Chapter 2: Devices, Circuits, and Systems

2.1 Introduction

The integrated circuit made modern day information processing and communications systems possible. Its basic functional element is the transistor, most commonly a silicon metal oxide semiconductor field-effect transistor (MOSFET). For the past forty years, MOSFET scaling (the reduction of its critical dimension by a factor of about $\sqrt{2}$ each technology generation, approximately 18 months) has driven Moore's Law (the doubling of the number of transistors per integrated circuit each technology generation). It now appears that the silicon MOSFET will reach its scaling limit within a decade or so [2.1, 2.2], and devices to complement or replace the silicon MOSFET are being explored [2.3].

This chapter is a brief overview of MOSFET essentials and a quick introduction to the bipolar transistor. The chapter provides some context for exploring new devices and an opportunity to discuss three important points. First, we discuss charge control by a gate electrode, which modulates the transistor's current. Electrostatics is likely to be similarly important for transistors that follow the MOSFET. Second, we discuss the characteristics of devices that make them useful in high-density, high-speed digital systems. Finally, we examine the fundamental limits that apply to any electronic switching device used for conventional, digital logic.

2.2 The MOSFET

Figure 2.1 illustrates the physical structure of two different kinds of MOSFETs. An n-channel bulk MOSFET is built on a p-type substrate with deep n^+ regions to facilitate contact to the source and drain. Shallow n^+ junctions connect the source and drain to the p-type channel. A thin gate oxide (typically still SiO_2 and about 1-2nm thick) separates the silicon channel from the gate electrode. Figure 2.1b shows a double gate MOSFET, which is built on a thin silicon film and with gates above and below the channel [2.2]. Numerous variations exist. Examples include the FinFET, a type of double gate MOSFET [2.3, 2.4], the tri-gate transistor [2.5], and the gate-all-around MOSFET [2.6].

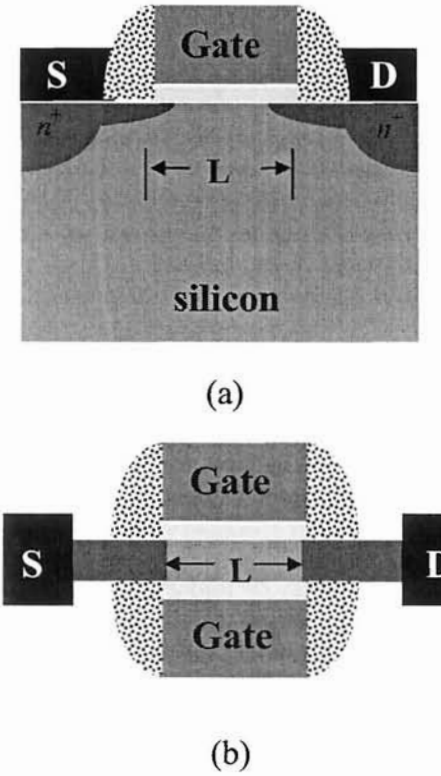


Fig. 2.1 Cross sectional sketches of: a) a bulk, silicon MOSFET and b) a double-gate MOSFET. The third dimension, the width, W, of the MOSFET, is into the page.

For any type of MOSFET, the gate voltage modulates the conductivity of the p-type channel by raising or lowering the height of an energy barrier between the source and channel, as shown in Fig. 2.2. Under low drain voltages (Fig. 2.2a), the device operates like a resistor with the gate voltage controlling the resistance, while under high drain bias (Fig. 2.2b), the device operates like a current source with the gate voltage controlling the magnitude of the current. The transistor designer's challenge is to engineer an appropriate energy barrier between the source and drain so that the device can be turned off while at the same time designing a gate structure that can effectively modulate the barrier and turn the transistor on. The design of a bulk MOSFET for proper electrical performance involves producing sophisticated two-dimensional doping profiles in the p-type channel, an ultra-thin gate oxide, and heavily doped, ultra-shallow source/drain extensions [2.7]. Double gate, tri-gate, and gate-all-around MOSFETs provides strong gate control of the channel conductivity, which allows the source and drain to be placed more closely. The channel length, L, sets the scale of the device. Device scaling refers to the process of shrinking L to reduce the device size, but a complete MOSFET is typically 10-15 times larger than L. The associated dimensions (oxide thickness, shallow extension junction depth, etc.) also need to be reduced accordingly to maintain good electrical characteristics.

