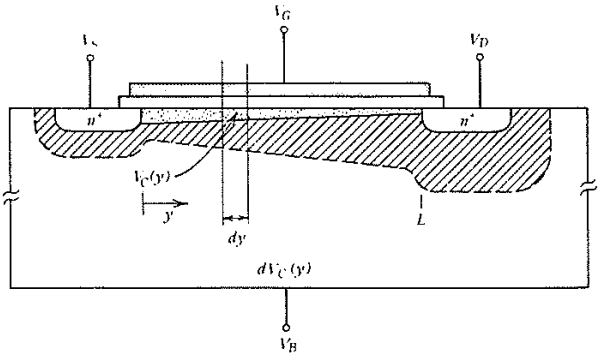


Fig. 4-9 Depletion region width variation from source to drain end of a MOSFET when V_{DS}>0V.

sufficient gate bias is applied to ensure that the MOSFET is being operated in inversion (i.e., $V_{GS} > V_T$). We then describe what happens as a dc V_{DS} is applied, with its value being increased in small steps, starting from $V_{DS} = 0V$.

As noted earlier, when $V_{DS} = 0V$, $I_D = 0$, even if $V_{GS} > V_T$ (see Fig. 4-10b). However, when V_{DS} is increased slightly, I_D starts to flow. When V_{DS} is small (i.e., a few tenths of a volt or less), the surface channel behaves like a simple resistor (Fig. 4-10c). That is, an I_D proportional to V_{DS} flows into the drain. If a plot of I_D versus V_{DS} is used to demonstrate the I_D-V_{DS} behavior, the response to small values of V_{DS} is shown as the line from the point 0 to the point A in Fig. 4-10a.

When V_{DS} is increased beyond a few tenths of a volt, the MOSFET exhibits a new phase of behavior. That is, current flowing in the channel gives rise to a significant voltage drop in the channel (unlike the case in the three-terminal structure introduced in section 4.2.1, where the channel voltage is constant along the entire channel length because essentially no current flows along it). Specifically, the *channel voltage* (introduced in section 4.2.1) is V_{DB}, at the drain end and V_{SB} at the source end. If $V_{SB} = 0V$, the channel voltage at the source end will equal zero. But even with $V_{SB} = 0V$, the channel voltage at the drain end will not be zero, since we assume $V_{DS} > 0V$. As shown in Fig. 4-9 such a non-zero V_{DB} causes the depletion region under the gate to widen toward the drain end of the channel. As described in section 4.2.1, this implies that the mobile charge concentration in the inversion layer simultaneously decreases. The smaller number of carriers causes the channel conductance to decrease, which in turn is manifested as a smaller slope in the I_D-V_{DS} characteristic as V_{DS} is increased. As the drain voltage is further increased, the depletion region continues to widen and the slope of the I-V characteristic continues to decrease (see Figs. 4-10a and 4-10c).

The energy-band and potential diagrams of the MOSFET when both a small V_{DS} and a $V_{GS} > V_T$ are applied are shown in Figs. 4-7d and 4-8d, respectively. Even if the drain voltage is small enough so that the entire surface region remains in inversion, we see that near the drain the band bending is larger than at the source (see the E_C vs. y diagram in Fig. 4-7d). This implies that the depletion region under the gate is wider there. Since the gate-to-body voltage is constant across the entire gated region, less voltage is

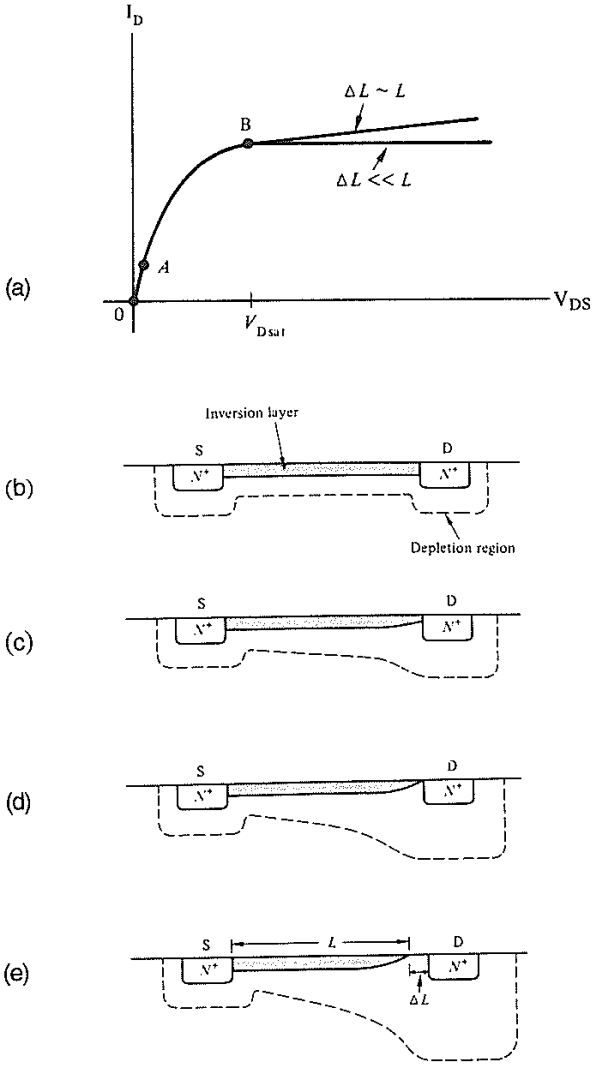


Fig. 4-10 (a) I_D-V_{DS} characteristic of a MOSFET being operated at a fixed V_{GS}, with $V_{GS}>V_T$. (b -e) Visualization of the various phases of $V_{GS}>V_T$ MOSFET operation: (b) V_{DS} = 0V; (c) channel (inversion layer) narrowing under moderate V_{DS} biasing; (d) pinchoff; and (e) post pinchoff ($V_{DS}>V_{DSsat}$).[30] Reprinted with permission of Addison-Wesley Pub. Co.

available near the drain to establish the inversion layer charge. Hence, inversion is less strong in the channel near the drain.

Eventually, if V_{DS} is sufficiently increased such that $V_{DS}>(V_{GS} - V_T)$, the inversion layer ceases to exist near the vicinity of the drain. The disappearance of the inversion layer due to an increase in V_{DS} is termed *pinch-off* (see Fig. 4-10d). The pinch-off point on the I_D-V_{DS} characteristic is shown as point B in Fig. 4-10a. We see that the slope of the I_D curve here has become approximately zero. Note, however, that pinch-off does *not* mean that I_D vanishes as V_{DS} exceeds the pinch-off voltage. In fact, it is the *gradual channel approximation* that is no longer valid in the region of the device in which pinch-off exists. That is, mobile charges still move through the depleted region near the drain, and drain current continues to flow. However, the equations for the charge distribution derived from the one-dimensional solution of the Poisson equation (which are valid for the case of the MOS-C) predict that no mobile charge exists in the pinch-off region. But in the depleted pinched-off region the one-dimensional Poisson equation is no longer valid. Instead, a two-dimensional numerical analysis (or an analytical expression based on a modified 1-D form of the Poisson equation - see section 5.4.3.3) must be used to calculate the potential and carrier concentrations in the pinched-off device region. A detailed numerical analysis of the saturation range has been carried out using the two-dimensional device simulator MINIMOS.[3] This work showed that for voltages greater than V_{DSsat} the current stream of the carriers moving into the pinched off region is no longer confined to the inversion layer region near the surface, but begins to move away from the surface and into the bulk. Note that this physical picture is plausible because once V_{DS} becomes larger than $(V_{GS} - V_T)$ the polarity of the vertical electric field in the pinched-off region is reversed. This tends to push any mobile carriers flowing in the channel *away* from the surface. Thus, in the region of the channel where $V_{DS}>(V_{GS} - V_T)$ the mobile carriers flow toward the drain *beneath* the surface.

Nevertheless, as the drain voltage increases beyond the pinch-off voltage V_{DSsat} the pinched-off section of the channel widens from just a point into a segment ΔL in length (see Fig. 4-10e). Most of the voltage drop in excess of V_{DSsat} is dropped across this pinched-off section ΔL. I_D remains approximately constant (*saturated*) for drain voltages greater then V_{DSsat} so long as $\Delta L<<L$. If ΔL becomes comparable to L (which is a condition that may occur in short-channel MOSFETs), the same voltage drop appears across a shorter channel $(L - \Delta L)$, and the post-pinch-off current I_{Dsat} in such devices will increase somewhat as V_{DS} is increased further beyond V_{DSsat}. Figures 4-7e and 4-8e show the energy-band diagrams and potential diagrams, respectively, for a long-channel MOSFET biased into pinch-off.

The response of the long-channel MOSFET to separate manipulation of gate and drain dc biases has been qualitatively described up to this point. This has allowed us to visualize the general variation of I_D with V_{DS} for a fixed value of $V_{GS}>V_T$. To allow the complete set of I_D versus V_{DS} characteristics to be obtained, we need to combine the results from the individual cases. These are as follows:

For $V_{GS}<V_T$, no channel exists and $I_D \approx 0$ for all drain biases below junction breakdown. For each value of $V_{GS}>V_T$, a characteristic of the form shown in

Fig. 4-10a will be observed. Since the conductance of the channel will increase as V_{GS} is made larger, the initial slope of I_D will also become steeper with increasing V_{GS}. Furthermore, since the inversion layer contains more electrons, a larger value of V_{DS} is needed to produce pinch-off (V_{DSsat} increases for increasing V_{GS}.). Based on these phenomena, the static I_D-V_{DS} characteristics assume the form shown in Fig. 4-11.

Let us next consider quantitative models which allow these static (i.e., dc) I_D-V_{DS} characteristics to be accurately simulated.

4.5 QUANTITATIVE BEHAVIOR OF MOSFETS: DC CIRCUIT MODELS OF LONG-CHANNEL MOSFETS

Although the discussion in the previous section provides a qualitative picture of how a MOSFET operates, we need to express this behavior in quantitative form. Specifically, to design circuits with MOSFETs we must be able to calculate how much current is carried in the "ON" state of a MOSFET and how much leakage current flows in the "OFF" state. In short, we need the current-voltage characteristics of the MOSFET, specifically the quantitative relationship between the drain current and terminal voltages.

In this section our goal is to derive relationships that allow the dc drain current I_D of long-channel MOSFETs to be quantitatively predicted. Such I_D-V_{DS} relationships are referred to as the *static (dc) circuit models* of the MOSFET. As is the case for the MOS-C, there exists a hierarchy of such long-channel MOSFET I_D-V_{DS} models in terms of their generality, accuracy, and complexity. Our discussion in this section will derive four of these models, using as our foundation two of the five classical semiconductor equations introduced in chap. 2: the one-dimensional Poisson equation, and the one-dimensional drift-diffusion current density equation (as well as the three models of the MOS-C derived in chap. 3).

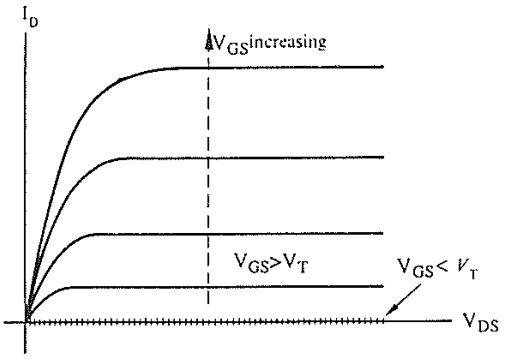


Fig. 4-11 General form of the I_D-V_{DS} characteristics expected from a long-channel MOSFET.

The first of these models is the *Pao-Sah* model, which is the most general and accurate of our four dc I_D long-channel MOSFET models. Like the general MOS-C model, it also the most computationally complex (in that it requires numerical analysis techniques to calculate I_D for a given set of terminal voltages). The three less-computationally complex long-channel models are derived from the Pao-Sah model by making suitable simplifying approximations. The first of these that we will study is the *charge-sheet MOSFET* model. It is based on the same assumptions used in the MOS-C charge-sheet model introduced in chap. 3. Unlike the Pao-Sah model, it provides an analytical expression for I_D in terms of V_{GB} and φ_{surf}, and thus involves much less computational effort. Nevertheless, it is only slightly less accurate than the Pao-Sah model provided the MOSFET is operated in inversion (i.e., this model is valid from weak through strong inversion). However, while this model provides an analytical expression for I_D, it still requires a numerical procedure to obtain the values of the surface potential needed to evaluate the I_D expression. That is, I_D is not explicitly stated in terms of the terminal voltages. (The terminal voltage values are used to first calculate φ_{surf}, which is then used to compute I_D.) Thus, the charge-sheet model is still too cumbersome to serve as a MOSFET model for most circuit-simulation applications.

By making suitable approximations that are valid over a more limited range of operation (namely, the strong inversion regime), two simpler long-channel MOSFET models have been derived that are extensively used in circuit analysis and design. They are the *variable depletion-charge* and the *square-law* models. These provide analytical expressions which give I_D explicitly in terms of the terminal voltages. We show how they can be derived from the charge-sheet model by making additional simplifying assumptions. At the end of the section we compare these four models in more detail.

4.5.1 General Approach to Deriving the DC Drain-Current Relationships in Long-Channel MOSFETs

The derivation of the dc drain current relationship recognizes that, in general, the current in the channel of a MOSFET can be caused by both drift and diffusion current. In an NMOSFET we can accurately assume that only electrons contribute to the drain current. Thus, we only need to use Eq. 1-5a to describe the drain current density J_D

$$J_D = J_n = q\ \mu_n\, n\, \mathcal{E} + q\, D_n\, \nabla n \qquad (4\text{ - }16)$$

Within the conducting channel of a MOSFET the current flows almost exclusively in the y-direction, thus Eq. 4-14 can be simplified to the following one-dimensional form:

$$J_D \approx J_{ny} = q\, \mu_n\, n\, \mathcal{E}_y + q\, D_n\, dn/dy = -q\, \mu_n\, n\, d\varphi/dy + q\, D_n\, dn/dy \qquad (4\text{ - }16a)$$

Since we also assume there are no current sources or sinks inside the device, the current flowing through any cross-sectional plane within the channel must be constant, having some value I_D. We also assume that the current density is constant in the z-direction, but not necessarily in the x-direction. Hence, if J_{ny} is integrated over Z we get

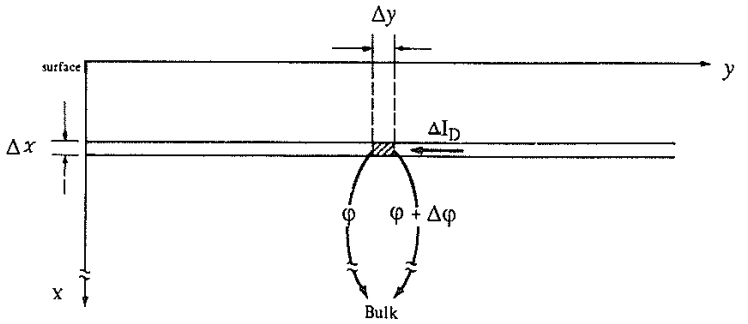


Fig. 4-12 The drain current I_D in a MOSFET can be modeled as the sum of currents in slices ΔI_D each having a thickness Δx and a width Z. These slices exist from the surface to the depth at which the electron concentration becomes negligible.[5] Reprinted with permission of McGraw-Hill Book Co.

$$I_D = -\int\int J_{ny}\ dx\ dz = -Z\int_0^{x_c(y)} J_{ny}\ dx \qquad (4 - 17)$$

If the flow of electrons can be considered to be laminar, we can assume that the total current I_D is the sum of elemental currents ΔI, each flowing horizontally in an inversion layer slice parallel to the surface, having width Z and depth Δx, as shown in Fig. 4-12. If we allow Δx to become a differential, the current in each slice dI_D is

$$dI_D = dI_{Ddrift}(x,y) + dI_{Ddiff}(x,y) \qquad (4 - 18a)$$

Since $\mathcal{E}(y) = -d\varphi/dy$, the drift component of Eq. 4-18a can be written as

$$dI_{Ddrift}(x, y) = q\,\mu_n\,Z\,n(x, y)\,\partial\varphi(x, y)/\partial y \qquad (4 - 18b)$$

The diffusion component of Eq. 4-18b can also be rewritten as,

$$dI_{Ddiff}(x, y) = -Z\,q\,D_n\,\partial n(x, y)/\partial y = -Z\,\mu_n\,kT\,\partial n(x, y)/\partial y \qquad (4 - 18c)$$

since $D_n/\mu_n = kT/q$.

An expression for the total drain current for each of these components can be obtained by integrating each expression over the depth from $x = 0$ to a point below the surface at which the electron concentration is negligible, x_c or

$$I_{Ddrift} = q\,\mu\,Z\frac{d\varphi}{dy}\int_0^{x_c(y)} n(x,y)\,dx \qquad (4 - 19)$$

$$I_{Ddiff} = -\mu Z kT \frac{d}{dy}\left[\int_0^{x_c(y)} n(x,y)\,dx\right] \qquad (4-20)$$

Since the charge per unit area due to carriers in the inversion layer, Q_I is given by

$$Q_I(y) = -q\int_0^{x_c(y)} n(x,y)\,dx \qquad (4-21)$$

Thus, Eqs. 4-19 and 4-20 can be written as

$$I_{Ddrift} = \mu Z(-Q_I)\frac{d\varphi}{dy} \qquad (4-22)$$

and

$$I_{Ddiff} = \mu Z \frac{kT}{q}\frac{dQ_I}{dy} \qquad (4-23)$$

Our goal is to derive a model in which I_D is expressed in terms of the terminal voltages V_{GS} and V_{DS}. To do this we must integrate the sum of Eqs. 4-22 and 4-23. The integration is carried out by changing the variables of integration by multiplying both sides of Eqs. 4-22 and 4-23 by (dy). Since we recognize that I_D is constant along the length of the channel, when we integrate the left hand side of the sum of Eqs. 4-22 and 4-23 (namely I_D[dy] from 0 to L) we get

$$I_D L = \int_0^L I_D\,dy \qquad (4-24)$$

Then we can write the total drain current I_D as

$$I_D = \frac{\mu Z}{L}\int_{\varphi_{surf0}}^{\varphi_{surfL}} (-Q_I)\,d\varphi + \frac{\mu ZkT}{Lq}\int_{Q_{I,source}}^{Q_{I,drain}} dQ_I \qquad (4-25)$$

We see that the first term on the right hand side of Eq. 4-25 is no longer integrated with respect to position, but with respect to voltage. (In the *variable-depletion charge model* [section 4.5.4, Eq. 4-58], and the *square-law model* [section 4.5.5, Eq. 4-62], we assume that only the drift current term is needed to describe I_D. Hence, only this integral needs to be carried out to calculate I_D. By changing the variable of integration from position to voltage this integration becomes easy to perform, yielding Eqs. 4-58 and 4-62.) The limits of integration of the first term of Eq. 4-25 are the value of the surface potential at the source end φ_{surf0} (which is usually set equal to zero), and the surface potential value at the drain end φ_{surfL} (which is usually set to V_{DS}). The limits of the integration

of the second term are the value of the mobile charge density at the source end of the channel $Q_{Isource}$, and the value at the drain end Q_{Idrain}.

4.5.2 Pao-Sah Model

The *Pao-Sah* model,[4] published in 1966, was the first advanced long-channel MOSFET model to be developed. While it retained the gradual-channel approximation, it did not invoke the depletion approximation and permitted carrier transport in the channel by both drift and diffusion current. The formulation of the I_D equation it develops is therefore quite general, but as a result requires numerical integrations in two dimensions. Such numerical complexity has limited the role of this model to theoretical analyses of the MOSFET. Other simpler models of the MOSFET have been developed for use in IC circuit design applications.

The Pao-Sah model follows the approach outlined in the previous section for deriving the I_D relationship from the semiconductor device equations. However, the derivation is usually carried out using a more compact formulation, in which the drift and diffusion current expression are combined into a single equation. When we derive the charge-sheet MOSFET model we will return to calculating both current components separately, as this results in a simpler derivation. Here, the opposite is true, and so we also elect the single equation approach to deriving the Pao-Sah model. Note that the presentation of the derivation in this section, and the next, closely mimics the method proposed by Tsividis.[5]

We begin by noting that the Pao-Sah approach invokes the gradual channel approximation (like all long-channel MOSFET models), and assumes that μ is constant along the channel. In our derivation we also assume that the substrate doping is uniform and that an n-channel MOSFET is being treated.

To simplify the expression for I_D so that it consists of one instead of two terms on the right hand side of Eq. 4-16, we start by noting that at the source end of the channel in the MOSFET (where $y = 0$), n can be expressed in terms of φ by

$$n\,(x, 0) = n_{po} \exp q[\varphi\,(x, 0) - V_{SB}]/kT \qquad (4 - 26)$$

and at the drain end by

$$n\,(x, L) = n_{po} \exp q[\varphi\,(x, L) - V_{DB}]/kT \qquad (4 - 27)$$

For an arbitrary position y along the channel, n is given by

$$n\,(x, y) = n_{po} \exp q[\varphi\,(x, y) - V_{ChB}(y)]/kT \qquad (4 - 28)$$

where $V_{ChB}\,(0) = V_{SB}$ and $V_{ChB}\,(L) = V_{DB}$. In Eq. 4-28, V_{ChB} represents the potential corresponding to the difference between the quasi-Fermi level of the electrons in the inversion layer and the quasi-Fermi level of holes in the bulk, as shown in Fig. 4-13. That is, because the source and drain will in general be biased with respect to the substrate, the channel is also at some potential $V_{ChB}\,(y)$ relative to the substrate. Clearly this voltage is a function of position y along the channel if $V_{DS} \neq 0$, because at the source end we have $V_{ChB}\,(0) = V_{SB}$ and at the drain end $V_{ChB}\,(L) = V_{DB}$.