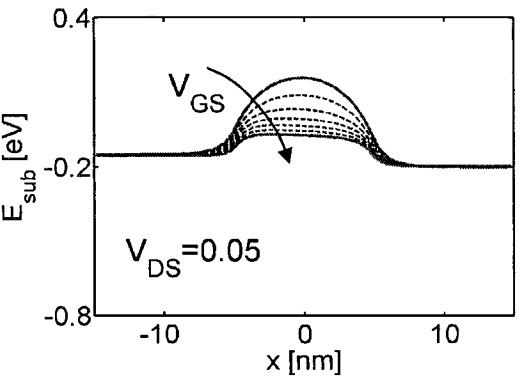
Figure 2.3 sketches the drain current, I_D, vs. drain-to-source voltage, V_{DS}, characteristics of the MOSFET. Because the MOSFET has four terminals there are several ways to plot these characteristics. In Fig. 2.3a, we plot I_D vs. V_{GS} on both linear and logarithmic axes. On a linear scale, essentially no current flows until the gate voltage reaches a critical value, the threshold voltage, V_T. On a logarithmic scale, we see that the drain current actually increases exponentially for $0 < V_{GS} < V_T$. Above threshold, I_D varies as the gate overdrive, $(V_{GS} - V_T)$ to a characteristic power, α. For low V_{DS}, $\alpha = 1$, but for high V_{DS}, $1 \leq \alpha \leq 2$. The maximum current, known as the *on-current*, occurs when the power supply voltage is applied between the drain and source and between the gate and source.

Figure 2.3b plots I_D vs. V_{DS} with V_{GS} as a parameter. For low V_{DS}, the MOSFET operates like a gate voltage dependent resistor, but for high V_{DS}, it operates more like a gate voltage controlled current source (with a finite output conductance). The voltage that separates these two regions is the so-called "drain saturation voltage," V_{Dsat}.

Figure 2.3c plots log I_D vs. V_{GS} at low and high drain voltages. The subthreshold region is characterized by its slope or, equivalently, the subthreshold swing, S, which is the number of millivolts of increase in gate voltage needed to increase the drain current by a factor of 10. For well-

designed MOSFETs, $S < 80$ mV/decade; the theoretical lower limit is 60 mV/decade at room temperature. Another important performance metric is the *off-current*, the current that flows when the $V_{DS} = V_{DD}$ and $V_{GS} = 0$. A good transistor should display a high on-current, a low off-current, and a rapid transition between the off and on states (i.e. a small S).

(a)



(b)

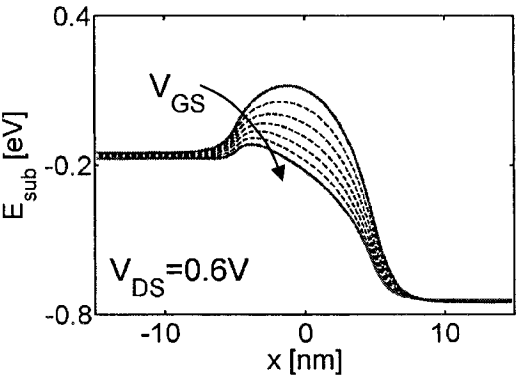


Fig. 2.2 Sketch of the minimum electron energy vs. position showing how an increasing gate voltage lowers the energy barrier between the source and drain. Two cases are shown: a) low drain voltage and b) high drain voltage.

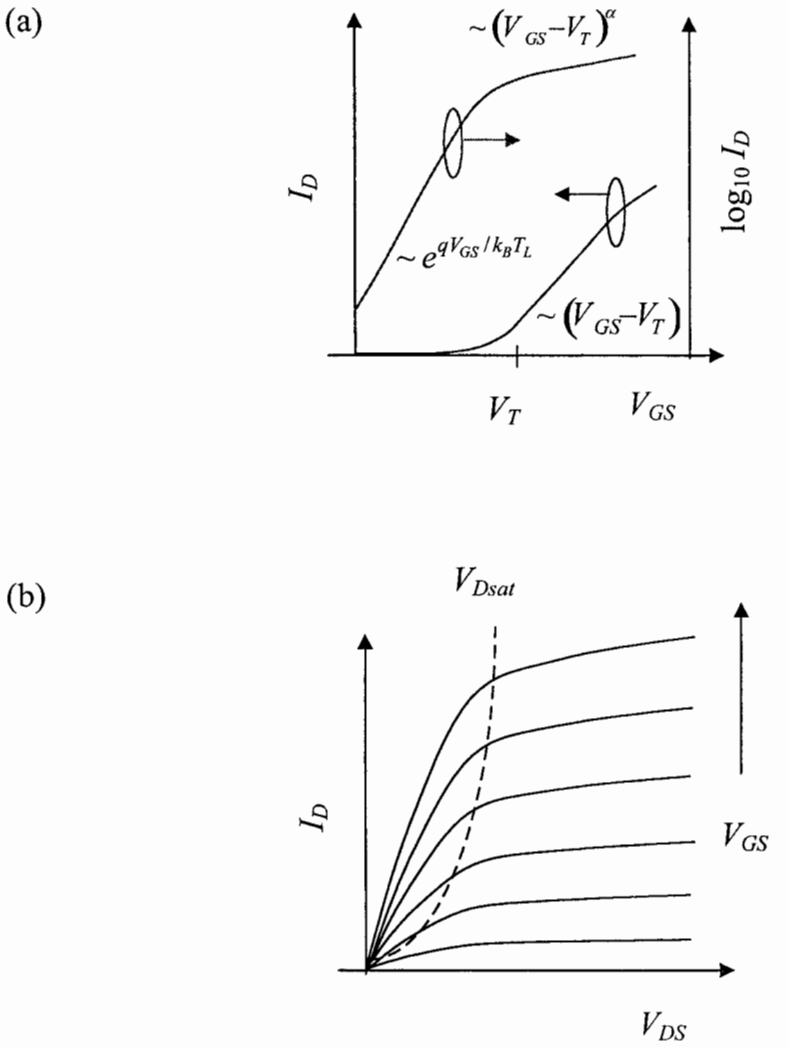
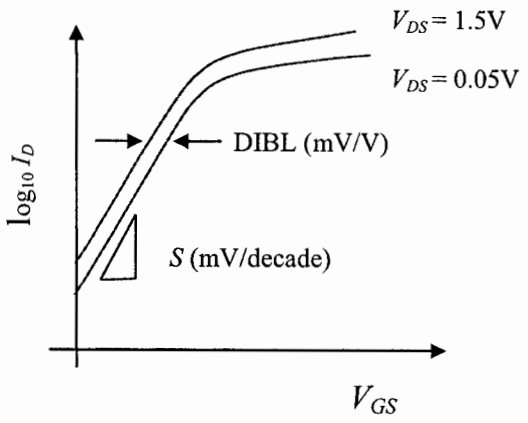


Fig. 2.3 The current vs. voltage characteristics of a MOSFET. a) I_D vs. V_{GS} for a fixed V_{DS}. b) I_D vs. V_{DS} with V_{GS} as a parameter.