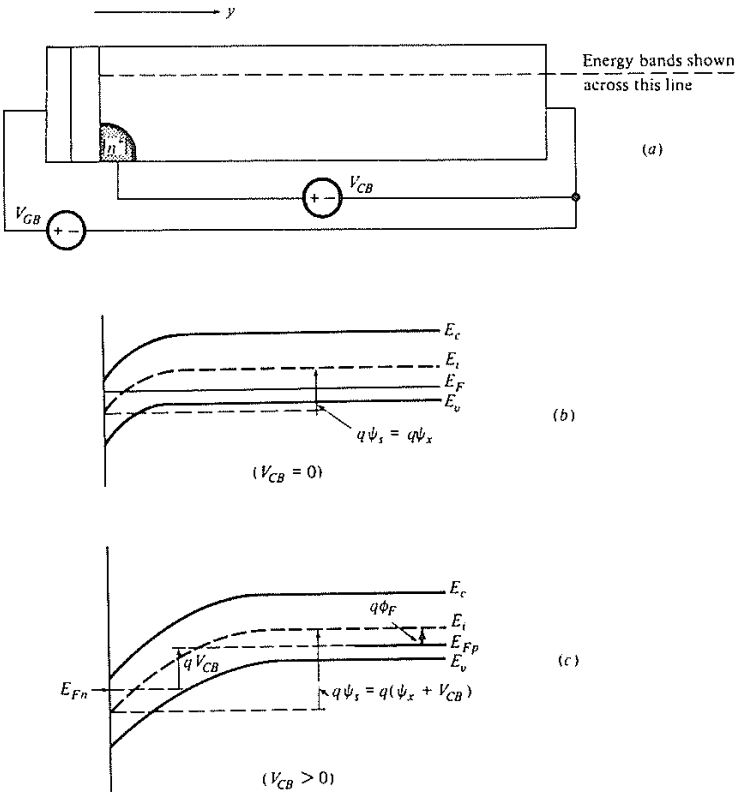


Fig. 4-13 Energy-band diagram of a three-terminal MOS structure with a bias V_{CB} (= V_{ChB} in a MOSFET) showing how the applied bias V_{CB} is related to the difference between the quasi-Fermi levels of the electrons in the inversion layer E_{Fn} and that of the holes E_{Fp}.[5] Reprinted with permission of McGraw-Hill Book Co.

If Eq. 4-28 is differentiated with respect to y we get

$$\frac{\partial n(x, y)}{\partial y} = \frac{q\, n(x, y)}{kT}\left[\frac{\partial \varphi(x, y)}{\partial y} - \frac{\partial V_{ChB}(y)}{\partial y}\right] \qquad (4\text{-}29)$$

Using Eq. 4-29 in Eq. 4-18 we obtain

$$d\, I_{Ddiff}(x, y) = -(Z\, dx)\, q\, \mu\, n(x, y)\left[\frac{\partial \varphi(x, y)}{\partial y} - \frac{\partial V_{ChB}(y)}{\partial y}\right] \qquad (4\text{-}30)$$

If Eq. 4-30 is inserted together with Eq. 4-17 into Eq. 4-16 it is seen that the $\partial\varphi/\partial x$ terms cancel out, leaving

$$d\,I_D\,(x, y) = (Z\,dx)\,q\;\mu\,n\,(x, y)\,\frac{dV_{ChB}(y)}{dy} \qquad (4 - 31)$$

As was done above for Eqs. 4-19 and 4-20, I_D from Eq. 4-31 can be obtained by integrating from the surface to a depth $x = x_c$ beneath which the electron concentration is negligibly small

$$I_D = \mu\,Z\,\frac{dV\,(y)}{dy}\,q\int_0^{x_c(y)} n\,(x, y)\,dx \qquad (4 - 32)$$

From Eq. 4-21 we can write Eq. 4-32 as

$$I_D = \mu\,Z\,(-Q_I)\,\frac{dV\,(y)}{dy} \qquad (4 - 33)$$

Now, integrating over the channel length and using Eq. 4-24, we get

$$I_D = \frac{\mu\,Z}{L}\int_{V_{SB}}^{V_{DB}} (-Q_I)\,dV \qquad (4 - 34)$$

We can now rewrite the expression for Q_I given by Eq. 4-21 as

$$Q_I = q\int_0^{x_c} n\,(x, y)\,dx = q\int_{\varphi_{surf}}^{\varphi_c} \frac{n\,(\varphi)\,d\varphi}{d\varphi/dx} \qquad (4 - 35)$$

If we define $\varphi_t = kT/q$ we can write $n\,(\varphi)$ in Eq. 4-35 as

$$n\,(\varphi) = N_A \exp\,[\varphi\,(x,y) - 2\varphi_B - V_{ChB}(y)]/\varphi_t \qquad (4 - 36)$$

because $p_{po} \approx N_A$, and $n_{po} = n_i^2/N_A$, and $\varphi_B = kT/q \ln\,(N_A/n_i)$. Thus,

$$(N_A/n_i)^2 = \exp\,(2\varphi_B)/\varphi_t \qquad (4 - 37)$$

Solving Eq. 4-37 for n_i^2 and substituting the result for n_i^2 into $n_{po} = n_i^2/N_A$, we get

$$n_{po} = N_A \exp\,(-2\varphi_B/\varphi_t). \qquad (4 - 38)$$

In addition, we recall that $\mathcal{E}(x) = -\,d\varphi/dx$. In chap. 3 we derived an expression for $\mathcal{E}(x)$ in a MOS-C in which both electrons and holes were allowed in the depletion region of the semiconductor. This expression, with φ being replaced by $\varphi \rightarrow \varphi + V_{ChB}$, is the expression we use for $d\varphi/dx$ in the denominator of Eq. 4-35. That is, $\mathcal{E}(x)$ is now

$$-\frac{d\varphi}{dx} = \mathcal{E}(\varphi) = \sqrt{\frac{2q\,N_A}{\kappa_{Si}\varepsilon_0}}\sqrt{\varphi_t e^{-(\varphi + V_{ChB})/\varphi_t} + \varphi - \varphi_t + e^{-2\varphi_B/\varphi_t}\left(\varphi_t e^{-(\varphi + V_{ChB})/\varphi_t} - \varphi - \varphi_t\right)} \tag{4 - 39}$$

Then substituting Eqs. 4-36 and 4-39 into Eq. 4-35 we obtain the *Pao-Sah I_D expression* *

$$I_D = \frac{q\,\mu\,Z\,N_A}{L}\int_{V_{SB}}^{V_{DB}}\int_{\varphi_{surf}}^{0}\left[\{e^{(\varphi - 2\varphi_B - V_{ChB})/\varphi_t}\}/\,\mathcal{E}(\varphi)\right]\,d\varphi\,dV \tag{4 - 40}$$

In this equation, we must first calculate the limit of integration φ_{surf} by using

$$V_{GB} = V_{FB} + \varphi_{surf} \tag{4 - 41}$$

$$+\ \gamma\sqrt{\varphi_t e^{-(\varphi_{surf} + V_{SB})/\varphi_t} + \varphi_{surf} - \varphi_t + e^{-2\varphi_B/\varphi_t}\left(\varphi_t e^{-(\varphi_{surf} + V_{SB})/\varphi_t} - \varphi_{surf} - \varphi_t\right)}$$

where $\gamma = \sqrt{2\kappa_{Si}\varepsilon_0 N_A/C_{ox}}$. Note that this equation must be solved iteratively for φ_{surf}.

To solve for I_D once φ_{surf} is obtained, Eq. 4-40 must be solved numerically. This involves numerically evaluating the double integral in Eq. 4-40. This is a computationally lengthy procedure. Nevertheless, the Pao-Sah model has been found to accurately predict the I_D in long-channel MOSFETs.

Figure 4-14 is an example of the I_D-V_{DS} characteristics of a long-channel MOSFET

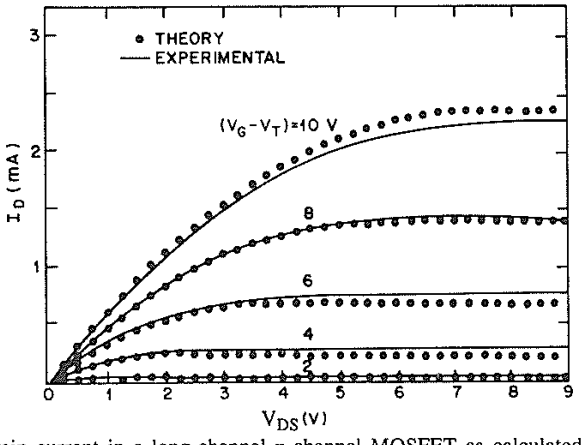


Fig. 4-14. Drain current in a long-channel *p*-channel MOSFET as calculated with the Pao-Sah model (dots) and as measured (solid lines).[4] Reprinted with permission of Solid-State Electronics.

* If $V_{BS} = 0V$, the limits of the voltage integral become 0 and V_{DS}, respectively.

calculated with the Pao-Sah model, together with experimentally measured results. We see that the agreement between measured and predicted I_D values is excellent.

The computational complexity of the Pao-Sah model, however, makes it impractical for use as a MOSFET model in circuit simulation applications. Thus, we show how simpler analytical models can be developed by limiting the validity of the model to reduced regimes of operation. We also mention that an approach has been proposed to reduce Eq. 4-40 to a single-integral formula (which, however, must still be evaluated numerically).[6] However, we do not discuss this approach here.

4.5.3 The Charge Sheet Model

The limited practical utility of the Pao-Sah model motivated a search for an approximate advanced analytical model that still was accurate over a wider range of operating conditions than the earlier *variable depletion-charge* (introduced in 1964) and *square law* models, which are only valid if the MOSFET is operated in strong inversion. The charge-sheet MOSFET, introduced separately by Bacarani[7] and Brews[8] in 1978, has become the most widely adopted long-channel MOSFET model that is accurate over the entire range of inversion.

The charge-sheet model of the long-channel MOSFET makes some simplifying assumptions so the numerical analysis needed in the Pao-Sah model to calculate I_D can be avoided. By doing so, it develops an analytical closed-form expression for I_D in terms of V_{GB} and the surface potential at the drain and source ends of the channel.* It is somewhat less general than the Pao-Sah model in that it is not valid if the MOSFET is operated in depletion or accumulation. However, it is only slightly less accurate than the Pao-Sah model if the MOSFET is being operated in inversion. Hence, by using this model it is possible to get quite accurate predictions of I_D in these regimes without having to resort to the complex numerical analysis required by the Pao-Sah model.

We introduced the charge-sheet approximation for the MOS-C in chap. 3. The MOSFET model is based on essentially the same approximations, with a few more added to account for the current flow behavior of the MOSFET. That is, the charge-sheet approximation basically assumes that the inversion layer (both in weak and strong inversion) can be treated as an infinitesimally thin sheet of charge at the Si surface, and that the depletion approximation is valid for the depletion region below the inversion layer. As such, the inversion-layer charge density Q_I can be found by subtracting the ionic charge in the depletion region Q_B from the total charge in the semiconductor Q_C.

The other assumptions of the charge-sheet model of the MOSFET are the following:

1. The gradual channel approximation applies to the MOSFETs being treated by this model.
2. The drain current consists of both drift and diffusion components.
3. The mobility μ in the channel is proportional to the electric field and is constant with position along the channel.

* Although, as we shall see, a numerical iteration procedure must be used to determine these surface potential values from the terminal voltages.

We begin the derivation of the charge-sheet I_D model by assuming that I_D consists of both drift and diffusion current components. Since we also assume that the mobile charge is located in a very thin region near the surface, all the current is located at one depth (i.e., there is no x-dependence of current). Hence, Eq. 4-16 can be expressed as

$$I_D(y) = I_{Ddrift}(y) + I_{Ddiff}(y) \tag{4 - 42}$$

Now, we consider each of these terms separately. To express I_{Ddrift}, we treat a small element of the inversion layer between y and Δy, as depicted in Fig. 4-15. The potential difference across this element is given by the difference of φ_{surf} at the points y and $y + \Delta y$, or $\varphi_{surf}(y + \Delta y) - \varphi_{surf}(y)$. Hence we can write an expression for the drift current using Eq. 4-22 with $\Delta\varphi(y)$ replaced by $\Delta\varphi_{surf}(y)$. In addition, we can express this equation in terms of the mobile charge density per unit area Q_I instead of in terms of n by multiplying the mobile charge density in the channel (qn) by the inversion layer thickness t_{inv}, or $Q_I = -q\, n\, t_{inv}$. Since $I_{Ddrift}(y) = A\, J_{drift} = Z\, t_{inv}\, J_{drift}$, $I_{Ddrift}(y)$ becomes

$$I_{Ddrift}(y) = \mu Z(-Q_I)\, \Delta\varphi_{surf}(y)/\Delta y \tag{4 - 43}$$

If we allow Δy to approach zero we get

$$I_{Ddrift}(y) = \mu Z(-Q_I)\, d\varphi_{surf}/dy \tag{4 - 44}$$

The diffusion component of Eq. 4-42 can likewise be rewritten as

$$I_{D\,diff}(y) = Z D\, dQ_I/dy = Z \mu (kT/q)\, dQ_I/dy \tag{4 - 45}$$

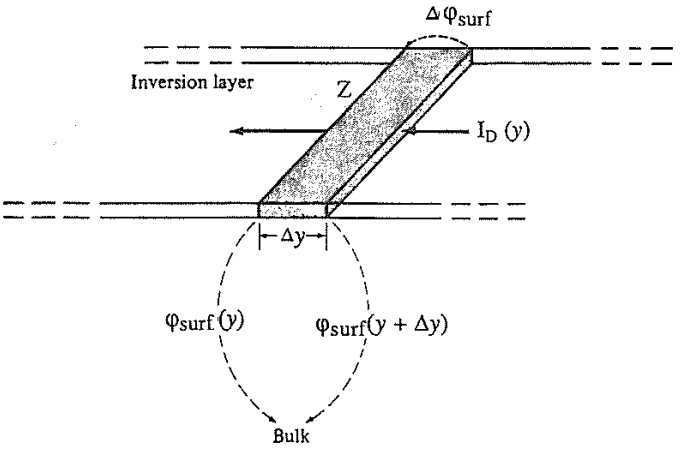


Fig. 4-15 Small element of the inversion layer in a long-channel MOSFET.[5] Reprinted with permission of McGraw-Hill Book Co.

In the case of dc biasing, the total current in the channel must be constant for all points in the y-direction. Thus, Eq. 4-42 can be written as

$$I_D(y) = \mu Z(-Q_I)\, d\varphi_{surf}/dy + Z\mu(kT/q)\, dQ_I/dy \qquad (4-46)$$

If we designate φ_{surf} at the source end of the channel ($y = 0$) by φ_{surf0}, and Q_I there by $Q_{Isource}$, and the corresponding quantities at the drain end of the channel ($y = L$) by φ_{surfL} and Q_{Idrain}, then by integrating from $y = 0$ to $y = L$ we get

$$\int_0^L I_D\, dy = \mu Z \int_{\varphi_{surf0}}^{\varphi_{surfL}} (-Q_I)\, d\varphi_{surf} + \frac{\mu Z kT}{q} \int_{Q_{I,source}}^{Q_{I,drain}} dQ_I \qquad (4-47)$$

or

$$I_D = \frac{\mu Z}{L} \int_{\varphi_{surf0}}^{\varphi_{surfL}} (-Q_I)\, d\varphi_{surf} + \frac{\mu Z kT}{L q} \int_{Q_{I,source}}^{Q_{I,drain}} dQ_I \qquad (4-48)$$

To perform the integrals in Eq. 4-47 and 4-48 we require an expression relating Q_I and φ_{surf}. Equation 4-6 derived in section 4.2.2.1 is the appropriate expression we seek, and is repeated here

$$Q_I = -C_{ox}(V_{GB} - V_{FB} - \varphi_{surf} - \gamma\sqrt{\varphi_{surf}}) \qquad (4-49)$$

Using this in Eq. 4-48 for the drift current component of the drain current, performing the integration from φ_{surf0} to φ_{surfL}, we get the following expression for I_{Ddrift}

$$I_{Ddrift} = (\mu C_{ox}Z/L)\,[(V_{GB} - V_{FB})(\varphi_{surfL} - \varphi_{surf0}) - (1/2)(\varphi_{surfL}^2 - \varphi_{surf0}^2)$$
$$- (2/3)\,\gamma\,(\varphi_{surfL}^{3/2} - \varphi_{surf0}^{3/2}) \qquad (4-50)$$

Using Eq. 4-49 for Q_I (in which we let $\varphi_{surf} = \varphi_{surf0}$ at the source and φ_{surfL} at the drain when we compute the values of Q_I at the limits of integration), in the term of Eq. 4-48 for I_{Ddiff}, we get

$$I_{Ddiff} = (\mu C_{ox}Z/L)(kT/q)(\varphi_{surfL} - \varphi_{surf0}) + (kT/q)\,\gamma\,(\varphi^{1/2}_{surfL} - \varphi^{1/2}_{surf0}) \qquad (4-51)$$

The last step needed when computing I_D with this model is to determine φ_{surf0} and φ_{surfL} from the terminal voltages V_{GB}, V_{SB}, and V_{DB}. It turns out that Eq. 4-4 is the appropriate expression for this task if φ_{surf} and V_{CB} are replaced with φ_{surf0} and V_{SB} at the source end of the channel, and with φ_{surfL} and V_{DB} at the drain end. Then we get

$$\varphi_{surf0} = V_{GB} - V_{FB} - \gamma\sqrt{\varphi_{surf0} + \varphi_t e^{(\varphi_{surf0} - 2\varphi_B - V_{SB})/\varphi_t}} \qquad (4-52a)$$

$$\varphi_{surfL} = V_{GB} - V_{FB} - \gamma\sqrt{\varphi_{surfL} + \varphi_t e^{(\varphi_{surfL} - 2\varphi_B - V_{DB})/\varphi_t}} \qquad (4-52b)$$

These equations for φ_{surfo} and φ_{surfL} however, are relatively complex transcendental equations which must be solved by iterative techniques to find φ_{surfo} and φ_{surfL}.

Now, I_D predicted by the charge-sheet model is a single expression given by the sum of Eqs. 4-50 and 4-51.

$$I_D\,(\text{charge sheet}) = I_{Ddrift}\,(\text{Eq. 4-50}) + I_{Ddiff}\,(\text{Eq. 4-51}) \qquad (4 - 53)$$

Experimental MOSFET characteristics are observed to agree very well with such predictions.

4.5.4 Variable Depletion-Charge (or Bulk-Charge) Model

As just described, the charge-sheet model for I_D in its full form includes drift and diffusion current components, making it accurate over the MOSFET operating regimes from weak to strong inversion. However, the fact that neither Eq. 4-50 nor 4-51 is explicitly expressed in terms of the terminal voltages makes the full charge-sheet formulation too cumbersome for use in circuit design applications.

Nevertheless, it is possible to simplify the charge-sheet model formulation to obtain analytical expressions that explicitly give I_D in terms of the terminal voltages. The price for such simplicity (while maintaining adequate accuracy) is that the resulting equations are only accurate over a limited operating range. Specifically, instead of being required to evaluate I_D from Eq. 4-53 (which is a single, relatively complex equation, valid from weak to strong inversion), we develop two less complex approximations to Eq. 4-53, one which is only valid over the weak inversion regime (subthreshold), and the other only over the strong inversion regime. In this section we derive the strong-inversion approximation to Eq. 4-53, which is termed the *variable-depletion-charge* or *bulk-charge* model, and was developed in 1964.[9] In section 4.7 we develop the simplified I_D equation valid for weak inversion. Note that in both of these models the charge-sheet and depletion approximations are still invoked.

We begin the derivation of the strong-inversion model by recalling that once a MOSFET is biased beyond the point of strong inversion, the diffusion current component of I_D is negligible compared to the drift current component; thus we can approximate I_D with only the drift current term of Eq. 4-53. In addition, we assume that beyond the onset of strong inversion φ_{surf} remains fixed. That is, the surface potential at the source and drain ends of the channel in strong inversion is assumed to be clamped at the value possessed at the onset of strong inversion. In that case, φ_{surf} can be approximated as

$$\varphi_{surfo} \approx 2\varphi_B + V_{SB} \qquad (4 - 54a)$$

and

$$\varphi_{surfL} \approx 2\varphi_B + V_{DB} \qquad (4 - 54b)$$

The final approximation we use to derive the variable-depletion-charge model is that if V_{GS} is biased below strong inversion (i.e., $V_{GS} < V_T$), I_D is assumed to be zero.

Invoking these approximations, I_D in a MOSFET can now be approximated by Eq. 4-50, with the values of φ_{surfo} and φ_{surfL} in this equation given by Eqs. 4-54a and 4-54b. Thus, by substituting these into Eq. 4-50 we get

$$I_D = \frac{\mu C_{ox} Z}{L}(V_{GB} - V_{FB})(V_{DB} - V_{SB}) - \frac{1}{2}[(V_{DB} + 2\varphi_B)^2 - (V_{SB} + 2\varphi_B)^2]$$

$$- \frac{2}{3}\gamma[(V_{DB} + 2\varphi_B)^{3/2} - (V_{SB} + 2\varphi_B)^{3/2}] \qquad (4 - 55)$$

By rearranging terms, this equation becomes

$$I_D = \frac{\mu C_{ox} Z}{L}(V_{GB} - V_{FB} - 2\varphi_B)(V_{DB} - V_{SB}) - \frac{1}{2}[(V_{DB}^2 - V_{SB}^2)$$

$$- \frac{2}{3}\gamma[(V_{DB} + 2\varphi_B)^{3/2} - (V_{SB} + 2\varphi_B)^{3/2}] \qquad (4 - 56)$$

We observe that I_D is now explicitly expressed in terms of the terminal voltages V_{GB}, V_{DB}, and V_{SB}. To put Eq. 4-56 into a form in terms of V_{GS}, V_{DS}, and V_{BS}, we recall that $V_{GB} = V_{GS} + V_{BS}$ and $V_{DB} = V_{DS} + V_{BS}$. This allows us to rewrite Eq. 4-56 as

$$I_D = \frac{\mu C_{ox} Z}{L}(V_{GS} - V_{FB} - 2\varphi_B)V_{DS} - \frac{1}{2}V_{DS}^2$$

$$- \frac{2}{3}\gamma\,[(V_{DS} + V_{BS} + 2\varphi_B)^{3/2} - (V_{BS} + 2\varphi_B)^{3/2}] \qquad (4 - 57)$$

If $V_{BS} = 0V$, Eq. 4-57 becomes

$$I_D = \frac{\mu C_{ox} Z}{L}(V_{GS} - V_{FB} - 2\varphi_B)V_{DS} - \frac{1}{2}V_{DS}^2 - \frac{2}{3}\gamma(2\varphi_B + V_{DS})^{3/2} - (2\varphi_B)^{3/2} \quad (4 - 58)$$

Equation 4-57 (and 4-58) is sometimes referred to as the *Ihantola and Moll model*, since it was initially developed by them in 1964.[9] The name *variable-depletion charge model* signifies that the model computes Q_I as the difference between Q_C and Q_B (*bulk* charge), taking into account the variation of the depletion region width along the channel to compute Q_B. We will see that the square-law model assumes the depletion region width is constant from source to drain.

Equation 4-57 predicts that for a given V_{GS} the drain current first increases linearly with V_{DS} (the *linear* regime), then gradually levels off, approaching a saturated value (the *saturation* regime). Thus, it behaves quantitatively as depicted by the basic output characteristic of the idealized MOSFET of Fig. 4-11.