(c)



(d)

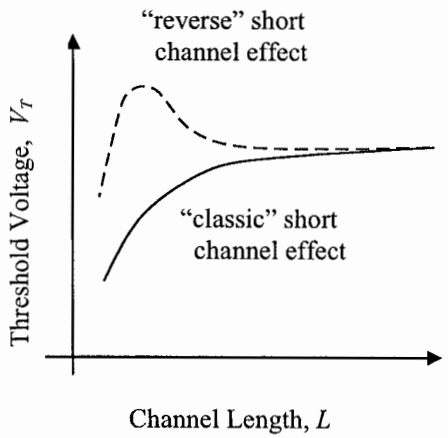


Fig. 2.3 The current vs. voltage characteristics of a MOSFET. c) log I_D vs. V_{GS} at low and high drain bias. d) The threshold voltage, V_T, vs. channel length.

Fig. 2.3c also shows that the I_D-V_{GS} characteristics for low and high V_{DS} are translated horizontally (for poorly designed MOSFETs, S also changes). The translation is known as DIBL (drain-induced barrier lowering) and is characterized by the number of millivolts of translation per volt of change in drain voltage. Well-designed MOSFETs typically have DIBL < 100 mV/V.

Finally, Fig. 2.3d sketches V_T vs. channel length, L. Two-dimensional electrostatic effects tend to reduce V_T as L decreases. Laterally non-uniform doping profiles can reverse this effect and produce an initial increase in V_T as L decreases. The goal of the transistor designer is to make V_T as nearly independent of L as possible.

In this section, we have described the physical structure and terminal I-V characteristic of the MOSFET. In the following sections, we highlight a few important concepts that we will make use of in later chapters. For a more extensive treatment, refer to Taur and Ning [2.7].

2.3 1D MOS Electrostatics

The most important thing to understand about a MOSFET is MOS electrostatics. We begin in 1D in equilibrium by examining a long channel MOSFET with $V_{DS} = 0$ (Fig. 2.4). Near the middle of the channel, there is no variation of potential with x. A positive voltage on the gate lowers the electron energy and bends the bands down by an amount, $q\psi_S$, as illustrated in Fig. 2.4b and 2.4c. Our goal is to determine how the charge in the semiconductor, Q_S in C/cm^2 varies with surface potential, ψ_S, or alternatively, with gate voltage, V_{GS}. Later, we will seek to understand two-dimensional electrostatics and the influence of V_{DS}.

A direct approach to finding $Q_S(\psi_S)$ is to solve Poisson's equation,

$$\frac{d^2\psi}{dy^2} = -\frac{\rho}{\varepsilon_{Si}}. \tag{2.1}$$

The charge density, ρ, is related to the mobile carrier densities, $n(y)$ and $p(y)$, which are related to band bending and, therefore, to $\psi(y)$. The result is the so-called "Poisson-Boltzmann equation," which can be solved numerically for $Q_S(\psi_S)$ [2.7, 2.8]. We seek a simpler approach in order to understand the essence of the problem.

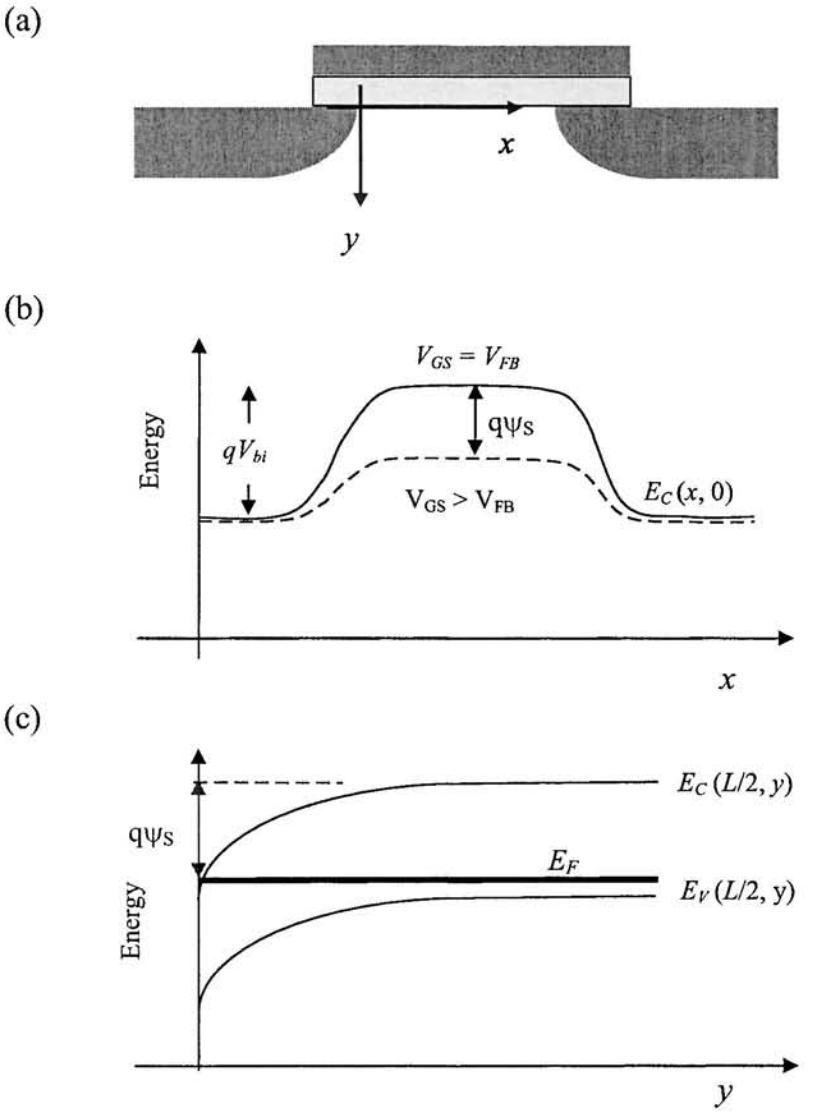


Fig. 2.4 Energy band diagrams for a silicon MOSFET in equilibrium. (a) the coordinate system, (b) $E_C(x,0)$, the conduction band edge at the oxide/silicon interface along the channel from source to drain. (c) $E_C(L/2,y)$, the conduction band edge in the middle of the channel vs. position into the silicon bulk.

The charge in the semiconductor has two components,

$$Q_S(\psi_S) = Q_B(\psi_S) + Q_i(\psi_S), \tag{2.2}$$

where Q_B, the bulk charge, is due to the depletion of majority carriers, and Q_i, the mobile charge, is due to inversion (or accumulation) layers of mobile carriers. Consider first a small, positive ψ_S for which we have only bulk charge described by depletion layer theory as [2.7, 2.8]

$$Q_B(\psi_S) = -qN_A W_D = -\sqrt{2q\varepsilon_{Si}N_A\psi_S}\ , \tag{2.3}$$

where W_D is the width of the surface depletion layer. The small mobile charge, $Q_i = -qn_S$ C/cm^2, does not affect the electrostatics, but it does give rise to the subthreshold current. Note that

$$n(y) = n_{op}e^{q\psi(y)/k_BT_L} = \frac{n_i^2}{N_A}e^{q\psi(y)/k_BT_L}\ \text{cm}^{-3}, \tag{2.4}$$

where n_{op} is the equilibrium minority electron density in the p-type bulk. The integrated electron density per cm^2 is

$$n_S = \int_0^\infty n(y)dy = \int_0^\infty \frac{n(y)}{d\psi/dy}d\psi\ \text{cm}^{-2}. \tag{2.5}$$