We noted in section 4.4.2 that as V_{DS} is increased, the output characteristic levels off. This leveling off occurs because as V_{DS} increases, the channel near the drain is driven toward pinch off. In terms of the gradual channel approximation, pinch off is the point at which Q_I drops to zero. (In fact, once Q_I gets very small, the gradual channel approximation breaks down and Eq. 4-57 no longer applies.) When the V_{DS} reaches the

value at which pinch off has been reached, the measured I_D value in long-channel MOSFETs *saturates*. That is, it becomes essentially constant as V_{DS} is increased further. The I_D value at that point is denoted as I_{Dsat}. Although Eq. 4-57 is no longer valid for this regime, we use the maximum value of I_D found using Eq. 4-57 as I_{Dsat}. We denote the V_{DS} value to just reach this condition as $V_{DS,sat}$.

Since $V_{DS,sat}$ is the V_{DS} value when I_D reaches its maximum, such a maximum can be found by setting $dI_D/dV_{DS} = 0$ from Eq. 4-57 and solving for $V_{DS} = V_{DS,sat}$

$$V_{DSsat} = V_{GS} - V_{FB} - 2\varphi_B + \frac{\gamma^2}{2} - \gamma\sqrt{V_{GS} - V_{FB} + V_{BS} + \frac{\gamma^2}{4}} \qquad (4 - 59)$$

The value of $V_{DS,sat}$ obtained from Eq. 4-59 is used in Eq. 4-57 to calculate I_{Dsat}.

Note that the linear and saturation regimes of MOSFET operation have been defined. Since we also assume in this model that $I_D = 0$ for $V_{GS}<V_T$, we define this bias range as the *cutoff* regime. In fact, a small but nonzero drain current exists even when $V_{GS} <V_T$, so it is more appropriate to refer to this regime as the *subthreshold* regime. Further details on this regime are given in section 4.7.

4.5.5 Square-Law Model

Equation 4-57 predicts I_D values in long-channel MOSFETs (operated in strong inversion) quite accurately. However, even this formulation is mathematically too complex for use in hand calculations. Furthermore, when it is necessary to simulate the circuit behavior of digital IC circuits containing a large number of MOSFETs, using Eq. 4-57 in SPICE significantly slows the simulation time. For such applications, a simplified version of Eq. 4-57 could be useful, even at the expense of somewhat less accuracy. This can be obtained if $2\varphi_B + V_{BS}>>V_{DS}$. In this case, we can use the binomial expansion to get

$$(2\varphi_B + V_{SB} + V_{DS})^{3/2} = (2\varphi_B + V_{SB})^{3/2}\left(1 + \frac{V_{DS}}{2\varphi_B + V_{SB}}\right)^{1/2} \qquad (4 - 60)$$

$$\approx (2\varphi_B + V_{SB})^{3/2} + \frac{3}{2}(2\varphi_B + V_{SB})^{1/2} V_{DS} \qquad (4 - 61)$$

Thus,

$$I_D = \frac{\mu C_{ox} Z}{L}\left(V_{GB} - V_{FB} - 2\varphi_B - [2\kappa_{si}\varepsilon_o N_A (2\varphi_B + V_{BS})/C_{ox}]^{1/2} - V_{DS}/2\right)V_{DS} \quad (4 - 62)$$

which can be written more succinctly as

$$I_D = [\mu C_{OX} Z/L]\,(V_{GB} - V_T - V_{DS}/2)V_{DS} \qquad V_{GS} \geq V_T \quad (4 - 63a)$$

$$V_{DS} \leq (V_{GS} - V_T)$$

since $V_T = V_{FB} + 2\varphi_B + [2\kappa_{si}\varepsilon_o N_A(2\varphi_B + V_{BS})/C_{OX}]^{1/2}$. Or, Eq. 4-63a can be written in an even more compact way (again assuming that $V_{BS} = 0$, in which case $V_{GB} = V_{GS}$),

as

$$I_D = k\,[(V_{GS} - V_T)V_{DS} - V_{DS}^2/2] \qquad (4\text{-}63b)$$

where k (the so-called device transconductance parameter) is defined as

$$k = \mu\, C_{OX}\, Z/L \qquad (4\text{-}64)$$

For very small V_{DS} values, Eq. 4-63 can be further approximated as

$$I_D = k\,(V_{GS} - V_T)\,V_{DS} \qquad (4\text{-}65)$$

For $2\varphi_B + V_{SB} >> V_{DS}$, the condition for saturation becomes

$$-C_{OX}\,(V_{GS} - V_{FB} - 2\varphi_B - V_{DS,sat}) + (2\kappa_{si}\varepsilon_0 q N_A[2\varphi_B + V_{BS}])^{1/2} = 0 \qquad (4\text{-}66)$$

so

$$V_{DS,sat} \approx V_{GS} - V_T \qquad (4\text{-}67)$$

In saturation, I_D in this model is given by Eq. 4-62 evaluated at $V_{DS} = V_{DS,sat}$; that is

$$I_{Dsat} = k/2\,(V_{GS} - V_T)^2 \qquad V_{GS} \geq V_T: \qquad (4\text{-}68)$$

$$V_{DS} > (V_{GS} - V_T)$$

Since Eq. 4-68 predicts that I_{Dsat} will vary as the square of the gate voltage above threshold, Eq. 4-62 is termed the *square-law model.*

Equation 4-68 is often used as the basis for measuring V_T. In this technique, the gate and drain of the transistor are tied together as shown in Fig. 4-16, and then the drain current is measured as a function of applied drain voltage. Because V_{DS} is set equal to V_{GS} in this circuit, the transistor is in the saturated region of operation. Now a plot of

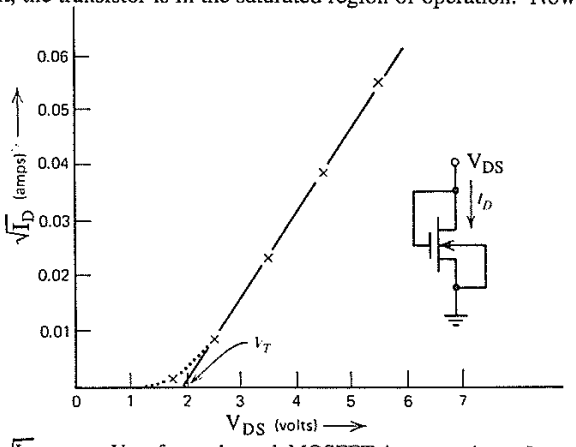


Fig. 4-16 Plot of $\sqrt{I_D}$ versus V_{DS} for n-channel MOSFET in saturation. Inset shows the circuit arrangement. The threshold voltage is indicated by the intercept of the straight line with the voltage axis.

$\sqrt{I_D}$ versus V_{GS} should be a straight line with a voltage axis intercept equal to V_T. In fact, plotting $\sqrt{I_D}$ versus V_{GS} is the most widely used method of determining V_T, yielding as will be discussed in section 4.6, the so-called *extrapolated threshold voltage* value.

From a physical perspective the square-law model can be derived by approximating that the channel depletion region has a constant width along the entire length of the channel (with a value equal to that near the source), even under $V_{DS} \neq 0$ biasing. This approximation is somewhat reasonable in strong inversion since the last term in the strong-inversion form of Q_I (Eq. 4-10), namely $[\sqrt{2\kappa_{si}\varepsilon_o q N_A (2\varphi_B + V_{ChB})}]$, is small compared to the other terms. Since it is the only term that depends on the depletion-region charge, and this charge is much smaller than the mobile charge concentration under strong inversion, the variation of this term will not change the value of Q_I much. That is, as long as the channel region remains inverted, Q_I does not vary much from the source to the drain ends of the channel. Hence, approximating this term by a constant value (namely with the charge that would exist if the depletion region had a constant width, equal to the value it at has at the source end) does not change the value of Q_I significantly. Yet it results in a much simpler form of the I_D model.

At the source the depletion region width is d_{max}(source), which can be calculated from Eq. 4-8

$$d_{max}\,(\text{source}) = [2\kappa_{si}\varepsilon_o(2\varphi_B + V_{BS})/qN_A]^{1/2} \qquad (4\text{-}69)$$

Then

$$Q_B = [2\kappa_{si}\varepsilon_o N_A(2\varphi_B + V_{SB}]^{1/2} \qquad (4\text{-}70)$$

and using Eq. 4-10

$$Q_I = -C_{ox}[V_{GB} - V_{FB} - 2\varphi_B - V_{ChB} - \gamma(2\varphi_B + V_{BS})^{1/2}]$$

$$= -C_{ox}[V_{GB} - V_T - V_{ChB}) \qquad (4\text{-}71)$$

Thus,

$$I_D = \frac{\mu\, C_{ox}\, Z}{L}\int_{V_{SB}}^{V_{SB}+V_{DS}} (V_{GB} - V_T - V_{ChB})\, dV_{ChB} \qquad (4\text{-}72)$$

Integrating 4-72 from source to drain gives 4-62.

4.5.6 Comparison of the Four MOSFET Circuit Models

Let us summarize the characteristics of the four long-channel MOSFET dc circuit models just described, and make some further comparisons about their relative accuracies and utility. Recall first off that all are based on the long-channel (GCA) approximation.

The most general and accurate of the models is the *Pao-Sah* model. While it retains the gradual-channel approximation, it does not invoke the depletion approximation and permits carrier transport in the channel by both drift and diffusion current. The formulation of the I_D equation it develops is therefore quite general, but also requires numerical integrations in two dimensions. Such numerical complexity has limited the role of this model role to theoretical analyses of the MOSFET. As a result the other three MOSFET models are more frequently used to calculate the I_D-V_{DS} characteristics.

The *charge sheet* model allows I_D to be calculated with much less computational effort than that of the Pao-Sah model and yet provides I_D values very close to those predicted by the more complex model over the regime of inversion (both weak and strong). The accuracy of the charge-sheet model when compared to measured dc I_D-V_{DS} characteristics of a MOSFET in inversion is about 1% or better. This model calculates the surface potential by solving the Poisson equation assuming the depletion approximation and the charge-sheet approximation. It also assumes that both drift and diffusion current can contribute to I_D. While not applicable to the depletion and accumulation regimes of operation, it is valid over all regimes of inversion. Although it accurately predicts I_D over these regions of operation, it requires values of the surface potential to be calculated at the source and drain ends of the channel. These equations for φ_{surf} are complex transcendental equations that require iterative techniques to solve them. As a result, the computational effort for determining I_D is still too great for most circuit simulation applications, and thus the charge sheet model has not been incorporated into such popular circuit simulation programs as SPICE.*

Since the charge-sheet model is too complex to be used in most circuit simulation applications, the simpler bulk-charge and square law models are used instead. They both assume that the MOSFET is biased to strong inversion, and that only drift current contributes to I_D. As long as the MOSFET is biased into strong inversion (V_H in Fig. 4-17) the results of the bulk-charge I_D equation agree very well with the charge-sheet model results. That is, to allow the simpler analytical expressions of the bulk-charge (Eqs. 4-57 and 4-59) and square-law models (Eqs. 4-63 and 4-68) to be used to accurately calculate I_D, the allowable regime of operation must be reduced to the strong-inversion regime. If the MOSFET is operated in weak inversion, an alternate analytical expression (as derived in section 4.7) must be used to get an accurate value of I_D in long-channel MOSFETs.

When comparing the utility of the *bulk-charge* and *square-law* models, the primary asset of the square-law model is its simplicity. It provides good insight into the device operation and is useful for obtaining first-order design data without an excessive amount of mathematical entanglement, particularly for the case of digital circuits. Most hand calculations for circuit design with MOSFETs therefore make use of the square-law model. It is also the model used for the simplest (Level-1) analyses of circuits containing MOSFETs in the SPICE program. The apparently arbitrary adjustment of k

* Note that the bulk-charge and square-law MOSFET models assume that the surface potential at inversion is fixed at $2\varphi_B + V_{CB}$. In reality this is the value of φ_{surf} at which the surface just enters strong inversion. If the electron charge in the channel is substantial, the surface potential must be greater than this value (but not much more than a few kT above $2\varphi_B + V_{CB}$). Nevertheless, both these models must therefore consistently underestimate the φ_{surf}.

Now, if φ_{surf} is underestimated, so is the magnitude of the depletion region charge Q_B. If Q_B is underestimated, then from Eq. 4-1b Q_I must be overestimated. The bulk charge model therefore consistently overestimates the amount of mobile electron charge in the channel, and thus overestimates the drain current - fortunately by less than 10 percent in most cases.

The charge sheet model, on the other hand does not make the above assumption, but instead calculates φ_{surf}. Hence, this error is avoided in the charge-sheet model approach.

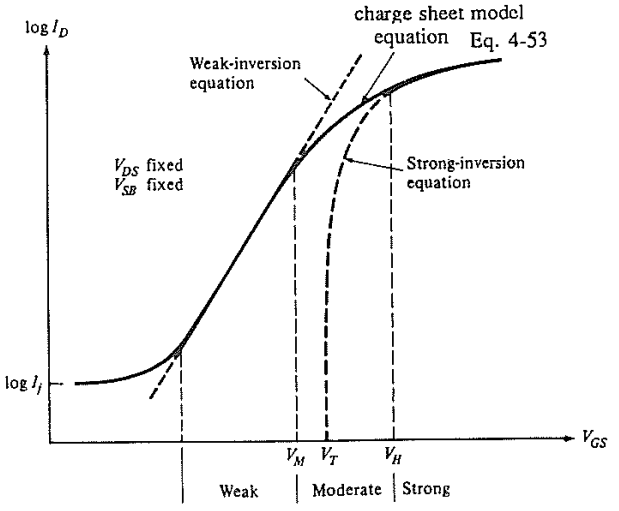


Fig. 4-17 Log I_D versus V_{GS} for fixed V_{DS} and V_{BS}, and comparison to weak- and strong-inversion equations.[5] Reprinted with permission of McGraw-Hill Book Co.

mentioned above is also not a practical problem since MOS processing procedures do not allow for the precise control of μ_n, and the parameter must be extracted from measurements.

There are many cases of circuit design for which the increased accuracy of the bulk-charge model is not necessarily an advantage when computational complexity is considered. Often, for example, the designer is interested only in MOSFET operation in the saturated region (i.e., only the value of I_{Dsat} is needed). To obtain I_{Dsat} from the bulk-charge model, it is first necessary to calculate V_{DSsat} from Eq. 4-59 and then use this voltage value in Eq. 4-58 to calculate I_{Dsat}. This lengthy process is greatly simplified in the square-law model in which I_{Dsat} is calculated directly from Eq. 4-68.

The bulk-charge model is used when more accurate modeling of circuit behavior is needed. Measurements of I_D versus V_{DS} characteristics of long-channel MOSFETs are in reasonable agreement with the predictions of this model (i.e., usually within 10% of the measured data). However, the computational complexity is considerably increased, and hence it is typically implemented only in computer analyses. This model is available as the Level-2 model of the MOSFET in SPICE. Note that when comparing Eqs. 4-57 and 4-59 with 4-63 and 4-68, the added terms in the bulk-charge model equations are always negative and act primarily to reduce I_D and V_{DSsat} for a given set of operational conditions.

A plot of I_D versus V_{DS} (with V_{GS} as a parameter) for a long-channel NMOS transistor calculated using both models is shown in Fig. 4-18. As can be seen the square-law model predicts larger values of I_{Dsat} and V_{DSsat} for each value of gate

voltage. The general behavior of both sets of curves is, however, similar. Note especially that in the linear region of operation both models give the same results, but that in saturation - for the same value of k in both models - the bulk-charge model predicts a smaller value of I_{Dsat}.

4.5.6.1 MOSFET Models of the SPICE Circuit Simulator. The SPICE circuit simulator contains four MOSFET models which differ in the way they formulate the I-V characteristic. The simplest is the Level-1 model which is based on the rudimentary square-law model (Equations 4-63 and 4-68), with empirical adjustment. The most important adjustment it employs is the use of a coefficient in these equations to bring I_{Dsat} into congruence with experiment and the more accurate models. That is, to obtain satisfactory results in circuit analysis and design using this model, measured data on samples of MOSFETs must be obtained. Measured data is fitted to the threshold voltage equation and Eqs. 4-63 and 4-65. Specifically V_T and k are determined from data obtained as shown in Fig. 4-16. Note that the extrapolated threshold voltage value V_{T0} is used in SPICE because it is obtained by extrapolating the experimental curve of $\sqrt{I_D}$ versus V_{DS} (with $V_{DS} = V_{GS}$) to zero drain current. The subscript "0" in the V_{T0} symbol signifies that this is the extrapolated threshold voltage when $V_{BS} = 0V$.

The more-accurate Level-2 SPICE MOSFET model is based on the bulk-charge model equations (Eqs. 4-57 and 4-59). The Level-2 model also incorporates an approximate*

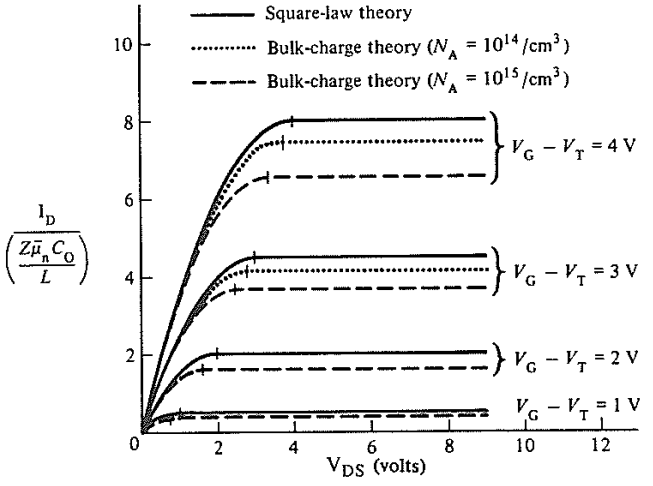


Fig. 4-18 Theoretical predictions of I_D versus V_{DS} using the bulk-charge model equations (solid lines) and the square-law model equations (dashed lines). I_D is taken to be equal to I_{Dsat} for $V_{DS}>V_{DSsat}$.[30] Reprinted with permission of Addison-Wesley.

* For example, in the Level-2 MOSFET model, the carrier mobility is not treated as a constant. Instead, the mobility is modeled as being impacted by the gate-induced electric field (see chap. 5). The value used in the Level-2 model μ_{eff} is found using the expression Eq. 5-26.

treatment of some of the short-channel effects discussed in chap. 5 and is capable of computing the subthreshold current (in the manner as described in section 4.7).

The basic device structure parameters required to describe a MOSFET when using either the Level-1 or Level-2 model are the substrate doping, the gate oxide thickness, and the channel length and width. SPICE computes the threshold voltage from N_A (NSUB in SPICE) and t_{ox} using Eq. 4-12 (assuming $V_{FB} = 0$). Alternatively, the threshold voltage (for the case of $V_{SB} = 0$) can be input directly by supplying a value to the model parameter V_{T0}, in which case the internally computed threshold voltage is overridden.

The Level-2 SPICE MOSFET model, being based on the bulk-charge MOSFET model, has a three-halves-power dependence of I_D, rather than the square-law dependence of the Level-1 SPICE MOSFET model. Because the numerical handling of a fractional power requires significantly more computation time than if a power in integer form must be computed, circuit designers sought a model that was more accurate than the Level-1 model, but was faster to execute than the Level-2 model. Hence, the Level-3 model was developed and incorporated into SPICE.*

The Level-3 model not only removed the computational shortcoming of the Level-2 model, but was formulated to deal empirically with many other effects as well, including short-channel and narrow-channel effects. In fact, the Level-3 model represents an attempt to pursue a semi-empirical modeling approach, i.e., which only approximates device physics and relies on the proper choice of the empirical parameters to accurately reproduce device characteristics. As a result, a more physically based short-channel MOSFET model has also been developed for SPICE, known as the Berkeley Short-channel IGFET Model (BSIM). It has been implemented into the SPICE program as the Level-4 MOSFET model. We will describe BSIM in more detail in chap. 5.

4.5.6.2 SPICE Models of the MOSFET Operated in Saturation. As we noted in section 4.4.2, the value of I_D may increase as V_{DS} exceeds the pinchoff voltage value V_{DSsat}. This was modeled by assuming that the pinched-off region with a length ΔL makes the channel length appear shorter (i.e., $L - \Delta L$). To predict I_D as V_{DS} exceeds V_{DSsat} the value of ΔL is needed. In the Level-1 and Level-2 MOSFET models

* The I_D equation that constitutes the dc long-channel Level-3 model of the MOSFET[10] in the SPICE program is derived from Eq. 4-57 by employing a three-term binomial-series expansion for the second term in Eq. 4-57. This yields the Level-3 MOSFET equation for I_D in the form:

$$I_D = \frac{\mu C_{ox} Z}{L}\left[(V_{GS} - V_T)V_{DS} - \left(\frac{1}{2} + \frac{\sqrt{\varepsilon_o q N_A/(-\varphi_B)}}{4 C_{ox}}\right)V_{DS}^2\right]$$

Notice by comparing it to the Level-1 expression, Eq. 4-63, that it differs only in the last term, a change that accounts for the ionic-charge variations that were ignored in the Level-1 model. Here, however, the square root term is only dependent on device structure parameters and not on V_{DS}, and hence this square-root term only needs to be evaluated once when calculating the I_D-V_{DS} characteristics.

SPICE offers two possible approaches to calculating ΔL. The first allows the *channel-length modulation parameter*, λ (LAMBDA) to be specified, in which case ΔL is obtained from the simple empirical expression $\Delta L = \text{LAMBDA} \cdot V_{DS}$. LAMBDA can be used in either the Level-1 or Level-2 model. The value of LAMBDA is typically in the range of 0.1 to 0.01 V^{-1}. Empirical values of $1/\lambda$ can be obtained by finding an approximate intercept with the V_{DS} axis of tangents to the I_D curves beyond V_{DSsat} in measured I_D-V_{DS} curves.

If the value of LAMBDA is not specified in the Level-2 model, however, then ΔL is computed by using Eq. 5-47, which is based on the 1-D Reddi-Sah model of the MOSFET operated in saturation (see chap. 5, section 5.4.3.1).

4.6 THRESHOLD-VOLTAGE CONTROL IN MOSFETS

4.6.1 Measuring Threshold Voltage

Equation 4-1 calculates V_T of real MOSFETs based on various device parameters. Yet, when devices are fabricated, they must be tested to ensure that the desired characteristics have been produced. Two approaches are used to obtain the V_T values of MOSFETs from measurements. The first yields an *extrapolated-V_T* value, and the second, a *constant-current-V_T* value.

As described in section 4.5.5, the extrapolated V_T is determined by connecting the drain and gate of a MOSFET together and measuring the drain current as a function of gate voltage (Fig. 4-16). Equation 4-68 predicts that a plot of $\sqrt{I_D}$ (with a linear I_D axis) versus V_{GS} should be linear. If $\sqrt{I_D}$ is plotted vs. V_{GS}, and the linear portion of the curve is extrapolated to intercept the voltage axis (corresponding to the condition that $V_{GS} = V_T$ in Eq. 4-68) the extrapolated-V_T value is yielded. (Note that the dotted section of the curve in Fig. 4-16 results from currents that flow below the threshold voltage. Such *subthreshold currents* are discussed in section 4.7.)