Since $n(y)$ falls rapidly for $y > 0$, we can approximate eq. (2.5) as

$$n_S(\psi_S) \cong -\frac{1}{E_S}\int_{\psi_S}^{0}\frac{n_i^2}{N_A}e^{q\psi/k_BT_L}d\psi = \left(\frac{k_BT_L/q}{\mathrm{E}_S}\right)\frac{n_i^2}{N_A}e^{q\psi_S/k_BT_L}, \tag{2.6}$$

where E_S is the electric field at the surface of the silicon. Finally, we have

$$Q_i(\psi_S) = -q\,n_S(\psi_S) = -q\left(\frac{k_BT_L/q}{\mathrm{E}_S}\right)n(0) \equiv -q\,n(0)W_{inv}, \tag{2.7}$$

where $n(0)$ is the electron concentration per cm^3 at $y = 0$, and $(k_BT_L/q)/\mathrm{E}_S$ is interpreted as the effective width of the inversion layer, W_{inv}.

As long as ψ_S is not too large, $Q_i(\psi_S) << Q_B(\psi_S)$ and

$$Q_S(\psi_S) \approx Q_B(\psi_S) \propto \sqrt{\psi_S} \qquad \psi_S < 2\psi_B \tag{2.8}$$

as illustrated in Fig. 2.5a. When ψ_S is greater than about

$$2\psi_B = \frac{2k_BT_L}{q}\ln(N_A/n_i), \tag{2.9}$$

then $Q_i(\psi_S) >> Q_B(\psi_S)$. In this case,

$$Q_S \approx Q_i \approx -\varepsilon_{Si}\mathrm{E}_S, \tag{2.10}$$

which can be used in eq. (2.7) to find

$$Q_S \propto e^{q\psi_S/2k_BT_L} \quad . \qquad \psi_S > 2\psi_B \tag{2.11}$$

Similar arguments apply to the heavily accumulated region, $\psi_S < 0$, where the charge is due to the accumulation of majority carrier holes, so putting it all together, we obtain the $Q_S(\psi_S)$ characteristic as sketched in Fig. 2.5a. Our simple arguments establish the shape of the $Q_S(\psi_S)$ characteristic in accumulation ($\psi_S < 0$), depletion ($0 < \psi_S < 2\psi_B$), and inversion ($\psi_S > 2\psi_B$). The complete $Q_S(\psi_S)$ can be evaluated by solving the Poisson-Boltzmann equation [2.7, 2.8].

The Semiconductor Charge vs. Gate Voltage:

Having understood the $Q_S(\psi_S)$ characteristic, we now turn to $Q_S(V_{GS})$. The voltage at the gate is [2.7, 2.8]

$$V'_{GS} = \psi_S + \Delta V_{ox} = \psi_S - \frac{Q_S(\psi_S)}{C_{ox}}, \tag{2.12a}$$

where

$$C_{ox} \equiv \frac{\varepsilon_{ox}}{t_{ox}}, \tag{2.12b}$$

and the second expression arises because $\Delta V_{ox} = \mathrm{E}_{ox}t_{ox}$ and $\varepsilon_{ox}\mathrm{E}_{ox} = -Q_S$. In Eqn. (2.12a), $V'_{GS} = V_{GS} - V_{FB}$ where V_{FB} is the flatband voltage, the voltage at which there is no band bending in the semiconductor. Its value is determined by the gate to semiconductor workfunction difference and by charges at the oxide-silicon interface [2.7, 2.8]. So, having determined $Q_S(\psi_S)$, we can use (2.12) to translate it to a $Q_S(V'_{GS})$ characteristic. The resulting $Q_S(V'_{GS})$ characteristic sketched in Fig. 2.5b should be compared to the $Q_S(\psi_S)$ characteristic of Fig. 2.5a.