The constant-current-V_T is obtained by measuring the gate voltage at which a specific small drain current flows and this gate voltage is defined as V_T. For active MOSFETs, the gate voltage at which 1 μA of drain current flows per μm of drain width for V_{DS} = 0.5V is the most common criterion applied when specifying constant-current-V_T. The constant-current-V_T is generally measured in the IC production environment, while the extrapolated V_T is preferred when the threshold voltage parameter is needed for circuit simulation (e.g., in a SPICE analysis). Note also that the extrapolated V_T is typically 100 mV larger than the 1 μA/μm constant-current V_T.

4.6.2 Adjusting Threshold Voltage

As pointed out earlier, in many MOS IC applications it is critical to be able to establish and maintain a uniform and stable V_T. The factors that impact V_T are given in Eq. 3-72. An examination of each term will reveal the device parameters that can be adjusted to provide practical control of V_T.

The φ_{ms} term depends on the work-function difference between the gate, $q\varphi_m$(gate), and the semiconductor, $q\varphi_B$ (sub). While $q\varphi_m$(gate) for metal and heavily doped silicon

gates is constant, the parameter φ_B (sub) depends on the substrate doping - but only in a logarithmic manner. Hence, each factor-of-10 increase in substrate doping will change the φ_{ms} term by only 2.3 kT/q, or ~0.06 V (kT/q = 0.026 V at 300K). Thus, changes in the substrate-doping concentration produce changes in V_T through the φ_{ms} term which are too small to provide the required degree of threshold-voltage control. The next term ($2\varphi_B$) also changes only slightly as a result of changes in the substrate doping concentration (for the same reason given for the φ_{ms} term); thus, the $2\varphi_B$ term is also ineffective for controlling V_T.

Since every attempt is made to keep Q_{tot} as low as possible through various processing procedures, and C_{ox} is relatively large (since t_{ox} is very thin in submicron MOSFETs), the Q_{tot}/C_{ox} term is also very small in modern MOSFETs. Hence, this term must also be ruled out as a candidate for controlling V_T.

While is true that C_{ox} could be varied (primarily by changing t_{ox}), this is not a practical approach to controlling V_T in active devices, since t_{ox} is normally made as thin as possible to maximize I_D. (In the field regions, however, large V_{TF} values are needed to prevent inversion under the field oxide. A thick field oxide makes C_{ox} small, allowing V_{TF} to be increased Thus, C_{ox} is one of the parameters normally used to control V_{TF} in the field regions of the circuits.)

This leaves the $2\sqrt{\kappa_{si}\varepsilon_o q N_{sub}\varphi_B}/C_{ox}$ term as the remaining candidate for controlling V_T in active devices. Since the change of V_T in this term depends on $\sqrt{N_{sub}}$, it indicates that V_T control is possible by exploiting this effect. In fact, Eq. 3-72 indicates that increasing the substrate doping (i.e., N_A in n-channel devices, and N_D in p-channel devices) will increase V_T. Figure 4-19 shows how V_T varies in MOSFETs (both PMOS and NMOS) with substrate doping concentration (assuming the substrate is uniformly doped - i.e., no threshold adjust implant is used - and that $Q_{tot} = 0$), for various gate-oxide thicknesses.

Merely increasing the substrate doping, however, is not desirable since it will adversely impact other MOSFET characteristics, such as lower junction-breakdown voltages, larger junction capacitances, and lower carrier mobilities. Yet, prior to the development of ion implantation in the early 1970s, adjustment of substrate doping was the only practical *processing* approach for significantly controlling V_T in active devices.

We showed earlier that a *circuit* approach based on applying a bias between the source and body (and hence known as body biasing) can also control V_T (section 4.2.2.2, Eq. 4-12). Generally, an increase in V_{BS} will raise V_T. Figure 4-20 gives an example of how the application of V_{BS} can change the V_T value in an NMOS device. The intercept of the channel conductance and the V_{GS} axis when $V_{BS} = 0$ in this example is +0.99V. As V_{BS} is increased from 0V to -5V, V_T increases up to +2.3V. However, body biasing is an added complication, and is now avoided whenever possible as a technique for adjusting V_T, in favor of the ion implantation V_T-adjust method.

4.6.3 Ion Implantation for Adjusting Threshold Voltage

The development of ion implantation for V_T adjustment removed the last obstacle to reliable production of n-channel devices for MOS ICs, as this procedure made it possible to select the substrate-doping value without having to consider its impact on V_T. That

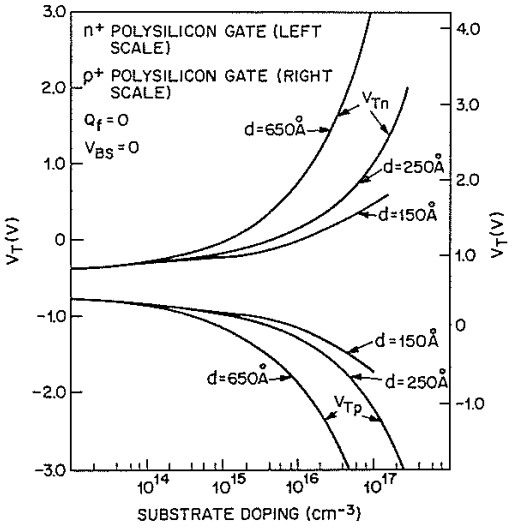


Fig. 4-19 Calculated threshold voltages of n-channel (V_{Tn}) and p-channel (V_{Tp}) transistors as a function of their substrate's doping, assuming n^+ polysilicon gate (left scale) and p^+ polysilicon (right scale). Curves for gate oxide thicknesses t_{ox} of 150 Å, 250 Å, and 650 Å are shown. From S. M. Sze, Ed., VLSI Technology, 2nd. Ed., Chap. 11, "VLSI Process Integration." Copyright, 1988 Bell Telephone Laboratories. Reprinted with permission of McGraw-Hill.

is, implantation can be used either to increase or to decrease (by compensation) the net dopant concentration at the silicon surface. As a result, substrate doping can be selected strictly on the basis of optimum device performance since V_T can now be set by the V_T-

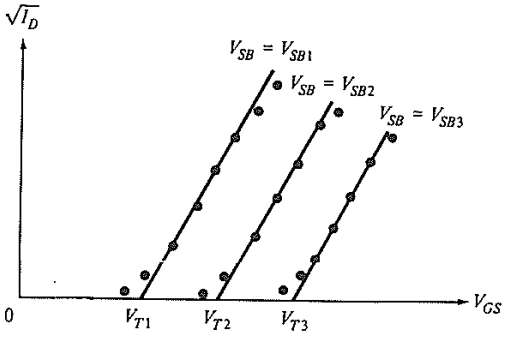


Fig. 4-20 Plot of $\sqrt{I_D}$ vs. V_{DS} for an NMOSFET showing the effects of a bias V_{BS} between the source and the substrate. The observed shift in the threshold is predicted by Eq. 3-71.[16] Reprinted with permission of John Wiley & Sons.

adjust implant process. In addition, since dopants can be selectively implanted into the *field regions*, high-performance NMOS circuits can be fabricated on lightly doped substrates, without the possibility of inadvertent inversion of the surrounding field regions.

As mentioned earlier, the V_T-adjust-implant technique involves implantation of boron, phosphorus, or arsenic ions into the regions under the gate oxide of a MOSFET. Boron implantation produces a positive shift in V_T, while phosphorus or arsenic implantation causes a negative shift. For shallow implants, the procedure has essentially the same effect as placing an additional sheet of "fixed" charge at the SiO_2-Si interface.

To first order, if the implant and substrate are of the same type (e.g., boron into a *p*-substrate), and the effective depth of the implant x_i (see Fig. 4-21) is less than d_{max}, the threshold-voltage change (ΔV_T) can be reasonably well estimated from[11]

$$\Delta V_T\,(\mathrm{II}) = q\, D_I / C_{ox} \qquad (4 - 73)$$

Example 4-1: If the V_T of a uniformly doped NMOS transistor is -0.41V and it is to be used in an application that requires a V_T = 1.0V, calculate the B dose (to first order) needed to adjust $V_{T(nud,lc)}$ to this value.

Solution: The $V_{T(ud,lc)}$ of the device is -0.41 V. We wish to have a V_T of 1.0 V. Thus, the threshold voltage must be shifted (ΔV_T) by +1.41 V. Using Eq. 4-73, we see that the boron dose needed to cause this ΔV_T is

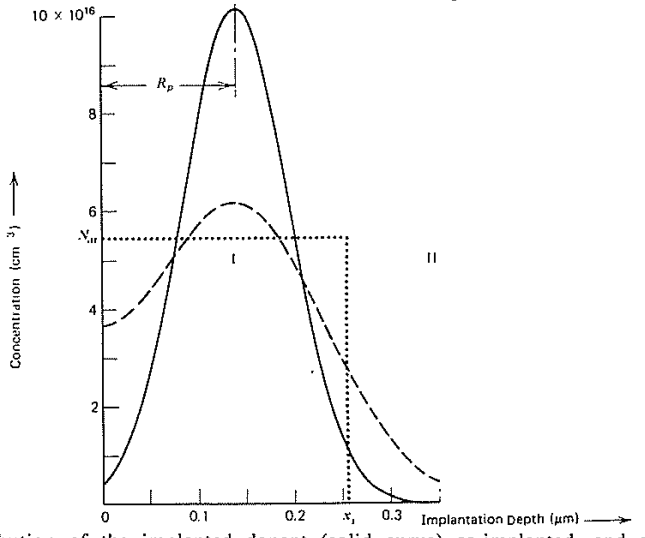


Fig. 4-21 Distribution of the implanted dopant (solid curve) as-implanted, and after activating anneal and diffusions (dashed curve). The dotted "box" curve shows the distribution used to calculate the threshold voltage. (After Rideout, Gaensllen & LeBlanc in ref. 27.)

$$D_I = \Delta V_T(II)\, C_{ox}/q = [1.41V \times 2.3\times10^{-7}\ F/cm^2]/1.6\times10^{-19}\ C$$

$$= 2.1\times10^{12}\ \text{boron atoms}/cm^2.$$

where D_I is the dose of the implanted ions (atoms/cm^2) which penetrate into the silicon. For example, Eq. 4-73 predicts that when $D_I = 5\times10^{11}$ ions/cm^2 and $t_{ox} = 25$ nm, a shift in V_T of 0.58 V will be produced. Thus, Eq. 3-72 must be modified to account for threshold-adjust implants, as (where *nud* stands for non-uniformly doped)

$$V_{T(nud,lc)} = V_{T(ud,lc)} + \Delta V_T(II) \qquad (4-74)$$

A more exact formula for calculating for V_T in this case (assuming $V_{BS} = 0$) is[12,13,14]

$$V_{T(nud,lc)} = V_{FB} + 2\varphi_B + qD_I/C_{ox} + \gamma_2\left(\sqrt{2\varphi_B + qD_I x_i/2\kappa_{si}\varepsilon_o}\right) \qquad (4-75)$$

where

$$\gamma_2 = \left(\sqrt{2q\kappa_{si}\varepsilon_o N_A}\right)/C_{ox} \qquad (4-76)$$

The V_T-adjust implant is often done through an oxide layer.* The implant energy is selected to place the peak of the implant slightly below the oxide-silicon interface (Fig. 4-21). After the implant-activating anneal, the implanted distribution is broader than the as-implanted profile.

Calculating the effect of the implant on V_T is greatly simplified by approximating the actual distribution via a "box" distribution, as is also shown in Fig. 4-21 (in which the implanted dopant is assumed to have a constant density N_A' from the surface to the depth x_i.) Figure 4-22 plots V_T vs. effective implant depth x_i, and we can see that V_T does not change much as x_i is varied. Since the second term under the square root in Eq. 4-75 is the only one that depends on x_i, if we ignore it when calculating V_T the results are still adequate for many applications. Thus, the first-order approximation of ΔV_T given by Eq. 4-73 is useful for hand calculations of $V_{T(nud,lc)}$.

Another slightly more complicated approach to calculating V_T for this case is to solve the Poisson equation using a high-low step doping profile, which results in an effective concentration N_{eff} given by[28]

$$N_{eff} = N_A' \frac{\varphi_s}{\varphi_i}\left[1 - \frac{N_A}{N_A'} + \frac{N_A'}{N_A}\sqrt{1 + \frac{N_A}{N_A'}\left(\frac{\varphi_s}{\varphi_i} - 1\right)}\right]^2 \qquad (4-77)$$

where $\varphi_s = kT/q\ [\ln(N_A'N_A/n_i^2)]$ and $\varphi_i = qN_A'x_i^2/2\kappa_{si}\varepsilon_o$. Here φ_i is the potential corresponding to x_i of the step profile. N_{eff} is then substituted for N_A in the body effect coefficient term γ of Eq. 4-12, and then Eq. 4-12 with this value of N_{eff} is used to calculate the value of V_T in MOSFETs with implanted channels. This method for calculating V_T in MOSFETs with non-uniformly doped channel regions is often called the *doping transformation procedure,* and reportedly gives excellent results.

Ion implantation can also be used to fabricate depletion-mode MOSFETs. Depletion-mode NMOS devices (i.e., in which $V_T<0V$) are commonly used in NMOS logic circuits. In order for the required negative threshold voltage for a depletion-mode NMOS device to be produced, *n*-type impurities are implanted into the *p*-type substrate to form a built-in channel between the source and drain. The dose required to shift the threshold voltage by the desired value may also be estimated using Eqs. 3-72 and 4-73.*

4.6.3.1 Impact of the V_T-adjust Implant on the Body Effect. The implanted layer also impacts the variation of V_T with V_{SB}. This is shown in Fig. 4-23. We observe that for uniformly doped substrates [curves (a) and (b)] the variation of V_T with V_{SB} follows the data shown in Fig. 4-20. For shallow implants [curve (c)], the V_T vs. V_{SB} behavior approaches that for a delta function dose at the surface. That is, V_T is observed to be merely shifted by qD_I/C_{OX} from the V_T values given by curve (a), and thus curve (c) is parallel with curve (a). However, for a deeper implant [curve (d)], the initial variation in V_T with V_{SB} is comparable to that for a uniformly doped region of concentration N_A' [i.e., curve (b)]. However, as V_{SB} is increased, the channel depletion region width d approaches and soon exceeds x_i, and the variation of V_T with V_{SB} thereafter behaves as if the substrate is doped with the bulk concentration N_A [like curve (c)]. Since it is generally desirable to have V_T exhibit a low substrate bias sensitivity

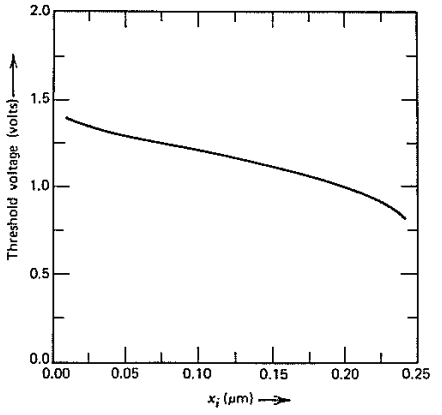


Fig. 4-22 Dependence of the threshold voltage on the effective implantation depth x_i as predicted by Eq. 4-75 for a MOSFET with an 87 nm-thick gate oxide, $Q_f/q = 10^{11}$ cm^{-2}, $D_I = 3.5\times10^{11}$ cm^{-2}, and $N_A = 2\times10^{15}$ cm^{-3}. Both the substrate and source terminal are assumed be grounded.[11] Reprinted with permission of John Wiley & Sons.

* Note that if x_i is greater than d_{max}, $V_{T(nud,lc)}$ can be calculated using Eq. 3-72, with N_A replaced by $N_A' = N_A + D_I/x_i$. However, this condition is generally avoided when choosing the V_T-adjust implant parameters, since the increased doping near the substrate edge of the source and drain junctions leads to higher junction capacitance, lower breakdown voltage, and a larger body effect.

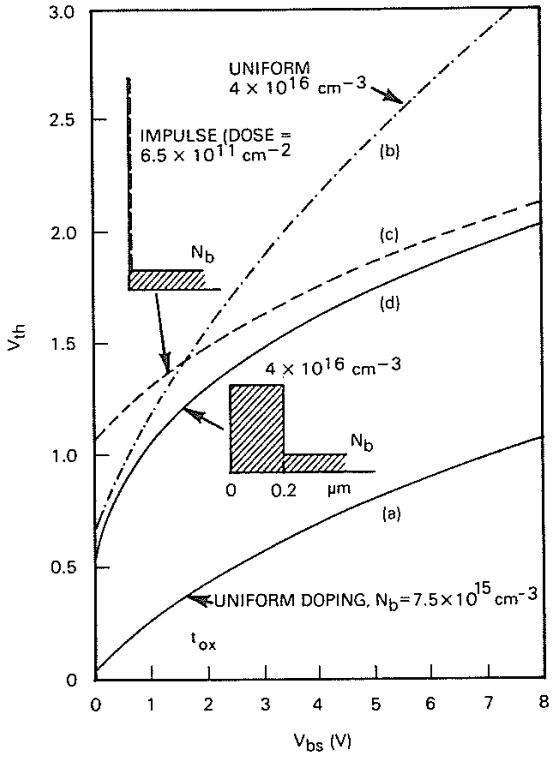


Fig. 4-23 Calculated substrate sensitivity for various doping profiles. (After Rideout, Gaensllen and LeBlanc in ref. 27.)

(i.e., small body effect), Fig. 4-23 also demonstrates that a shallow implant with an appropriate dose can raise V_T while maintaining a small body effect.

4.6.4 Impact of V_T-Adjust Implant on Other MOSFET Characteristics

By implanting ions near the Si surface, a nonuniformly doped channel region is produced. Besides shifting V_T, this alters the MOSFET behavior in several ways compared to devices having uniformly doped channels. First, the V_T-adjust implant introduces extra ions into the channel depletion region, causing its width to be modified. This effect is discussed here. In addition, the non-uniform doping profile changes the long-channel subthreshold characteristics (i.e., the slope of the log I_D vs. V_{GS} curves) as well as the punchthrough behavior of short-channel devices. These two latter effects will be discussed later, in chap. 5.

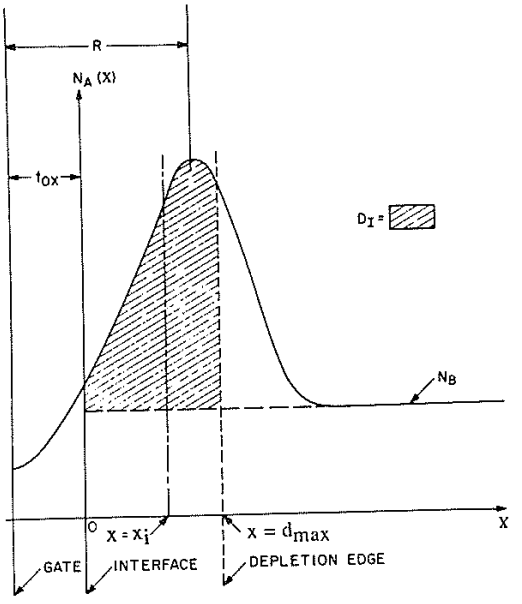


Fig. 4-24 Schematic of the threshold adjust doping profile in an implanted MOSFET. The shaded portion of the profile is the portion of the implanted dose in the depletion region, whose depth extends to d_{max}. The effective depth of the implant (or centroid) is labeled x_i.[8,15] Reprinted with permission of Solid-State Electronics.

However, all of the above effects depend only upon the so-called *exposed dose* D_I (i.e., the portion of the implant dose that ends up in the depletion region), and upon the *centroid* x_d of the exposed dose. That is, those implanted ions which do not penetrate the screening oxide or which penetrate beyond the edge of the channel depletion region have no impact. Thus, these terms must be defined. D_I (the shaded portion in Fig. 4-24), and x_d are given in terms $N_{sub'}(x)$ and N_{sub} by[15]

$$D_I\,(\text{cm}^{-2}) = \int_0^d [N_{sub'}(x) - N_{sub}]\,dx \qquad (4\text{-}78)$$

and

$$x_d = \frac{1}{D_I}\int_0^d [N_{sub'}(x) - N_{sub}]\,x\,dx \qquad (4\text{-}79)$$

where $N_{sub'}(x)$ is the true channel profile, including the implanted ions, and N_{sub} is the uniform background doping of the substrate, as depicted in Fig. 4-24.

The change in width of the channel depletion region is most easily illustrated by replacing the actual implant profile shown in Fig. 4-24 with a simpler delta-function equivalent implant. That is, the equivalent doping profile $N_{sub'}(x)$ is given by

$$N_{sub'}(x) = N_{sub} + D_I\,\delta(x - x_d) \qquad (4 - 80)$$

Using such a delta-function replacement for the actual profile, the effects of the V_T-implant are illustrated in Fig. 4-25. Figure 4-25a shows a potential versus distance diagram [$\varphi(x)$ versus x] of a long-channel MOSFET with a non-implanted (i.e., uniformly doped) channel region. The channel depletion region width is d and the band bending is φ_{surf}. In Fig. 4-25b, the effect of varying the *dose* D_I on the same depletion region is shown. The presence of D_I at x_d causes a larger voltage drop to occur between x_d and the surface, leaving a smaller portion of φ_{surf} to deplete the region behind the implant. Hence, the total width is reduced to $d_I = d_{DI}$. Doubling D_I further reduces the width (in Fig. 4-25b to d_{2DI}). Eventually, for a value of dose termed the *effective* dose, D_{eff}, d_I is reduced to a point $d_I = x_d$. Any increase of dose beyond D_{eff} simply is neutralized in the bulk, and the condition of $d_I = x_d$ is maintained. In other words, D_{eff} is the minimum dose needed to bring the depletion edge to that depth, making $d_I = x_d$.

Fig. 4-25c depicts the result of increasing implant *depth*. In this case, increasing x_d also increases the voltage drop between x_d and the surface, again reducing the width to a

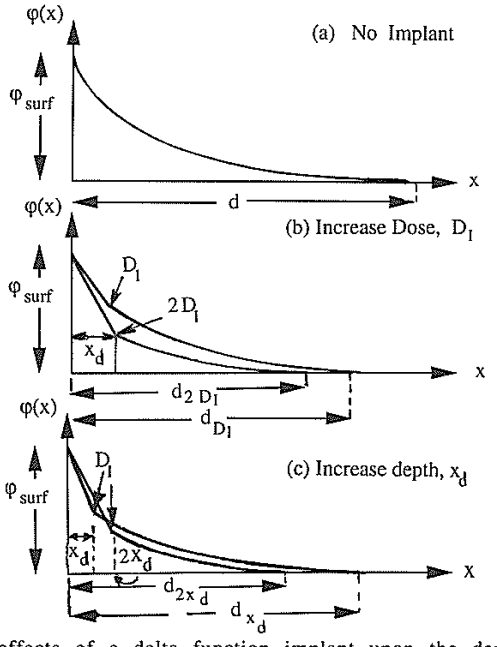


Fig. 4-25 The effects of a delta function implant upon the depletion width. (a) Unimplanted, (b) Effects of changing dose, and (c) Effect of changing depth.[13] (© IEEE 1978)

smaller d_I. As a deeper x_d is used, d_I moves toward the implant. For example, in Fig. 4-25c, we see that for an implant with depth x_d, $d_I = d_{xd}$. This is deeper than d_I due to a an implant with a depth $2x_d$ (here $d_I = d_{2xd}$). Ultimately, $d_I = x_d$, and $D_I = D_{eff}$, corresponding to this depth. Any further increase in x_d carries the depletion region edge with the implant. The depletion region edge becomes *clamped* to the implant.

The quantitative variation of d_I can also be simply expressed by again assuming a delta-function implant. That is, the depletion-region width in the presence of a V_T-adjust implant d_I is calculated from

$$d_I = \sqrt{2}\, L_D\, [\varphi_{surf}/\varphi_t - D_I x_d/(N_{sub} L_D^2) - 1]^{1/2} \qquad (4 - 81)$$

where L_D is the Debye length $\{[\kappa_{si}\varepsilon_o\varphi_t/qN_{sub}]^{1/2}\}$. Eq. 4-81 shows that the implant reduces the width, since in a uniformly doped channel the width d is given by

$$d = \sqrt{2} L_D\, [\varphi_{surf}/\varphi_t - 1]^{1/2} \qquad (4 - 82)$$

Likewise, D_{eff} is given by

$$D_{eff} = N_{sub}\, (d^2 - x_d^2)/2x_d. \qquad (4 - 83)$$

4.7 SUBTHRESHOLD CURRENTS IN LONG-CHANNEL MOSFETs (I_{Dst} when $V_{GS} < |V_T|$)

The small drain current which flows in the MOSFET channel below threshold (i.e., in weak inversion) is called *subthreshold current*, I_{Dst}. In most applications I_{Dst} is far too small to be useful as a drive current. However, it can represent an unwanted leakage current, especially in ICs designed for low-power applications. It can also give rise to premature discharging of memory-cell capacitors or dynamic-logic circuit nodes, preventing these circuits from operating properly. The issues involving leakage currents in MOSFETs will become even more important with the proliferation of battery-powered (i.e., portable) electronic systems. The common range of V_T in submicron digital CMOS ICs is 0.6-0.8V. Thus, when V_{GS} = 0V the MOSFETs in these circuits may be close to weak inversion. Consequently, for many applications (especially those requiring low standby-power dissipation), subthreshold leakage must be well characterized so that the total IC leakage current of the chip can be predicted during the design phase of the product.

In actual MOSFETs, when $V_{GS}<V_T$ the minority carrier concentration in the channel (e.g., n_{ps}) is small (but not zero). That is, as V_{GS} is increased from 0V and approaches V_T, the mobile charge concentration due to n_{ps} remains smaller than the ionic charge of the bulk doping concentration, but continues to increase as V_{GS} rises. In section 3.3.5 we defined *weak inversion* in terms of φ_{surf}, namely as $\varphi_B<\varphi_{surf}<2\varphi_B$.

The bulk-charge and square-law models of I_D are not suitable for treating subthreshold behavior because they were developed to describe MOSFET behavior only in strong inversion. That is, both models assume that $I_D = 0$ if $V_{GS}<|V_T|$. While the Pao-Sah model can provide accurate predictions of I_{Dst}, it requires too much computation to be practical for circuit simulation applications. Thus, the charge-sheet model becomes the candidate for characterizing I_{Dst}. We noted in section 4.5.3,

however, that the I_D expression of the charge-sheet model in its full form is also still too cumbersome for most circuit-design simulation tasks. Fortunately, by making some appropriate simplifying approximations, the general form of the charge-sheet model I_D equation can be reduced to a less complex formulation that still provides accurate predictions of I_{Dst} in long-channel MOSFETs.

Consider the case of a long-channel NMOSFET in which: (1) the source is grounded; (2) V_{GS} approaches (but remains less than) V_T, and; (3) $|V_{DS}| \geq 0.1V$. Now recall that in inversion (i.e., when $V_{GS} > |V_T|$), the free-carrier concentration n_{ps} in the channel is large (e.g., $n_{ps} \geq N_A$), and I_D in this case is dominated by drift current. However, when $V_{GS} < |V_T|$, n_{ps} is much smaller than N_A, but is not zero In addition, in weak inversion the drain voltage V_{DS} drops almost entirely across the reverse-biased drain-substrate depletion region.

Figure 4-26a illustrates the variation of inversion-layer charge concentration versus channel position in a long-channel NMOSFET in weak inversion. Figures 4-26b and c depict $\varphi_{surf}(y)$ and a 2-D perspective plot of the potential $\varphi(x,y)$ diagrams in weak inversion, respectively. We see that the band bending in the *x*-direction is too small to establish a strongly inverted channel [i.e., $\varphi_B < \varphi_{surf} < 2\varphi_B$], thus the channel mobile charge density Q_I is very small. In the *y*-direction most of the band bending occurs at the drain junction. Since the value of φ_{surf} along the *y*-direction remains essentially constant over most of the length of the channel, the $\mathcal{E}$-field along the *y*-direction is also very small [i.e., $\mathcal{E}(y) = -\partial\varphi_{surf}/\partial y$ is very small]. The combination of a small lateral $\mathcal{E}$-field acting on a relatively small number of free carriers implies that the drift component of I_{Dst} is negligibly small in weak inversion. In addition, in long-channel MOSFETs, the *gradient* of the $\mathcal{E}$-field in the *y*-direction along most of the channel length must also be very small. This means it is valid to apply the GCA when calculating I_{Dst} in such devices. In short-channel MOSFETs, however, the gradient of $\mathcal{E}(y)$ can no longer be neglected, even in weak inversion, and thus the long-channel formulation of I_{Dst} given here may not yield accurate estimates of I_{Dst} in short-channel MOSFETs.

On the other hand, we shall show that the *gradient of the free carriers* along the *y*-direction in the channel in weak inversion can be relatively large. Since diffusion current is proportional to the carrier-density gradient, in such cases carrier diffusion can produce a significant I_{Dst}. In fact, diffusion is the dominant mechanism that establishes subthreshold current flow in long-channel devices. The expression for the diffusion current component is given by:

$$I_{diff} = A\, J_{diff} = A\, [q\, D_n\, dn\,(y)/dy] = Z\, t_{inv}\, [\, q\, D_n\, dn\,(y)/dy\,] \qquad (4 - 84)$$

where D_n is the electron diffusion constant, A is the cross-sectional area of the electron flow, and t_{inv} is the inversion layer thickness (or $A = Z\, t_{inv}$). We can express this equation in terms of the mobile charge density per unit area Q_I instead of in terms of *n* by multiplying the mobile charge density in the channel (*qn*) by the inversion layer thickness t_{inv}. As shown in section 4.3.1.1, t_{inv} can be approximated as $(kT/q)\mathcal{E}_{si}$ (x = 0), where $\mathcal{E}_{si}$ (x = 0) is given by $\sqrt{2qN_A\varphi_{surf}/\kappa_{si}\varepsilon_o}$, so

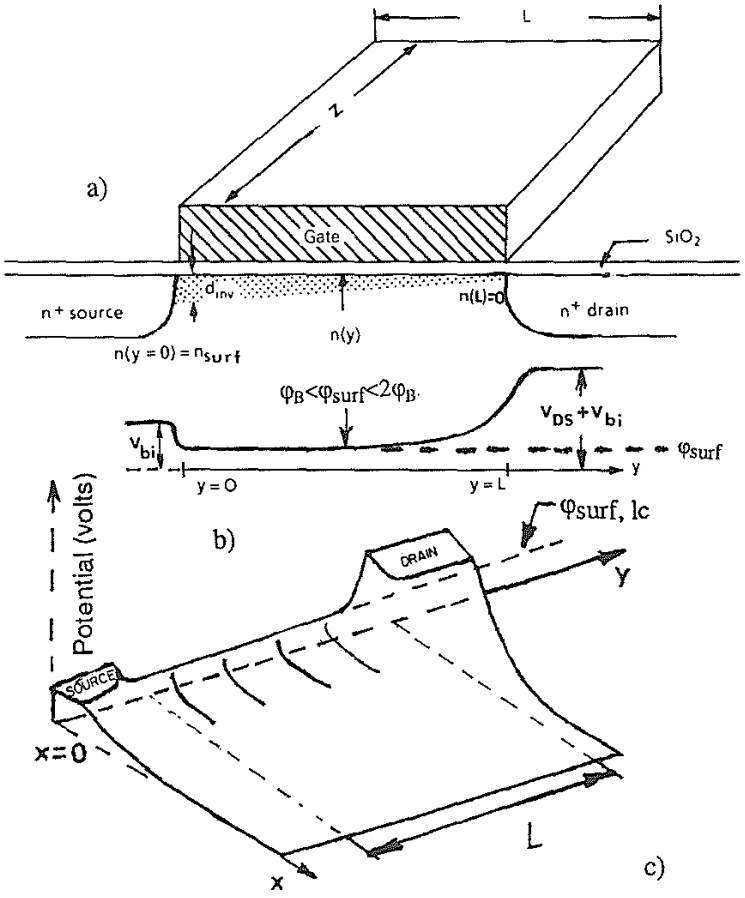


Fig. 4-26 (a) Cross-sectional area of an NMOSFET in weak inversion showing the variation of inversion layer charge concentration versus position along the channel; (b) Plot of $\varphi_{surf}(y)$ versus y for the device of part (a); (c) 2-D perspective plot of potential versus position in a long-channel NMOSFET in weak inversion.

$$t_{inv} \approx \frac{kT}{q} \frac{\sqrt{\kappa_{si}\varepsilon_o}}{\sqrt{2q\,N_A\,\varphi_{surf}}} \qquad (4 - 85)$$

In addition, the carrier concentration in the inversion layer can be found using Eq. 3-13 (with $n_{po} = n_i^2/N_A$), or $n = (n_i^2/N_A)\exp(q\varphi_{surf}/kT)$. Using these relationships, $Q_I(y)$ can be expressed as

$$Q_I = q\,t_{inv}\,n(y) \approx q\,\frac{kT}{q}\frac{\sqrt{\kappa_{si}\varepsilon_o}}{\sqrt{2q\,N_A\,\varphi_{surf}}}\frac{n_i^2}{N_A}\exp\left(\frac{q\,\varphi_{surf}}{kT}\right) \qquad (4 - 86a)$$

By some algebraic manipulation this expression for Q_I can be shown to be identical to the equation for the mobile charge density per unit area Q_I in a MOS-C under weak inversion given in chap. 3 (i.e., Eq. 3-58, which we repeat here)

$$Q_I \text{ (weak inv)} \approx \frac{\sqrt{2q\, N_A \kappa_{Si} \varepsilon_o}}{2\sqrt{\varphi_{surf}}} \varphi_t\, e[\varphi_{surf}(GB) - 2\varphi_B]/\varphi_t \qquad (4 - 86b)$$

where the surface potential in the MOS-C is established just by the gate-body bias V_{GB} (and $\varphi_t = kT/q$). We can include the effect of a bias between the channel and substrate in Eq. 4-85. That is, with the source grounded (i.e., $V_{BS} = 0V$, implying that $V_{DB} = V_{DS}$), Q_I at the source and drain ends of the channel in a MOSFET under weak inversion, can be written using Eq. 4-86a as follows:

$$Q_{I,source} \approx q\, \frac{kT}{q} \frac{\sqrt{\kappa_{Si}\varepsilon_o}}{\sqrt{2\, q\, N_A\, \varphi_{surf}(source)}} \frac{n_i^2}{N_A} \exp\left(\frac{q\, \varphi_{surf}(source) - 2\varphi_B}{kT}\right) \qquad (4 - 87a)$$

$$Q_{I,drain} \approx q\, \frac{kT}{q} \frac{\sqrt{\kappa_{Si}\varepsilon_o}}{\sqrt{2\, q\, N_A\, \varphi_{surf}(source)}} \frac{n_i^2}{N_A} \exp\left\{\frac{q\, [\varphi_{surf}(source) - 2\varphi_B - V_{DS}]}{kT}\right\} \qquad (4 - 87b)$$

where $\varphi_{surf}(GB) = \varphi_{surf}(source)$ in this case, as $V_{BS} = 0V$. Since the diffusion current is constant along the channel, the gradient of Q_I (i.e., dQ_I/dy) is also constant and can be approximated by the following expression, using Eqs. 4-87a and b:

$$\frac{dQ_I}{dy} = \frac{Q_{I,source} - Q_{I,drain}}{L} \approx \frac{Q_{I,source}}{L}\left[1 - \exp\left(\frac{-q\, V_{DS}}{kT}\right)\right] \qquad (4 - 88)$$

Then the expression for the diffusion component of the drain current in weak inversion I_{Dst} can be written as

$$I_{Dst} = D_n\, Z \frac{dQ_I}{dy} = \frac{D_n\, Z\, Q_{I,source}}{L}\left[1 - \exp\left(\frac{-q\, V_{DS}}{kT}\right)\right] \qquad (4 - 89)$$

Now, if $|V_{DS}| \geq 0.1V$, at room temperature the term $\exp(-qV_{DS}/kT)$ in Eq. 4-88 is smaller than e^{-4}. Hence, this term in Eq. 4-89 will be much smaller than 1. As a result, if V_{DS} exceeds a few φ_t (i.e., a few kT/q, or at 300K, $\geq 0.1V$), we can consider $Q_{I,drain}$ to be negligibly small compared to $Q_{L,source}$.

In such cases, the expression for I_{Dst} in long-channel MOSFETs can be approximated as $I_{Dst} \sim D_n Z Q_{Isource}/L$, or, in expanded form:

$$I_{Dst} \approx q\, D_n \frac{Z}{L} \frac{kT}{q} \frac{\sqrt{\kappa_{Si}\varepsilon_o}}{\sqrt{2\, q\, N_A\, \varphi_{surf}(source)}} \frac{n_i^2}{N_A} \exp\left(\frac{q\, \varphi_{surf}(source)}{kT}\right) \qquad (4 - 90)$$

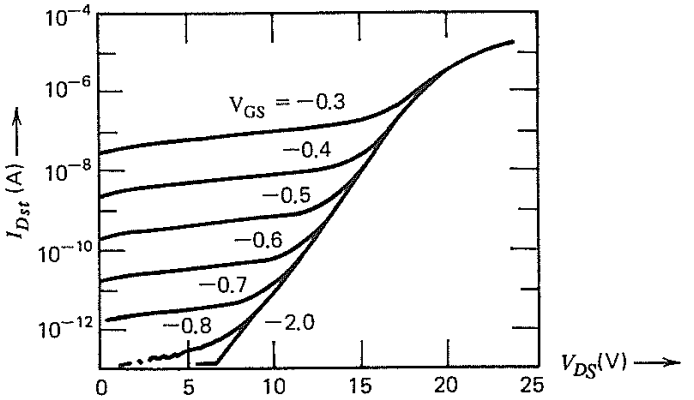


Fig. 4-27 Experimental low-current characteristics for a MOSFET with $L = 2.1$ μm, with the source and substrate both tied to ground. For lower values of V_{DS}, I_{Dst} is reduced by roughly one order of magnitude for each 0.1V reduction in V_{GS} below the threshold value. At higher values of V_{DS}, a subsurface current flows that is independent of V_{GS}.[24] (© IEEE 1974).

Since φ_{surf}(source) is roughly proportional to V_{GS} in the subthreshold region,[16] Eqs. 4-87 and 4-90 imply that $Q_{I,source}$ and I_{Dst}, respectively, are both exponentially dependent on V_{GS}.

In summary, Eq. 4-90 predicts that in long-channel MOSFETs I_{Dst} will exponentially increase with V_{GS} but will remain substantially independent of V_{DS}. Figure 4-27 illustrates this subthreshold behavior in a MOSFET with $L = 2.1$ μm. For values of V_{DS} smaller than the punchthrough voltage (i.e., roughly the V_{DS} value where the I_D curves intercept the more steeply rising portion of the curve in Fig. 4-27), I_{Dst} is seen to be essentially independent of V_{DS} and increases exponentially with V_{GS}.

It turns out, however, that devices with gate lengths of ~2 μm are at the transition between long-channel and short-channel device behavior. That is, when V_{DS} is sufficiently increased in MOSFETs with $L \leq 2$ μm, the source and drain depletion regions merge, and at that point I_{Dst} starts to become dependent on V_{DS} (as is also illustrated in Fig. 4-27 in the right side of the characteristics, where I_{Dst} no longer depends on V_{GS}). This short-channel effect, known as *punchthrough*, is discussed in chap. 5. In general, however, punchthrough must be prevented from occurring in normal device operation, even in short-channel devices. That is, at the maximum operating V_{DS} value the punchthrough current should be kept smaller than the long-channel I_{Dst} value. One simple rule of thumb to accomplish this in a MOSFET that employs a V_T-adjust implant is to set the substrate doping concentration N_{sub} at $>0.1N_I$, where $N_I = D_I/x_d$, and D_I and x_d are the threshold-adjust implant dose and the effective implant depth (or centroid), respectively.

It should also be noted that the V_T-adjust implant generally increases the subthreshold swing S_t, and this is one reason that S_t in real devices is larger (~100 mV/dec) than the theoretically predicted value of ~60 mV/dec at 300K. A more complete discussion of S_t is given in the upcoming section.

4.7.1 Subthreshold Swing S_t

Another important device characteristic in the subthreshold regime is the slope of log I_D versus V_{GS}. This characteristic indicates how effectively a MOSFET can be turned OFF as V_{GS} is decreased below V_T. The inverse of the slope is defined as the subthreshold swing S_t.[17] As voltage levels are reduced in an effort to save power and to use smaller devices, this characteristic becomes a limitation on how small a power-supply voltage can be used. In MOSFETs with uniformly doped substrates, S_t is calculated from

$$S_t = \ln 10(d\ln I_D/dV_{GS})^{-1} = 2.3\ (kT/q)\ (1 + C_d/C_{ox}) \qquad (4\text{-}91a)$$

$$= 2.3\ (kT/q)\ [1 + \kappa_{si}t_{ox}/\kappa_{ox}d)] \qquad (4\text{-}91b)$$

S_t is expressed in units of mV per decade. (The factor 2.3 comes from the conversion of ln(x) to 2.3 $\log_{10}[x]$). Let us consider the S_t parameter in more detail.

Ideally, an abrupt change in I_D should occur as V_{GS} passes through V_T. If a MOSFET exhibits a steep decline in I_D as V_{GS} is decreased below V_T, its S_t value will be small. Conversely, if a MOSFET structure is known to possess a small S_t, the implication is that only a small reduction of V_{GS} below V_T will effectively turn off the device; whereas if the device exhibits a large S_t value (meaning a gradual slope in log I_D vs V_{GS}) a significantly large I_{Dst} may still flow in the OFF state (e.g., when $V_{GS} = 0$).

According to Eq. 4-91b, at 300K in the ideal limit of t_{ox} --> 0, $S_t = (2.3kT/q)$ ~60 mV/dec. However, any nonidealities in an actual device (e.g., the presence of interface states, or a nonzero value of t_{ox}) cause S_t to become larger than the 60 mV/dec value (also, note that if T increases, so does the value of S_t). The S_t value of typical submicron NMOSFETs operating at 300K is ~100 mV/decade, and in such devices I_D will drop from 1 μA to 1 pA for a 0.6V decrease in V_{GS}. Thus, if a MOSFET exhibits a *constant-current-V_T* of 0.6V (i.e., $I_D = 1\ \mu A/\mu m$ when $V_{GS} = V_T = 0.6V$), when $V_{GS} =$ 0V, I_D will be reduced to 1 pA/μm. Thus, if S_t is known, Eq. 4-91a can be used to estimate I_{Dst} in long-channel MOSFETs in subthreshold operation (Fig. 4-28).[18] That is, circuit designers can readily calculate the gate bias required to keep the subthreshold leakage current below its specified maximum. Typically, to ensure that I_{Dst} will be negligibly small, the maximum bias applied to the gate when the device is in the OFF state should be kept at least 0.5 V below V_T.

Equation 4-91 also indicates that S_t can be made smaller by using a thinner t_{ox} or a lower substrate doping concentration (which will make the channel depletion-region width d larger). In addition, Eq. 4-91 implies that S_t increases with temperature (for example, at 100°C, the minimum theoretical value of S_t is increased to ~100 mV/dec).*

* Note that in a non-uniformly doped channel (e.g., one which uses a V_T-adjust implant), the expression for S_t is modified to

$$S_t = (kT/q) \ln 10\ [\ 1 + C_d/C_{ox}]/\{1 - [(\kappa_{si}/\kappa_{ox})(t_{ox}/L_D)]^2[C_d/C_{ox}]^2\} \qquad (4\text{-}92)$$

Basically, this equation indicates that the V_T-adjust implant causes S_t to be somewhat increased (since the additional dopant in the channel reduces the depletion-region width, thus making C_d in Eq. 4-91 larger). However, if the implant is deep enough, the depletion region

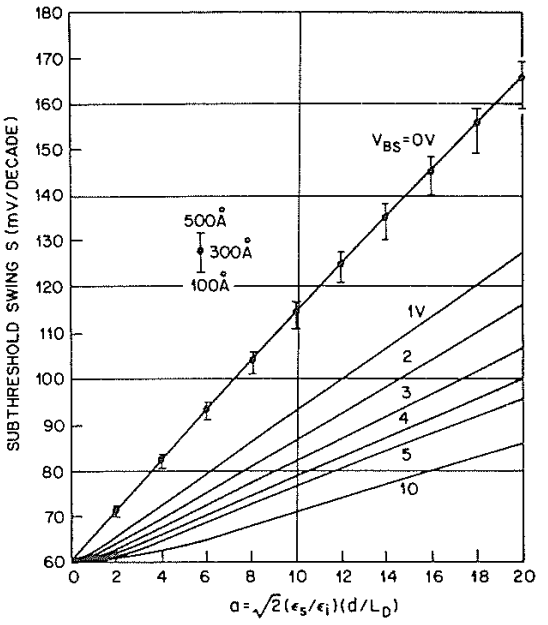


Fig. 4-28 Subthreshold gate swing S_t required to reduce the subthreshold current one decade versus the parameter $a = [\sqrt{2}\, t_{ox}/\kappa_{ox}q)][\kappa_{si}kT/\varepsilon_0 N_A)^{1/2}$. Vertical bars indicate the variation in S_t at constant a when the oxide thickness is varied from 500 to 100 Å, and doping level and mobility adjusted accordingly. $V_{DS} >> kT/q$. [13] (© IEEE 1979).