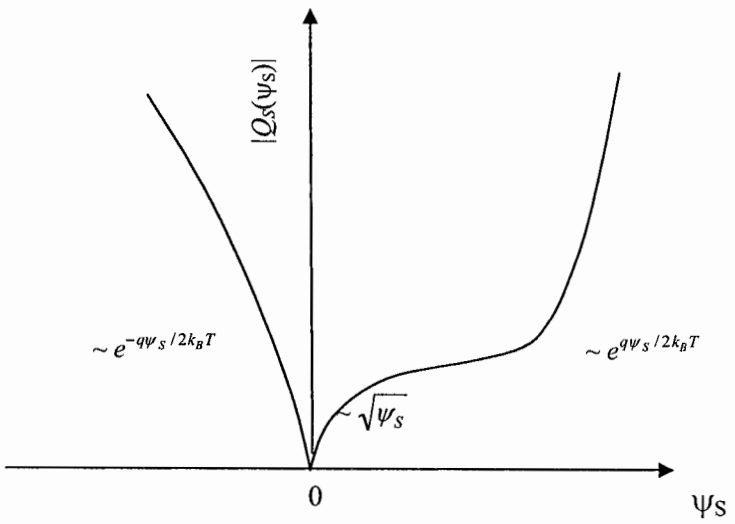


Fig. 2.5a Charge in the semiconductor, Q_S, vs. surface potential for p-type silicon.

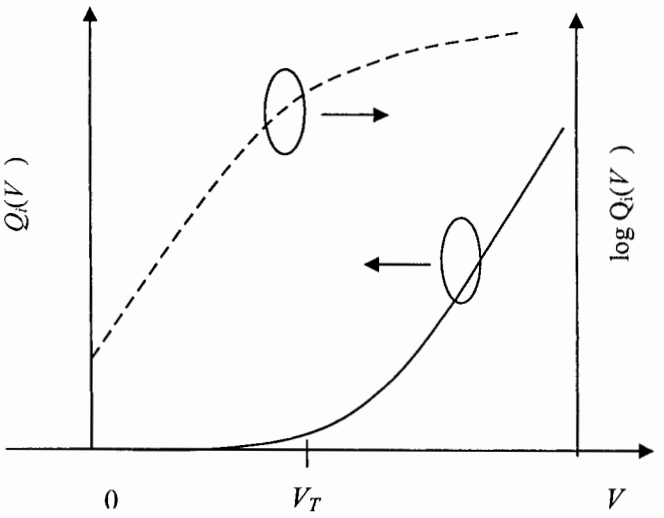


Fig. 2.5b Inversion layer charge vs. gate voltage. On a linear scale (solid line), we see that the inversion layer charge becomes significant above the threshold voltage, V_T where it varies linearly with $(V_G - V_T)$. On a logarithmic scale (dashed line), we see that the inversion layer charge varies exponentially with gate voltage below threshold.

Above threshold, where $Q_i >> Q_B$, the $Q_i(V_{GS})$ relation becomes simple. Consider a Taylor series expansion of Q_i about V_T

$$Q_i(V_{GS}) = Q_i(V_T) + \frac{dQ_i}{dV_{GS}}(V_{GS} - V_T). \tag{2.13a}$$

Assuming $Q_i(V_T) \cong 0$ and using the chain rule, eq. (2.13a) becomes

$$Q_i(V_{GS}) = \frac{dQ_i}{d\psi_S} \cdot \frac{d\psi_S}{dV_{GS}}(V_{GS} - V_T) \ . \tag{2.13b}$$

The *semiconductor capacitance* is

$$-\frac{dQ_S}{d\psi_S} = C_S, \tag{2.14a}$$

and above threshold,

$$C_S \approx -\frac{dQ_i}{d\psi_S} = C_{inv}. \tag{2.14b}$$

From eq. (2.12a), we also have

$$\frac{dV_{GS}}{d\psi_S} = 1 + \frac{C_{inv}}{C_{ox}} \ . \tag{2.15}$$

After using eqs. (2.14b) and (2.15) in (2.13b), we find

$$Q_i \cong -C_G(V_{GS} - V_T), \qquad V_{GS} > V_T \tag{2.16a}$$

where

$$C_G \equiv \frac{C_{ox}C_{inv}}{C_{ox} + C_{inv}} < C_{ox} \ . \tag{2.16b}$$

The gate capacitance is the series combination of the oxide capacitance and the semiconductor capacitance. Phenomenologically, we can write

$$C_S = \frac{\varepsilon_S}{W_{inv}}, \tag{2.17}$$

which provides another way to define the inversion layer width. (See eqn. (2.7) for the other definition.) Because inversion layers are typically thin (~ 5 nm), MOSFET analysis has traditionally assumed that $C_{inv} >> C_{ox}$ so that $C_G \approx C_{ox}$ for above threshold operation.