Finally, it should be noted that because the depletion width increases when a substrate bias is applied, the subthreshold swing decreases according to Eq. 4-91b. This effect is shown in Fig. 4-29.[17,24]

Note that when the channel length gets small, the values of I_{Dst} are larger than those predicted by the above model. This is due to so-called short-channel effects (discussed in chap. 5). As a result, measurements of S_t versus channel length are used to detect the onset of these short-channel effects (punchthrough and surface drain-induced barrier lowering). Since the measurable I_{Dst} is the sum of both the normal subthreshold and short-channel-induced components, an increase in the value of S_t will signal the onset of the latter.

* (cont.) width will eventually be equal to that of the unimplanted device, and S_t returns to the value exhibited by a device with no V_T-adjust implant. Nevertheless, the general dependences of S_t on t_{ox}, T, and channel substrate doping exhibited in devices with uniformly doped channels still apply.

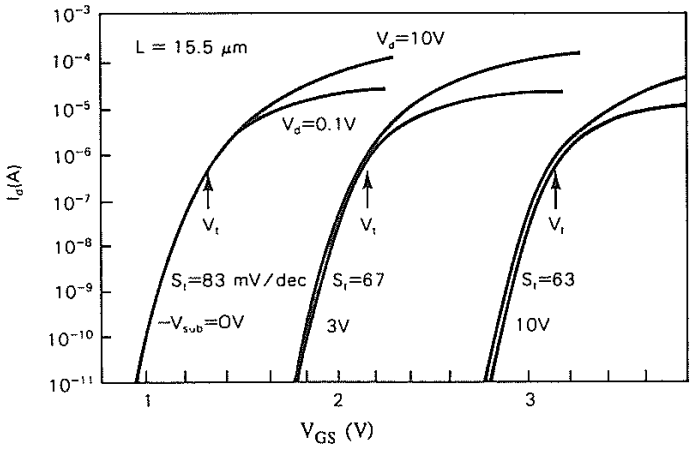


Fig. 4-29 Experimental subthreshold characteristics for a long-channel device (L = 15.5 μm).[24] (© IEEE 1974).

4.7.2 Gate-Induced Drain Leakage (GIDL)

Another form of leakage current observed in OFF-state MOSFETs is *gate-induced drain leakage (GIDL).*[19] The carriers responsible for GIDL originate in the region of the drain that is overlapped by the gate, and GIDL is occurs when the gate is grounded and the drain is at V_{DD}. A large electric field then exists across the oxide ($\mathcal{E}_{ox}$), which must be supported by charge in the drain region. This charge is provided by the formation of a depletion region in the drain as shown in Fig. 4-30a. If $\mathcal{E}_{ox}$ becomes sufficiently large, in addition to the depletion region, an inversion layer will attempt to form at the silicon surface of the drain. However, as the minority carriers arrive at the surface to form the inversion layer, they are immediately swept laterally to the substrate (which is a region of lower potential for the minority carriers). Hence, in this case, the depletion region under the oxide in the drain instead becomes a *deep depletion layer.** The zone near the surface where an inversion layer should be formed is referred to as an "incipient inversion layer." The current that flows as a result of the carriers being swept from this

* *Deep depletion* is a non-equilibrium condition where there are fewer minority carriers in a depletion region than would exist in equilibrium. Consider the case of an n^+MOS structure (i.e., the condition that exists in an NMOS drain region), that is biased so that an inversion layer of holes should exist at the surface (and a depletion layer exists below it). Under equilibrium conditions, this condition can persist indefinitely under appropriate dc bias. In the structure described above, a dc bias is applied that attempts to establish such an inversion layer of holes at the surface of the n^+drain. However, the potential of the substrate is lower for holes than the surface potential of the drain, and thus any holes arriving at the drain surface move laterally to the substrate region. Thus, the depletion region under the overlap area becomes a zone of deep depletion.

incipient inversion layer constitute the GIDL current.

The magnitude of GIDL current is independent of temperature, and hence it is believed that thermal generation of carriers in the incipient inversion region plays only a minor role in GIDL. Instead, it is proposed that when $\mathcal{E}_{ox}$ becomes large enough, the voltage drop in the deep-depletion layer becomes so large that various trap-assisted carrier-generation events become possible (Fig. 4-30b).[20] These include thermal emission of an electron from the valence band to a trap site (expressed as a vertical transition in the diagram) and then tunneling to the conduction band (horizontal transition), and trap-assisted tunneling from valence band to conduction band. For even larger fields (a few MV/cm), trap-free band-to-band tunneling may also occur (see also Fig..4-30b).

Like other forms of leakage current, GIDL will contribute to standby power, giving rise to the previously mentioned problems of excessive heat dissipation at large device counts and heavy current drain in portable systems. Hence, GIDL must be controlled so that it does not exceed some specified maximum value (typically, ≤10 pA/μm).

In order to control GIDL, the most important mechanisms that may cause it need to be identified. This allows processes involving these mechanisms to be avoided or modified to minimize GIDL. For example, it has been observed that shallow junction

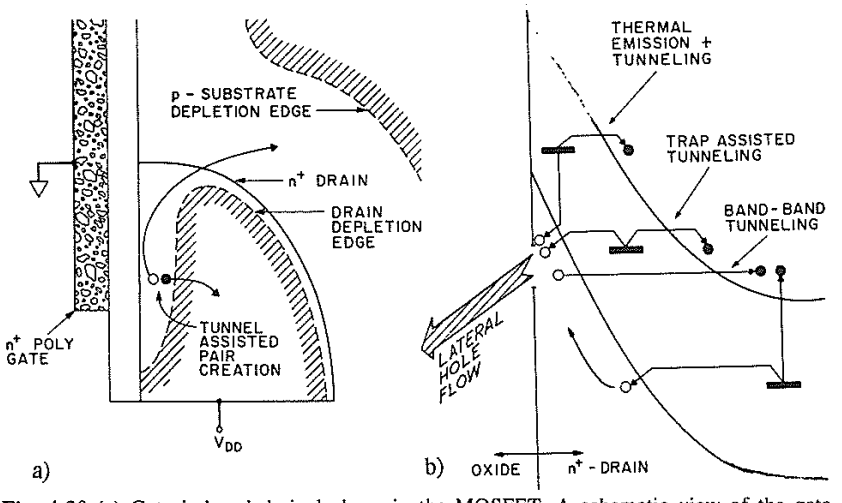


Fig. 4-30 (a) Gate induced drain leakage in the MOSFET. A schematic view of the gate-drain overlap region for a grounded gate and the drain biased at V_{DD}. Tunneling created pairs lead to a lateral hole flow in the n^+ drain. This flow prevents formation of an inversion layer inside the drain. (b) Various mechanisms that can contribute to gate-induced drain leakage. Vertical transitions are thermal-emission processes. Horizontal transitions are tunneling processes. Holes are prevented from forming an inversion layer by lateral transport to the substrate. This flow is normal to the plane of the diagram, as indicated by the slanted arrow, lateral hole flow.[20] From J.R. Brews, "The Submicron MOSFET," Chap. 3, in *High-Speed Semiconductor Devices*, Ed. Sze. Copyright 1990, John Wiley & Sons. Reprinted with permission.

fabrication using Ge preamorphization introduces bulk midgap traps that increase GIDL.[21] Other reports indicate that interface traps caused by hot carrier injection,[22] or Fowler-Nordheim tunneling[23] increase GIDL by stimulating additional tunneling.

There are several general procedures that appear to be effective in reducing GIDL. First, the oxide thickness can be increased to reduce $\mathcal{E}_{ox}$ for a given voltage. But this is usually not implemented, since a thicker gate oxide will adversely impact other device characteristics. Second, the trap density in the near-surface area can be reduced, but such trap elimination generally requires a carefully controlled fabrication sequence. Third, the doping in the drain can be increased, as this will decrease the depletion region width and the tunneling volume. Unfortunately, the latter approach favors abrupt junctions, because graded or lightly-doped-drain (LDD) structures will lead to a significant lateral extension of the drain within which the doping is lighter, inviting greater GIDL. Tradeoffs of field reduction against GIDL may need to be considered. On the other hand, it has also been reported that LDD devices can suppress GIDL by suppressing the lateral field.

Nevertheless, it has been reported that GIDL sets a limit on the power supply voltage V_{DD} of an LDD MOSFET of[25,29]

$$V_{DD} = 4 \times 10^6 \text{ V/cm} \times t_{ox} + 1.2\text{V} - V_{FB} \qquad (4 - 93)$$

A non-LDD MOSFET would have a much higher GIDL current. This implies that suppression of GIDL may be a strong reason to retain the LDD structure in MOSFETs even after V_{DD} is reduced to suppress hot-carrier effects. (Note that in Eq. 4-93 V_{FB} is ~0V for an n^+ polysilicon gate on an n^+ drain and about 1.1V for an n^+ polysilicon gate over a p^+ drain.)

4.8 SUMMARY OF LONG-CHANNEL MOSFET BEHAVIOR

The behavior of the long-channel as detailed in this chapter can be summarized by examining the I-V characteristics of the long-channel MOSFET in strong inversion (Fig. 4-31a), and in the subthreshold region of operation (weak inversion, see Fig. 4-31b).

In strong inversion we plot the output characteristics of the MOSFET (I_D versus V_{DS} with V_{GS} as the parameter) on an axis with linear I_D and V_{DS} scales. We see that for a fixed value of $V_{GS}>V_T$, the drain current first increases linearly with V_{DS}. Eventually, as shown in Fig. 4-31a, the drain current saturates. Beyond the value of saturation voltage (i.e., when $V_{DS}>V_{DSsat}$), I_D remains constant with increasing V_{DS}. The value of I_{Dsat} is also proportional to $(V_{GS} - V_T)^2$ - as predicted by charge-control model. The values of I_D versus V_{DS} and V_{GS} in the linear region, as well as the values of I_{Dsat} and V_{DSsat} can be calculated using any one of the four long-channel MOSFET models given in section 4.5, depending on the accuracy desired.

In the subthreshold region of operation, the behavior is described by plotting I_D (with a logarithmic scale) versus V_{GS} (with a linear scale). As shown in Fig. 4-31b, in long-channel MOSFETs the drain current I_D drops exponentially once V_{GS} becomes ~0.2V below V_T. Furthermore, in long-channel MOSFETs the slope of this log I_D-V_{GS}

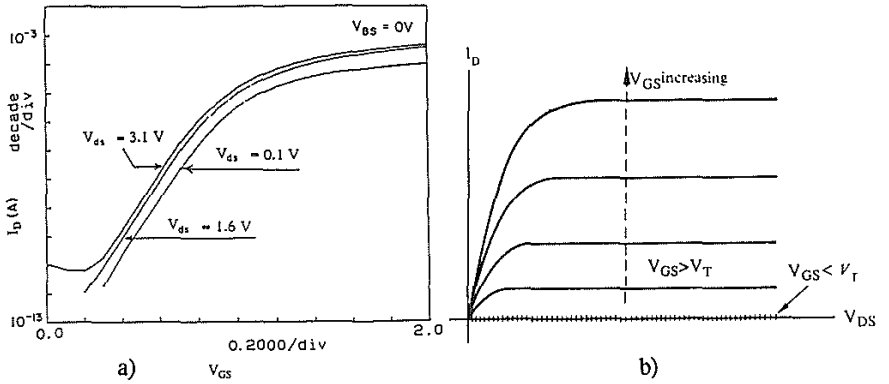


Fig. 4-31 (a) I_D versus V_{DS} characteristics of long-channel MOSFETs beyond strong inversion. (b) I_D versus V_{GS} characteristics of a long-channel MOSFET below threshold (weak inversion).

characteristic does not change with increasing V_{DS}, nor does the value of I_D increase at a fixed value of V_{GS} as V_{DS} is increased.

Finally, the threshold voltage of the long-channel MOSFET does not vary as L is decreased, nor does it change with increasing V_{DS}, although this V_T behavior is not indicated by either Fig. 4-31a or Fig. 4-31b.

REFERENCES

1. D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits,* McGraw-Hill Book Co., New York, 1983.
2. S. M. Sze, *Semiconductor Devices - Physics and Technology,* John Wiley & Sons, New York, 1985, p. 444.
3. S. Selberherr, A. Schutz, and H.W. Potzl, "MINIMOS-A two dimensional MOS analyzer," *IEEE Trans. Electron Dev.,* **ED-27**, p. 1540, 1980.
4. H.C. Pao and C.T. Sah, "Effects of Diffusion Current on Characteristics of Metal-Oxide (Insulator)-Semiconductor Transistors," *Solid-State Electron.* **9**, 927, (1966).
5. Y.P. Tsividis, *Operation and Modeling of the MOS Transistor,"* McGraw-Hill, New York, 1987.
6. R.F. Pierret and J.A. Shields, "Simplified long-channel MOSFET Theory," *Solid-State Electronics,* vol. 26, p. 143, 1983.
7. G. Baccarani et al., "Analytical i.g.f.e.t.model including drift and diffusion currents," *IEE Journal on Solid-State and Electron Devices,* vol. 2, p. 62, March 1978.
8. J.R. Brews, "A Charge-Sheet Model of the MOSFET," *Solid-State Electronics,* vol. 21, p. 345, 1978.
9. H.K. Ihantola and J.L Moll, "Design Theory of a Surface Field-Effect Transistor," *Solid-State Electronics* 7, 423 (1964).

10. A. Vladimirescu, A.R. Newton, and D.O. Pederson, *SPICE Version 2G.2 Users Guide,* Electronics Research Laboratory, University of California, Berkeley.

11. R.S. Muller and T.I. Kamins, *Device Electronics for Integrated Circuits,* 2nd. Ed., New York, John Wiley & Sons, 1986.

12. V.L. Rideout, F.H. Gaensslen, and A. LeBlanc, "Device Design Consideration for Ion-implanted *n*-Channel MOSFETs," *IBM J. Res. Dev.,* p. 50, (1975).

13. J.R. Brews *IEEE Trans. Electron Dev.*, **ED-26(11)**, 1696 (1979),

14. D. Antoniadis "Calculation of threshold voltage in nonuniformly doped MOSFETs," *IEEE Trans. Electron Dev.*, **ED-31,** p. 303, March 1984.

15. J.R. Brews, "The Physics of the MOS Transistor," in *Silicon Integrated Circuits*, Part A, Ed. D. Kahng, New York, Academic Press, 1981.

16. S. M. Sze, *Semiconductor Devices: Physics and Technology,*, John Wiley & Sons, New York, p. 213, 1985.

17. J.R. Brews, J. "Subthreshold Behavior of Uniformly and Nonuniformly Doped Long-Channel MOSFET," *IEEE Trans. Electron Dev.,* **ED-26** (9) p. 1282, Sept. 1979.

18. *ibid.* , (**16**), p. 470.

19. T. Y. Chan et al., 1987 IEDM Tech. Digest, p. 718.

20. J.R. Brews, "The Submicron MOSFET," chap. 3 in *High-Speed Semiconductor Devices,* S.M. Sze Ed., Wiley- Interscience, New York, 1990.

21. D.-S. Wen et al., "Tunneling Leakage in Ge-PreAmorphized Shallow Junctions," *IEEE Trans. Electron Dev.,* **ED-35** (7), 1107 (1988).

22. C. Duvvury et al., "Leakage Current Degradation in *n*-MOSFETs Due to Hot-Electron Stress," *IEEE Electron Dev. Lett.* **EDL-9** (11), 579 (1988).

23. I. C. Chen et al., "Interface-Trap Enhanced Gate-Induced Leakage Current in MOSFET," *IEEE* Electron Dev. Lett., **EDL-10**, (5), 216 (1989).

24. R.R. Troutman, "Subthreshold Design Considerations for IGFETs," *IEEE J. Solid-State Circuits,* **SC-9**, p. 55, (1974).

25. S. Parke et al., "Design for suppression of gate-induced leakage in LDD MOSFETs."

26. *ibid.,* reference **16** p. 444.

27. V.L. Rideout, F.H. Gaensslen, and A. LeBlanc, *IBM J. Res. Dev.,* **19**, p. 50, (1975).

28. N.D. Arora, *Solid-State Electronics*, **30**, p. 559, (1987).

29. H.-J Wann, P.K. Ko, and C. Hu, "Gate-Induced Band-to-Band Tunneling Leakage Current in LDD MOSFETs," *Tech. Dig. IEDM,* 1992, p. 147.

30. R.F. Pierret, *Field Effect Devices*, Vol. IV Modular Series on Solid-State Devices, Addison-Wesley Publishing Co. Reading, MA., 1983.

PROBLEMS

1. Another way to model the drain current in a MOSFET is to note that it can be represented as mobile charge induced in the channel divided by the *transit time* (or average time required for an electron to move from the source to drain). For an *n*-channel MOSFET with a 250Å oxide, 1 μm channel length, 10 μm channel width, and $V_T = 1V$, operated at $V_{GS} = 5V$ and $V_{DS} = 0.1V$, calculate the total electron charges in the channel and the electron transit time. Then, show that this charge divided by the transit time is equal to the current calculated at the linear region as given by Eq. 4-65.

2. Evaluate the coefficient *k* of Eq. 4-64 for a MOSFET in which $\mu_n = 580$ cm^2/V sec, Z/L = 1 and (a) $t_{ox} = 50$ nm, and (b) $t_{ox} = 10$ nm.

3. Account qualitatively for the differing drain current I_D predictions of the bulk-charge and square-law models of the long-channel MOSFET.

4. Describe how the saturation voltage V_{DSsat} is defined and derive $V_{DSsat} \sim V_{GS} - V_T$ as a first order expression.

5. The sheet resistances of the source (R_S) and the drain (R_D) are both parasitic resistances (see Appendix B). Discuss why R_S decreases I_D more than does R_D in the saturation region of operation. Also, derive an analytic expression of g_m in the linear region as a function of R_S and R_D.

6. If a circuit designer wants to ensure that a MOSFET exhibits no more than no more than 1 pA of leakage current in the off state, and also wants $V_T = 0.9V$ (where V_T is defined as the gate voltage that causes 1 μA current), what maximum S_t value must the device exhibit.

7. Verify that the expression for the capacitance of the surface space-charge region C_d in Eq. 4-91 is correct (i.e., that $C_d = \kappa_{Si}\varepsilon_0/d$).

8. The I_D-V_{GS} characteristics for a long *n*-channel MOSFET are shown in Fig. 4P-1. Using the square-law model and information available from the graph, compute the drain current I_D when $V_{GS} = 3V$ and $V_{DS} = 5V$.

9. A MOSFET has been fabricated on a *p*-type substrate with doping concentration $N_A = 4\times10^{15}$ cm^{-3}. The gate oxide thickness is 50 nm, the gate is n^+ polysilicon, the oxide charge is $Q_{tot} = 2\times10^{10}$ cm^{-2}, and V_T-adjust implant of 2×10^{12} boron atoms/cm^2 is performed through the gate oxide. Assume that Z/L = 1 and $\mu_n = 800$ cm^2/V sec, and $V_{BS} = 0V$. (a) Calculate the threshold voltage of the device; and (b) Calculate V_{DSsat} and I_{Dsat} for the MOSFET using: (i) the bulk-charge model and; (ii) the square-law model. Take $V_{GS} = 5V$. Comment on the accuracy of the square-law model.

10. Explain why we can approximate the thickness of the inversion layer as the depth beneath x = 0 at which $\varphi(x)$ has dropped to a value of $0.38\varphi_{surf}$.

11. Explain why I_D does not become zero if $V_{GS}>V_T$ and $V_{DS}>V_{DSsat}$, as might be the conclusion based on the long-channel approximation used to calculated I_D.

12. Derive the I-V characteristics of a MOSFET with the drain and gate connected together and the source and substrate grounded. Can one obtain the threshold voltage from these characteristics?

13. An NMOSFET with $Z/L = 15$, $t_{ox} = 80$ nm and $\mu_n = 600$ cm^2/V sec is to be used as a controlled resistor: (a) What charge density is required for the device to exhibit a dc resistance of 3.5 kΩ between the source and drain in the saturation region? (b) What value of $(V_{GS} - V_T)$ is needed to obtain the desired resistance?

14. An Al gate enhancement-type NMOSFET has a substrate doping of $N_A = 6\text{x}10^{15}$ cm^{-3} and a gate oxide thickness of 100 nm, and $Q_{tot} = 5\text{x}10^{10}$ cm^{-2}. Determine V_T of this device. When the device is operated in the linear region with $V_{DS} = 0.5$V, calculate the gate voltage needed to obtain a drain current of 2 mA. Assume $L = 10$ μm, $Z = 100L$, and $\mu_n = 500$ cm^2/V sec.

CHAPTER 5

MOS TRANSISTOR DEVICE PHYSICS: PART 3

THE SUBMICRON MOSFET

The two major goals of MOSFET scaling are to increase the *density* and *speed* of the digital ICs in which such scaled-down devices are used. Increasing *density* obviously means using smaller channel lengths and widths. To increase the *speed* of digital ICs, the MOSFET saturation drain current I_{Dsat} must be increased (i.e., to allow faster charging and discharging of parasitic capacitances). From the long-channel MOSFET models given in chap. 4, a decrease in either the channel length L or gate oxide thickness t_{ox} will lead to an increase in I_{Dsat}. In fact, the long-channel MOSFET models predict that I_{Dsat} should continue to increase indefinitely as L and t_{ox} are decreased, seeming to imply that only the limitations of process technology (and not device effects) prevent the manufacture of even smaller, higher-performing MOSFETs. In addition, the MOSFET models of chap. 4 also accurately describe many other device characteristics of long-channel MOSFETs, including threshold voltage, subthreshold current, and the I_D-V_{DS} characteristics in strong inversion.

However, as process technology improved to the point where devices could be fabricated with gate lengths smaller than ~2 μm, it turned out that MOSFETs began to exhibit phenomena not predicted by the long-channel MOSFET models. Such phenomena were thus termed *short-channel effects.* One of the more surprising of such short-channel effects (which becomes especially pronounced as L decreases below 0.5 μm) is that the drain current in saturation I_{Dsat} shows far less increase as L is decreased than is predicted by the long-channel models (see Fig. 5-1).[1,47] In fact, in the simple model to be presented in section 5.4, I_{Dsat} is predicted to become independent of L in extremely small MOSFETs, approaching instead, a constant value given by $Q_I v_{sat} Z$ (where v_{sat} is the saturation velocity of the carriers, as described in section 5.4.1.2). In contrast, reducing t_{ox} (which has the effect of increasing the oxide field, and in turn causes Q_I to increase), yields a considerably greater increase in I_{Dsat}. Furthermore, as shown in Fig. 5-1 I_{Dsat} will also increase more rapidly as L is decreased if the MOSFET has a thinner gate oxide. This benefit provides even a greater impetus for making t_{ox} as thin as possible as gate lengths are decreased further in the drive for higher density.

Since the quest for higher density still requires L (and Z) to be further reduced, it will nevertheless be necessary to confront the other short-channel effects (besides the

Dell Ex. 1008C
Page 144

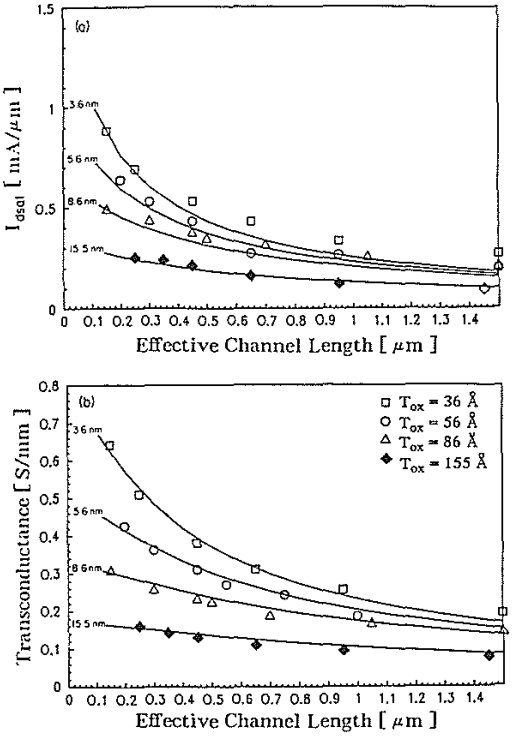


Fig. 5-1 Measured and calculated I_{Dsat} for an array of deep submicron NMOS transistors with four different gate-oxide thicknesses.[47] (© 1987 IEEE).

one of the I_{Dsat} dependence on L mentioned above). These other short-channel effects can be classified into two groups: (1) increased leakage current when the MOSFET is "off," and; (2) reliability problems associated with short-channel device structures.

Increased "off-state" leakage in short-channel MOSFETs is due to several phenomena, including: (a) lowering of the threshold voltage V_T as L is decreased and/or V_{DS} is increased; (b) the onset of punchthrough at smaller drain biases as L is decreased, and; (c) an increase in isolation leakage current as the isolation spacing is decreased. We will discuss V_T lowering and punchthrough in this chapter, and isolation leakage phenomena in chap. 6.

The reliability problems that arise in short-channel MOSFETs include: (a) thin gate-oxide breakdown (to be discussed in chap. 7); (b) device degradation due to hot-carrier effects (to be discussed in chap. 9); and (c) reliability problems associated with the interconnects between MOSFETs (e.g., electromigration failures in the metal lines). The latter problem is treated in Vol. 2, chaps. 3 and 4.

Table 5-1. Comparison of Long-Channel and Short-Channel MOSFET Characteristics

Long-Channel MOSFET Behavior	Short-Channel MOSFET Behavior
1. The threshold voltage V_T is independent of channel length L and width Z.	1. V_T decreases as L is decreased. V_T may also be impacted by changes in Z.
2. V_T is independent of drain bias voltage.	2. V_T decreases with increasing V_{DS}.
3. V_T depends on V_{BS} according to Eq. 4-12 (with V_{BC} in Eq. 4-12 replaced by V_{BS}).	3. V_T increases less rapidly with V_{BS} than predicted by Eq. 4-12.
4. The subthreshold current I_{Dst} increases linearly as L decreases.	4. I_{Dst} increases more rapidly than linearly as L decreases.
5. I_{Dst} is independent of drain bias.	5. I_{Dst} increases with increasing V_{DS}.
6. The subthreshold swing S_t is independent of gate length.	6. S_t increases with decreasing L.
7. The drain current in saturation I_{Dsat} is independent of V_{DS}.	7. I_{Dsat} increases as V_{DS} increases.
8. I_{Dsat} is proportional to $(V_{GS} - V_T)^2$	8. I_{Dsat} is proportional to $(V_{GS} - V_T)$.
9. I_{Dsat} is proportional to $1/L$.	9. As $L \rightarrow 0$, I_{Dsat} becomes independent of L.

In this chapter we will also cover several device design scenarios developed to keep leakage currents sufficiently small as the MOSFET is scaled down. In addition, we describe the issues related to submicron PMOSFET design.

5.1 DEVICE CHARACTERISTICS OF LONG-CHANNEL VERSUS SHORT-CHANNEL MOSFETS

From the perspective of MOSFETs as circuit components, long-channel behavior has been specified in a number of ways. As a result, short-channel device effects can be so designated when a deviation from long-channel MOSFET behavior can be correlated with a reduction in the gate length (and/or gate width) dimension. Table 5-1 gives a list of long-channel device characteristics that undergo variation as the gate dimensions are decreased. As the behavior of the short-channel MOSFET is described later in this chapter, the terms associated with these characteristics will be defined and discussed in more detail.

From Table 5-1 we can see that the short-channel effects can be divided into the following categories: (a) those that impact V_T (Items 1-3): (b) those that impact subthreshold currents (Items 4-6); and (c) those that impact I_D when the MOSFET is operated in saturation [i.e., $V_{DS}>(V_{GS} - V_T)$], namely Items 7-9. In this chapter we

discuss these effects in detail, as well as measures that can be adopted to minimize their impact on MOSFET behavior.

5.2 EFFECT OF GATE DIMENSIONS ON THRESHOLD VOLTAGE

In this section we will describe three short-channel effects on the threshold voltage of MOSFETs, namely: (1) the short-channel threshold-voltage shift; (2) Narrow gate width effects on threshold voltage; and (3) the reverse short-channel threshold voltage shift.

5.2.1 Short-Channel Threshold-Voltage Shift

Experimentally, it is observed that as the dimensions of the gate are reduced, the threshold voltage of MOSFETs become less well predicted by the long-channel V_T equation (Eq. 4-74). The error becomes significant as the *channel length dimension L* is reduced to less than 2 μm. Usually, V_T decreases as L is reduced.* In addition, in short-channel MOSFETs V_T also decreases as V_{DS} is increased. To get good agreement with measured data, a term ΔV_{TSC} must be subtracted from the value of V_T obtained from Eq. 4-74. (In *n*-channel enhancement-mode MOSFETs the measured value of V_T becomes less positive than that predicted by Eq. 4-74, while in *p*-channel enhancement-mode MOSFETs, V_T becomes less negative.)

The decrease of V_T with L and V_{DS} in short-channel devices is crucial because enhancement-mode FETs in CMOS are generally designed to operate with V_T magnitudes of 0.6-0.8V. If the magnitude of V_T drops even slightly below its designed value, the device may exhibit excessive drain leakage current when V_{GS} = 0V. As discussed earlier, V_T values of 0.6-0.8V in long-channel NMOS devices with lightly doped substrates can only achieved by increasing the doping concentration at the surface of the channel (i.e., with a V_T-adjust implant). If short-channel effects reduce V_T below the long-channel values, the channel doping concentration must be increased even more to re-establish the desired V_T. Higher channel doping, however, generally degrades other device characteristics, such as carrier mobility and off-current. Thus, it is important to be able to model ΔV_{TSC} to facilitate reliable tradeoffs between process and electrical parameters and leakage current levels.

The failure of Eq. 4-74 to give good agreement with measured V_T values in short-channel MOSFETs arises because this equation was derived assuming one-dimensional theory. That is, Eq. 4-74 assumes the space charge under the gate is controlled only by the vertical $\mathcal{E}$-field, $\mathcal{E}_X$. In that case, the channel depletion region charge will be influenced only by the charge on the gate. When the channel of a MOSFET is long, this is a reasonable assumption, as the influence of the drain and source junctions (and the influence of V_{DS}) on the quantity of channel charge can be neglected (Figs. 5-2a and b).

* In section 5.2.3, however, we will see that V_T may initially increase as L shrinks, before ultimately decreasing as described here. This unexpected effect is referred to as a *reverse short-channel effect* (RSCE), and will be described in section 5.2.3.

However, as L approaches the dimensions of the source/drain junction depletion-region widths, a greater part of the channel-depletion region begins to consist of the space charge in the junction depletion regions (Fig. 5-2c and d). Hence, less gate charge (and a smaller vertical $\mathcal{E}$-field) is needed to cause inversion in short-channel MOSFETs than in long-channel MOSFETs (assuming each has the same substrate doping). This makes it appear that a smaller V_T is needed to turn on a short-channel device. Furthermore, as the reverse bias applied to the drain (V_{DS}) is increased, the drain depletion region grows wider, leaving even less charge under the control of the gate.

There are four approaches that have been taken to establish quantitative values of ΔV_{TSC}. They are the following:

1. *Numerical analysis* is used to find solutions of the two-dimensional Poisson equation for calculating the electric field and potential in the channel region.
2. Analytical expressions for ΔV_{TSC} are derived using the *charge sharing model.*
3. Analytical expressions for ΔV_{TSC} are derived using the drain-induced barrier lowering model, in which the 2-D Poisson equation is simplified to a one-dimensional form, making it solvable with analytical functions.
4. Analytical expressions for ΔV_{TSC} are derived using the drain-induced barrier lowering model in which a *quasi 2-D* approach is employed to analytically solve the Poisson equation.

While the numerical solution approach is the most accurate, this technique is not feasible for circuit simulation and statistical modeling. As a result, substantial effort has been expended to develop analytical models which provide rapid, inexpensive, and sufficiently accurate simulation of ΔV_{TSC}. Hence, we devote most of our discussion to the latter three approaches.

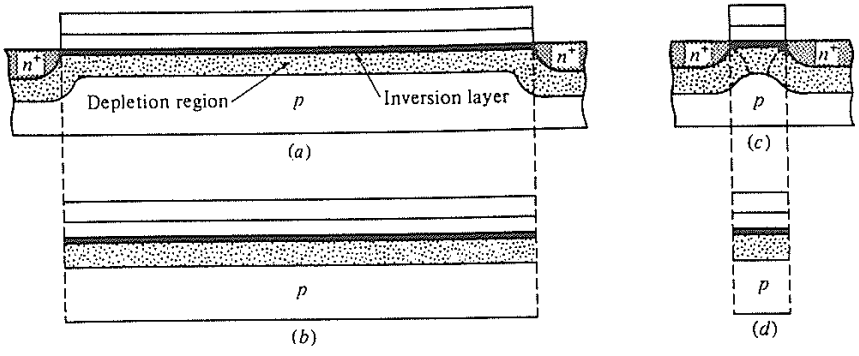


Fig. 5-2 (a) Long-channel transistor; (b) the channel of (a) with edge effects neglected; (c) a short-channel transistor; (d) the channel of (c) with edge effects neglected. From Y.P Tsividis, *Operation and Modeling of the MOS Transistor*, McGraw-Hill, New York, 1987. Reprinted with permission.

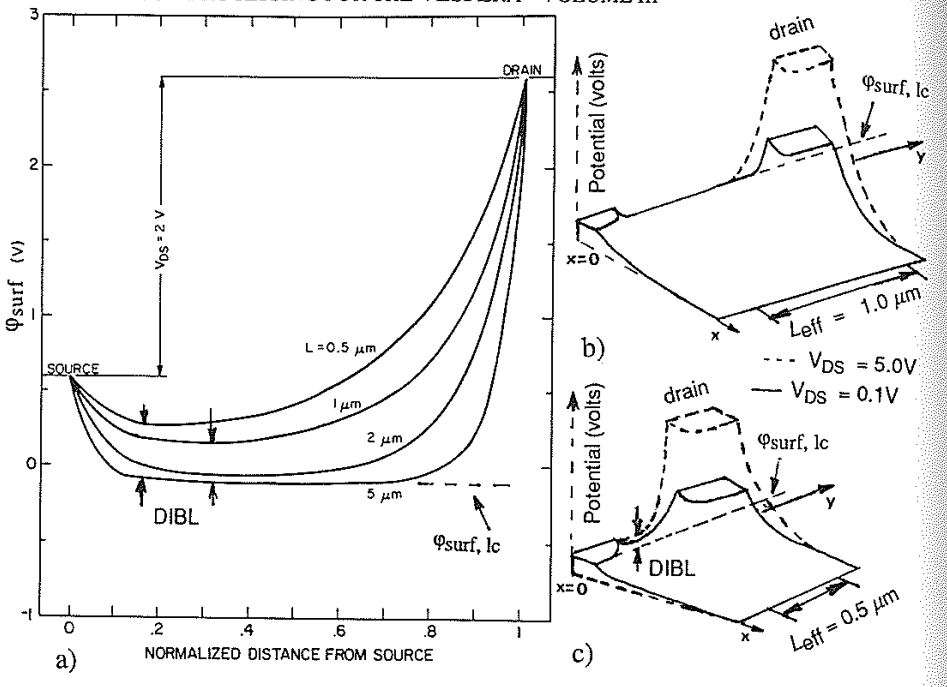


Fig. 5-3 Simulated values of the surface potential φ_{surf} along the channel for NMOSFETs with different channel lengths (note channel length is normalized). $N_A = 10^{15}$ cm^{-3}, t_{ox} = 500Å, r_j = 0.33μm, V_{GS} = 0V, V_{DS} = 2.0V, V_{BS} = 0V.[109] (© 1979 IEEE). (b & c) 2-D potential variation for NMOSFETs with (b) L_{eff} = 1.0 μm, and (c) L_{eff} = 0.5 μm (see also Fig. 5-5). Solid line, V_{DS} = 0.1V and dashed line, V_{DS} = 5.0V.

5.2.1.1 Numerical Solution of the 2-D Poisson Equation for Calculating ΔV_{TSC}. The most accurate method for simulating this effect is to solve the 2-D Poisson equation (i.e., as given here)

$$\frac{\partial^2\varphi(x,y)}{\partial y^2} + \frac{\partial^2\varphi(x,y)}{\partial x^2} = -\frac{\rho(x,y)}{\kappa_{si}\,\varepsilon_o} \qquad (5-1)$$

for the electrostatic-potential φ and the electric-field ℰ in the channel region. While exact analytical solutions of the 2-D Poisson equation are not possible, numerical methods can be used to obtain solutions of Eq. 5-1. As noted in chap. 2, several software programs have been developed which allow these methods to be implemented.[2,21] Specifically, such numerical Poisson solvers are available in device simulators (e.g., GEMINI, MINIMOS, and PISCES-II). However, numerical solutions of 2-D partial differential equations are computationally intensive (and hence time-consuming and expensive), and therefore as noted, are not suitable for use in circuit simulation applications in which the electrical behavior of large circuits must be simulated. Nevertheless, they serve as accurate benchmarks of ΔV_{TSC} against which the

values derived from the analytical models can be compared. Thus, they are a valuable tool for helping to develop more accurate analytical ΔV_{Tsc} models. Figure 5-3 shows such numerically calculated values of φ_{surf} versus y in NMOSFETs operated in subthreshold as L is decreased from 5.0 μm to 0.5 μm.

5.2.1.2 Charge-Sharing Models. The *charge-sharing* perspective assumes that not all of the electric-field lines emanating from the charge under the gate terminate on the gate charge. Instead, some terminate on space charge in the source and drain depletion regions. The depletion charge under the gate is thus fractionally established by the source and drain (and in this sense, the channel charge is considered to be "shared" among the gate, source, and drain charge, see Fig. 5-4). Hence, less gate charge is required to cause inversion, making V_T appear smaller. The fraction of the charge induced by the source and drain becomes significant as the channel length is of the order of the junction depletion region widths. Charge-sharing models seek to quantitatively determine the fractional charge induced by the gate.

In 1974, Yau proposed a simple, widely used analytical model based on the *charge-sharing* perspective.[3] This model assumes that the charge controlled by the gate is

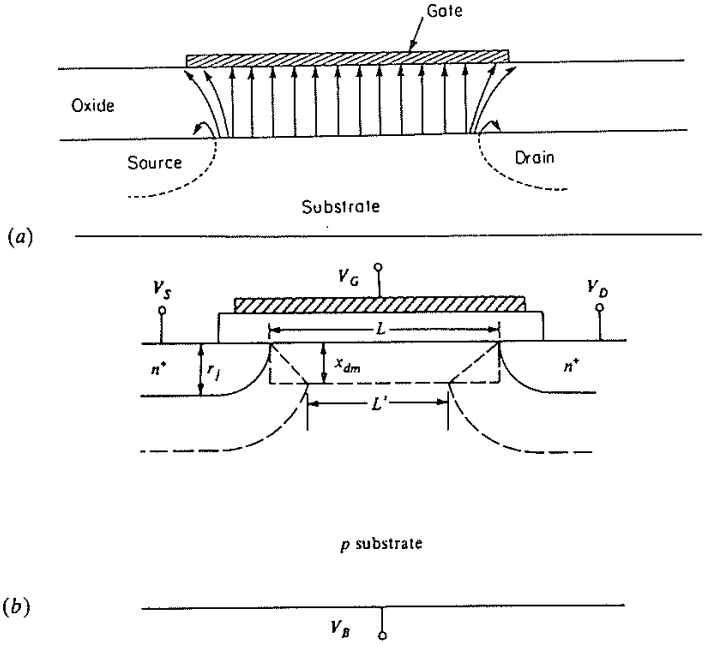


Fig. 5-4 (a) Schematic showing perturbation of the electric field lines in the oxide near the source and the drain.[5] (© 1985 IEEE). (b) Yau's model of charge sharing.[3] Reprinted with permission of Solid-State Electronics.

reduced to the charge in the trapezoidal region of the channel shown in Fig. 5-4b. The resulting analytical expression based on this assumption allows the value of ΔV_{Tsc} to be calculated in devices with uniformly doped channels:

$$\Delta V_{Tsc} = \frac{q\, N_{SUB}\, d_{max}\, r_j \left(\sqrt{\left\{ 1 + \frac{2\, d_{max}}{r_j} \right\}} - 1 \right)}{C_{ox}\, L} \qquad (5 - 2)$$

where d_{max} is the maximum width of the depletion region in the channel, and r_j is the junction depth of the source/drain regions. According to Eq. 5-2, as L decreases, ΔV_{Tsc} increases. Eq. 5-2 also implies that to keep ΔV_{Tsc} as small as possible, shallow source/drain junctions and thin gate oxides should be used ($C_{ox} = \varepsilon_{ox}/t_{ox}$). In addition, the channel depletion-region width should also be as small as possible. Although it will be seen that some of the assumptions used in Yau's model are generally invalid, the qualitative guidelines it yields for minimizing ΔV_{Tsc} appear to be correct as channel lengths are shrunk, regardless of the model employed for predicting ΔV_{Tsc}. Yau's model also permits qualitative insight of V_T lowering with decreasing L to be gained, even if the quantitative ΔV_{Tsc} values do not accurately reflect experimental observations.

By modifying Eq. 5-2, Yau's model can also be used to estimate the effect of drain voltage on ΔV_{Tsc}. That is, in Eq. 5-2 it is assumed that the width of the source, drain, and channel depletion regions are equal. However, under drain bias the drain depletion region grows wider. Thus these three depletion regions no longer have equal widths. To reflect the effect of drain-width variation on ΔV_{Tsc}, Eq. 5-2 is reformulated as follows:[4]

$$\Delta V_{Tsc} = \frac{q\, N_{sub}\, d_{max}\, r_j \left(\left[\sqrt{\left\{ 1 + \frac{2\, W_S}{r_j} \right\}} - 1 \right] + \left[\sqrt{\left\{ 1 + \frac{2\, W_D}{r_j} \right\}} - 1 \right] \right)}{2\, C_{ox}\, L} \qquad (5 - 3)$$

Despite the appealing simplicity of the charge-sharing models, they fail to give good quantitative agreement with measured V_T values in devices with very short channels (e.g., < 1 μm) or under large drain voltages. This apparently occurs because several of the assumptions used to derive the model are not valid, especially under short-channel and large V_{DS} conditions. These include the following:

1. In Yau's model the induced charge is arbitrarily apportioned among the gate, source, and drain, whereas an accurate apportionment can only be accomplished by numerical integration of Poisson's equation and graphical analysis.

2. In the long-channel MOSFET model we assume that in subthreshold operation the entire V_{DS} is dropped across the drain depletion region. Thus, the channel potential in the rest of the channel region is essentially independent of V_{DS} and depends only on the gate bias. This also implies φ_{surf} in this region is independent of position (i.e., its value is constant with position along the channel). However, as we shall see this assumption breaks down in the case of short-channel MOSFETs (and as is evident from the data shown in Fig. 5-3). Yau's model,

however, continues to assume that the potential in the channel region is constant with position. The portion of the charge that is shared is assumed to be distributed evenly under the gate to uniformly reduce the depletion charge density in the gate area. Obviously, this is not what actually happens physically.

These and other questionable assumptions of Yau's model are examined in ref. 5.

Nevertheless, despite the fact that charge-sharing models generally do not provide accurate quantitative results, they are still useful for providing first-order estimates of ΔV_{TSC}, as well as for helping to intuitively visualize the physical basis of changes in V_T as L is decreased.

5.2.1.3 Barrier-Lowering Models and DIBL. The *barrier-lowering* perspective introduced by Troutman in 1979, provides a more rigorous foundation for deriving analytical models of ΔV_{TSC}.[6] It seeks to determine the change in the potential energy barrier between the source and channel at the Si interface caused by the junction depletion region charges (which in turn arise from both the built-in *pn* junction $\mathcal{E}$-field and the reverse-bias drain voltage) as well as the gate charge.

Troutman made measurements of DIBL,[6] and he and others[1,2,6] performed numerical analysis of the Poisson equation (Eq. 5-1) to match experimental and theoretical results. As noted earlier, numerical analysis is necessary to obtain more accurate results. However, because of the extensive computational procedures required, this solution approach is not appropriate for applications involving computer-aided circuit design. As a result, other workers sought to develop approximate analytical solutions of Eq. 5-1 which would be suitable for such applications.

This is done by finding analytical models of the surface potential φ_{surf} which relate it to the MOSFET structural parameters (e.g., t_{ox}, r_j, L), substrate doping levels, and applied biases. The value of φ_{surf} from these models is used in the threshold voltage equation to calculate the modified V_T value. The mathematical approach is to try and discover approximate analytical solutions to the 2-D Poisson equation, generally by converting it to a form that contains a differential term in only one dimension. This approach may be fruitful in this case, since we are only interested in finding the minimum value of φ at the surface (i.e., of φ_{surf}) along the y-direction from source to drain, i.e., we want to determine $[\varphi_{surf}(y)]_{min}$. In most cases analytical solutions with approximate boundary conditions are proposed as the model. The accuracy depends on the degree of approximation in the boundary conditions.

Many DIBL-based analytical models have been developed. In general, they are more complex than the charge-sharing models. One of the simplest and most widely adopted is that of Hsu et al.[7] In this model, the 2-D Poisson equation is reduced to a 1-D equation by essentially approximating the $(\partial^2\varphi/\partial x^2)$ term of the 2-D equation as a constant. With the Poisson equation thereby converted into a 1-D form, it becomes possible to find an analytical solution, and Hsu's solution becomes the model for calculating ΔV_{TSC}. Jain and Balk extend this model to include the effect of non-uniform channel width.[8] They also point out that both the Hsu model and their extension are expected to be accurate only for small values of DIBL (i.e., for devices having channels that are not too short and in which V_{DS} is not too large). In addition, these models were derived for MOSFET structures with uniformly doped channel regions.

More recently, Skotnicki et al., reported another approach which allows the 2-D Poisson equation to be reduced to a 1-D equation using a so-called *voltage-doping transformation.* Using this technique they derived an analytical barrier-lowering model that is applicable to MOSFETs with non-uniformly doped channels.[9] Good agreement with measured data is also reported for devices with L as small as 0.8 μm (and in which V_{DS} values are as large as 3V). A final approach, as described in the next section, also provides good agreement with measured data over the ranges of voltages studied.

Nevertheless, the accuracy of the barrier-lowering-based analytical models has not yet been confirmed for predicting the value of ΔV_{Tsc} in all device structures. As a result, it may still be prudent to use empirical measurements and/or numerical analysis to obtain accurate estimates of barrier lowering in specific device structures.

5.2.1.4 Quasi-Two-Dimensional Analytical Solution for ΔV_{Tsc}. By using a quasi-2-D approach to solve the 2-D Poisson equation in the channel of the MOSFET, an analytical model has been derived that reportedly predicts ΔV_{Tsc} accurately, even for devices with channel lengths below 0.5 μm.[22] It begins by approximating the 2-D Poisson equation in the channel with the expression:

$$\frac{\varepsilon_{si}\, d_{max}}{\eta} \frac{d\mathcal{E}_y(y)}{dy} + \varepsilon_{ox} \frac{[V_T - V_{FB} - \varphi_{surf}(y)]}{t_{ox}} = q\, N_{sub}\, d_{max} \qquad (5 - 4a)$$

or

$$\frac{\varepsilon_{si}\, d_{max}}{\eta} \frac{d^2\varphi_{surf}(y)}{dy^2} + \varepsilon_{ox} \frac{[V_T - V_{FB} - \varphi_{surf}(y)]}{t_{ox}} = q\, N_{sub}\, d_{max} \qquad (5 - 4b)$$

where the parameters of this equation have been defined previously (except for η, which is a fitting parameter). The second term on the left-hand side is an approximation of the term $\partial\mathcal{E}_x/\partial x$ in the 2-D form of the Poisson equation. This quasi-2-D approach is also used in deriving a model for I_{Dsat} and thus, the details involved in transforming Eq. 5-1 into a form akin to Eq. 5-4 are given in section 5.4.3.

The solution to Eq. 5-4, namely $\varphi_{surf}(y)$, under the boundary conditions $\varphi_{surf}(0) = V_{bi}$ and $\varphi_{surf}(L) = V_{DS} + V_{bi}$ is:

$$\varphi_{surf}(y) = V_{sL} + (V_{bi} + V_{DS} - V_{sL}) \frac{\sinh(y/l)}{\sinh(L/l)} + (V_{bi} - V_{sL}) \frac{\sinh(L - y)/l)}{\sinh(L/l)} \qquad (5 - 5)$$

In Eq. 5-5 $V_{sL} = V_{GS} - V_T$, V_{bi} is the built-in potential between the source-substrate and drain-substrate junctions, and l is the characteristic length defined as

$$l = \sqrt{\frac{\varepsilon_{si}\, t_{ox}\, d_{max}}{\varepsilon_{ox}\, \eta}} \qquad (5 - 6)$$

Equation 5-5 can now be plotted for a sufficient number of points between $y = 0$ and $y = L$ (e.g., 10 or more). From such plotted data, and a fitted curve, the minimum value of $\varphi_{surf}(y)$ can be found. By subtracting the long-channel value of φ_{surf} at V_T from $\varphi_{surf}(y)_{min}$, we can find ΔV_{Tsc}. Figure 5-5 plots Eq. 5-5 for φ_{surf} at a given V_{GS}, and for two values of V_{DS} [i.e., $V_{DS} = 0.05$V (dashed lines) and $V_{DS} = 1.5$V (solid lines)] for

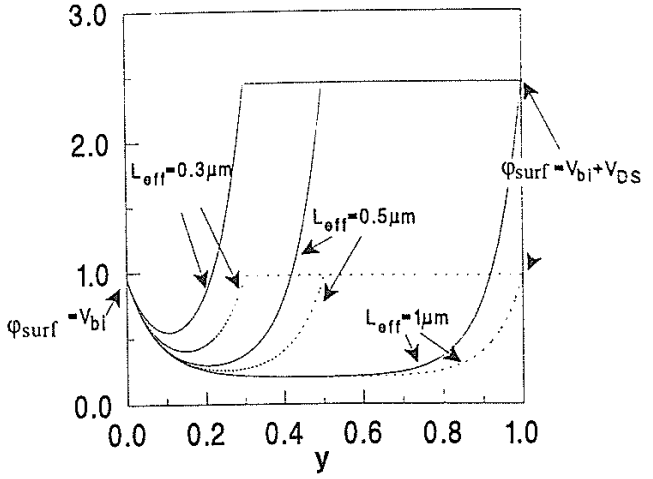


Fig. 5-5 (a) Calculated surface potential along channel for different channel lengths using Eq. 5-5. The device parameters are t_{ox} = 100Å, N_{sub} = N_A = 10^{16} cm^{-3}, n^+ = 10^{20} cm^{-3}, and η = 1, i.e., l = 0.1 μm.[22] (© IEEE 1993).

MOSFETs with different channel lengths. In this example, the device with L_{eff} = 1 μm has a region where φ_{surf} is constant as a function of position along a large fraction of the channel length (i.e., at a value of φ_{surf} ~0.2V in this figure), and this value is determined merely by the gate bias (and not V_{DS}). A constant value of φ_{surf} versus y is indicative of long-channel behavior and V_T in this device can thus be accurately found using Eq. 4-74.

In Fig. 5-5, devices with channel lengths 0.5 μm or smaller have a minimum value of φ_{surf} greater than the long-channel φ_{surf} value (i.e., $|2\varphi_B|$), and no region of constant $\varphi_{surf}(y)$ is evident, even for small values of V_{DS}. (This is different than the charge sharing approach, in which a region of constant [but somewhat larger] φ_{surf} versus distance is assumed to exist even in very short length devices.) The results of Fig. 5-5 have been verified by 2D numerical simulation using such programs as MINIMOS and PISCES (e.g., as shown in plots of φ_{surf} versus y for various L values, see Fig. 5-3).

The minimum value of $\varphi_{surf}(y)$ is also seen in Fig. 5-5 to increase (i.e., the potential energy barrier for electron flow between source and drain will decrease) with decreasing channel length and increasing V_{DS}. If the minimum long-channel value of φ_{surf} is subtracted from the minimum value calculated from Eq. 5-5 we get an expression for ΔV_{TSC}. When V_{DS} is small, this expression can be approximated as

$$\Delta V_{TSC} = [2(V_{bi} - 2\varphi_B) + V_{DS}](e^{-L/2l} + 2e^{-L/l}) \qquad (5-7)$$

Figure 5-6a shows the value of ΔV_{TSC} versus L as calculated with Eq. 5-7 as well as with 2D numerical simulation. Agreement with the numerical simulation results is seen to be very good. It should also be noted that if $L > 5l$, then this equation reduces to a single exponential term, much like the one derived from the simpler barrier-lowering

models described in the previous section.* For most technologies l is about 0.1-0.15 μm, hence these simple models give accurate results for ΔV_{Tsc} only if the channel length exceeds 0.5-0.8 μm.

When V_{DS} is not small, the expression for ΔV_{Tsc} from this model is more complex, with its value when $L > 5l$ given approximately as

$$\Delta V_{Tsc} = 3(V_{bi} - 2\varphi_B) + V_{DS})\, e^{-L/l} + 2\sqrt{(V_{bi} - 2\varphi_B)(V_{bi} - 2\varphi_B + V_{DS})}\, e^{-L/2l} \quad (5 - 8)$$

Eq. 5-8 reduces to Eq. 5-7 for large L/l and small V_{DS} as expected. Figure 5-6b shows the results of calculating V_T versus V_{DS} for $L_{eff} = 0.3$ μm with Eqs. 5-7 and 5-8 as well as with 2D numerical simulation. Eq. 5-8 is seen to compare well with the 2D numerical simulation results, but Eq. 5-7 overestimates ΔV_{Tsc} at high V_{DS}. In Fig. 5-7 the value of V_T is calculated versus L using the charge sharing model, 2D numerical simulation, and Eq. 5-8. The results using Eq. 5-8 and numerical simulation are comparable. Finally, Fig. 5-8 compares the values of V_T versus L (with V_{BS} as the parameter) calculated with Eq. 5-8 and with measured V_T values in non-LDD

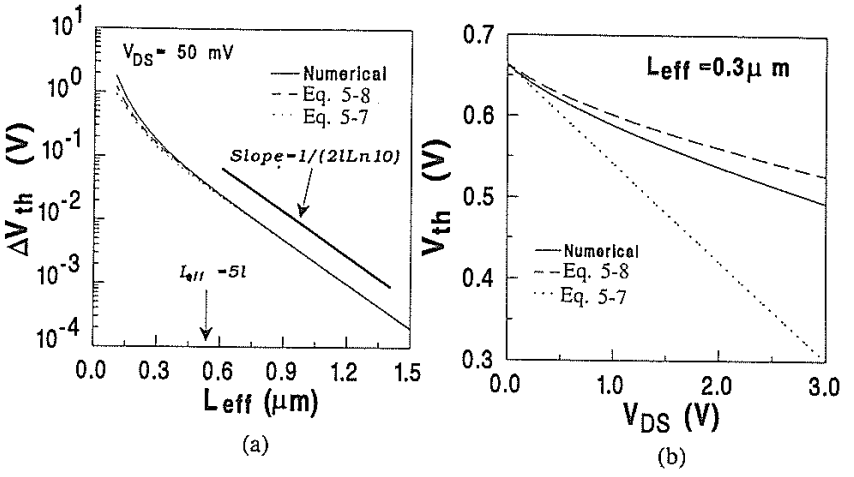


Fig. 5-6 (a) The calculated ΔV_T shifts versus channel length at V_{DS} = 0.005V. Note that the values obtained from the simple analytical models agree well with the numerical solution. The device parameters are the same as those in Fig. 5-5. Note that when $L > 5l$, all the curves have the same slope of l ($2l$ in Eq. 5-8). (b) Comparison between simple analytical solutions and numerical solution of threshold voltage versus drain voltage. The solution based on $y_O = L/2$ (Eq. 5-7) overestimates the variation in threshold voltage.[22] (© 1993 IEEE).

* If $L = 5l$ then the first exponential term in Eq. 5-7 is $e^{-2.5}$ while the second term is e^{-5}, which is much smaller than $e^{-2.5}$. Thus, for $L>5l$, the second exponential term in Eq. 5-7 can be neglected, leaving an expression with only one exponential term. For cases where L is smaller than ~$5l$, the second exponential term cannot be neglected, and the full Eq. 5-7 must be used to get accurate values of ΔV_{Tsc} when V_{DS} is small.

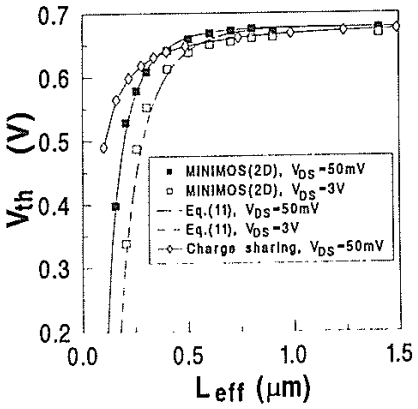


Fig. 5-7 Comparison of the V_T calculated using the charge sharing model, the two-dimensional numerical simulation (using MINIMOS), and the model of ref. 22. The device parameters are the same as those used in Fig. 5-5.[22] (© 1993 IEEE).

MOSFETs. Again, agreement is very good. The report describing this model also describes its application to LDD MOSFETs.

The characteristic length l defined in Eq. 5-6 needs to be more exactly determined in order for this model to give accurate values of ΔV_{TSC}. That is, due to the fitting parameter η in Eq. 5-6, the value of l from Eq. 5-6 only gives its correct order of magnitude. The exact value of l needs to be extracted from V_T data taken from actual MOSFETs.

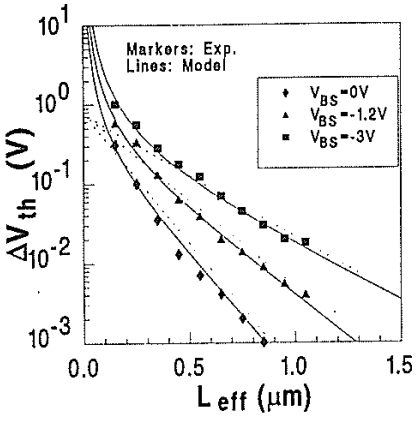


Fig. 5-8 ΔV_T versus L_{eff} at V_{DS} = 0.005V and different V_{BS} values for an non-LDD device. The solid lines are calculated results and the dashed lines are lines best fitting the experimental data of $L_{eff} >> 5l$. [22] (© 1993 IEEE).

An alternative form to Eq. 5-6 for expressing l which allows its exact value to be empirically extracted is based on the empirical expression of Brews et al[23] for the minimum channel length L_{min} a MOSFET must have in order to exhibit acceptable subthreshold behavior (see section 5.6.1.2 for more details on this expression):

$$L_{min} = 0.41\,(r_j\ t_{ox}\ d_{max}^2)^{1/3} \qquad (5 - 9)$$

If we assume L_{min} is about $4l$, using Eq. 5-9 yields an empirical expression for l as

$$l = 0.1\,(r_j\ t_{ox}\ d_{max}^2)^{1/3} \qquad (5 - 10)$$

According to this model, for an n-channel MOSFET with an n^+ poly gate, it can be shown that in order to maintain V_T at 0.7V, i.e.,

$$V_T = 0.7V = V_{FB} + 2\varphi_B + \frac{t_{ox}}{\varepsilon_{ox}}\frac{4\,\varepsilon_{si}\,\varphi_B}{d_{max}} \qquad (5 - 11)$$

it is necessary that

$$d_{max} = \frac{t_{ox}}{\varepsilon_{ox}}\ \frac{4\,\varepsilon_{si}\,\varphi_B}{(V_T - V_{FB} - 2\varphi_B)} \approx \frac{2\,\varepsilon_{si}}{\varepsilon_{ox}}\ t_{ox} \qquad (5 - 12)$$

By substituting Eq. 5-12 into Eq. 5-10, l can be rewritten as

$$l = 0.0007\ r_j^{1/3}\ t_{ox} \qquad (5 - 13)$$

where r_j and l are in μm and t_{ox} is in Å. On the other hand, for an n-channel MOSFET with a p^+ poly gate,

$$d_{max} = 4\,(\varepsilon_{si}/\varepsilon_{ox})\ t_{ox} \qquad (5 - 14)$$

(assuming $V_T = 1.2V$). Therefore one has $l \sim 0.0011\ r_j^{1/3} t_{ox}$. This helps explain why buried-channel devices, either NMOSFETs or PMOSFETs generally have worse short-channel effects than comparable surface-channel N- or PMOSFETs.

5.2.1.5 Measuring Surface-DIBL Effects. Surface-DIBL in MOSFET structures with a fixed L can be empirically characterized by measuring and plotting a set of log I_{Dst} versus V_{GS} curves with increasing values V_{DS} as the parameter (as shown in Fig. 5-9). When analyzing such a set of curves, it must first be ascertained that subsurface-punchthrough current is not flowing (see section 5.3 for a discussion on punchthrough). For example, in Fig. 5-9 there is no indication that such punchthrough current occurs over the applied V_{DS} range, since S_t remains constant as V_{DS} is increased.*

Once it is verified that punchthrough current is not significant, a quantitative estimate of surface-DIBL can be extracted from the shift in gate voltage ΔV_{GS} at a fixed drain current (typically 0.1 μA/μm) as the drain-source voltage is increased. That is, the

* If a device exhibits punchthrough current within the normal operating V_{DS} range, the structure must be redesigned to eliminate this behavior. Devices exhibiting punchthrough will exhibit excessive leakage current at $V_{GS} = 0V$ (i.e., when the device is nominally OFF).

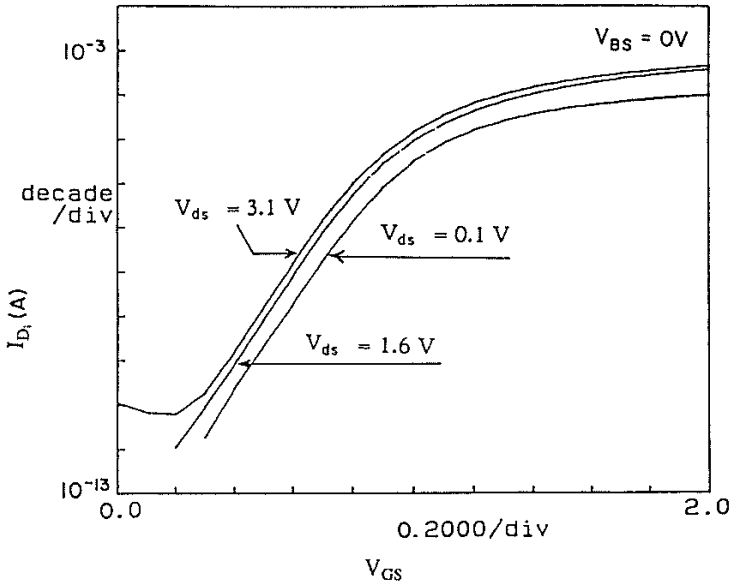


Fig. 5-9 Plot of log I_D versus V_{GS} for an NMOSFET operated below threshold, with V_{DS} as the parameter. The value of $\Delta V_{GS}/\Delta V_{DS}$ for a fixed value of I_D (e.g., 1×10^{-7} A/μm) is used to characterize DIBL.

ΔV_{GS} value represents a good first-order estimate of the expected shift in V_T as a function of drain bias. The shift in V_{GS} is normally expressed in normalized form as $\Delta V_{GS}/\Delta V_{DS}$. An acceptable value of $\Delta V_{GS}/\Delta V_{DS}$ is ~25 mV/V, depending on the circuit application and specific leakage requirements. $\Delta V_{GS}/\Delta V_{DS}$ is a strong function of channel length (Fig. 5-10) and must be guaranteed to at least 3-σ below the target channel length.

An alternative method of evaluating DIBL is to measure the change in drain current as V_{DS} is increased at some fixed gate voltage below V_T (e.g., the change ΔI_D in Fig. 5-10 when V_{DS} is increased from 0.1 [solid line in Fig. 5-10] to 5.0V [dashed line]). Because of its simplicity, this measurement is easily implemented on automated test equipment.

Another perspective on this last technique for measuring surface-DIBL is that such barrier lowering raises the electron density at the surface in subthreshold operation. That is, the electron density at the source end of the channel n_S(source) - where the surface potential is at a minimum - is increased due to surface-DIBL from its long-channel device value $n_S = (n_i^2/N_A)\exp(q\varphi_{surf}/kT)$ to

$$n_{SDIBL} = \left(\frac{n_i^2}{N_A}\right) e^{q\varphi_{surf}/kT}\, e^{q\Delta V_{Ts}/kT} \qquad (5 - 15)$$

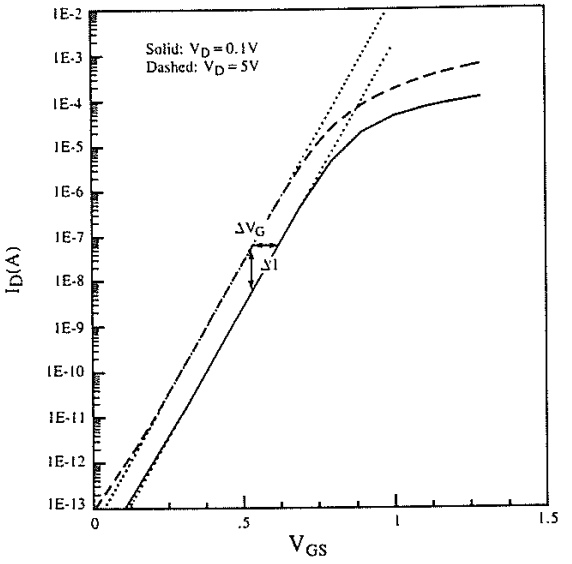


Fig. 5-10 Simulated values of I_D versus V_{GS} for two values of V_{DS} (0.1V and 5V) in a MOSFET with L_{eff} = 0.7V. From J. Teplik, Chap. 2, "Device Design," in *BiCMOS Technology and Applications*, A. Alvarez Ed., Klewer Academic Publishers, Boston MA, 1989. Reprinted with permission.

where ΔV_{TSC} will become larger as V_{DS} is increased. This also implies that surface-DIBL will cause I_{Dst} at a fixed value of V_{GS} and L to increase - by a factor of $\exp(q\Delta V_{TSC}/kT)$ - above that expected in a long-channel device.* Experimental data can be plotted to depict this effect in the following way: As shown in Fig. 5-11, log I_D is plotted versus V_{DS} with V_{GS} as the parameter. Figure 5-11a shows such characteristic curves for an NMOS device with $L = 2.1$ μm and Fig. 5-11b for a device with $L = 0.8$ μm. In both figures the curves exhibit two distinct regions. At low values of V_{DS}, I_D depends primarily on V_{GS} and increases only slightly with V_{DS} (*but it is this change in I_D that represents surface-DIBL*).** Eventually the I_D curves intercept a more steeply

* Note, however, that as V_{GS} is decreased, the value of n_s at the source will decrease as quickly with decreasing V_{GS} as in a long-channel device, because, the gate still has control of the charge giving rise to I_D. So, S_t will not change, even though V_T is smaller. But, when the device is in punchthrough, the gate no longer has the same control of charge flow as V_{GS} is reduced. Hence, drain current doesn't decrease as quickly with shrinking V_{GS}, or S_t gets larger in punchthrough.

** Note that according to the model of subthreshold behavior in long-channel MOSFETs at a fixed value of V_{GS}, I_{Dst} is independent of the drain voltage V_{DS} for V_{DS}>0.1V (see Eq. 4-90). But because of surface-DIBL, for a fixed value of L and V_{DS}, I_{Dst} increases as V_{DS} increases, as described in earlier sections.