An aside on the quantum capacitance:

The inversion layer capacitance, is closely related to the "quantum capacitance," and is becoming increasingly important as device scaling decreases t_{ox} and increases C_{ox}. We can evaluate C_{inv} from eqns. (2.7) and (2.14), but (2.7) assumes Boltzmann statistics, and above threshold, degenerate statistics should be used. As an illustration, consider a fully degenerate (T = 0K) case for a quantum well with one subband occupied. As shown in Fig. 2.6, a gate potential raises or lowers the subband. The quantum capacitance is

$$C_S = q\frac{\partial n_S}{\partial \psi_S}, \tag{2.18}$$

and

$$n_s = D_{2D} \times (E_F - \varepsilon_1) \tag{2.19}$$

where D_{2D} is the constant 2D density-of-states and

$$\varepsilon_1 = \varepsilon_{10} - q\psi_S. \tag{2.20}$$

Using eqs. (2.19) and (2.20) in (2.18) we find

$$C_S = q^2 D_{2D}\,, \qquad (T_L = 0\text{K}) \tag{2.21}$$

so the quantum capacitance is proportional to the density of states. According to eqn. (2.16), a large inversion layer capacitance is beneficial for inducing charge in a semiconductor. From this perspective, a large effective mass is beneficial, but as we shall see later, transport suffers when the effective mass is large.

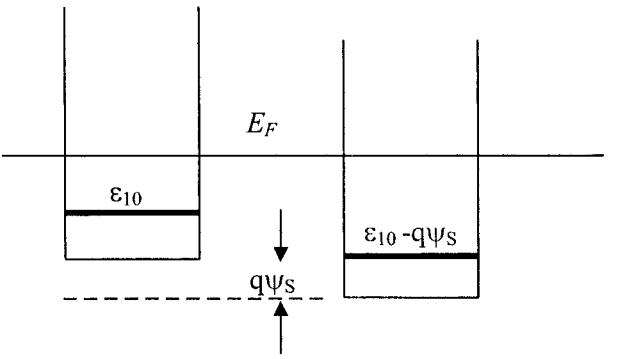


Fig. 2.6 Illustration of the origin of quantum capacitance.

Equation (2.21) shows that the T_L = 0K quantum capacitance is proportional to the density of states. To derive the quantum capacitance for $T_L > 0$K, we generalize eqn. (2.19) to

$$n_S = \int_{\varepsilon_1}^{\infty} D_{2D} f(E - E_F) dE = \int_0^{\infty} D_{2D} f(E' - E_F + \varepsilon_{10} - q\psi_S) dE' \qquad (2.22)$$

from which we can derive the quantum capacitance as in eqn. (2.18) as

$$C_S = q\frac{\partial n_S}{\partial \psi_S} = q\int_0^{\infty} D_{2D}\frac{\partial f}{\partial \psi_S} dE = q^2 \int_0^{\infty} D_{2D}\left(-\frac{\partial f}{\partial E}\right) dE = q^2 \langle D_{2D}(E_F)\rangle . \qquad (2.23)$$

The factor, $-(\partial f/\partial E)$ acts like a δ-function with a width of about $k_B T_L$ at the Fermi level, so the quantum capacitance is proportional to the average density of states at the Fermi level.

Inversion layer charge below threshold:

Having obtained $Q_i(V_{GS})$ above threshold [eqn. (2.16)] we seek a corresponding relation below V_T. Below V_T $Q_i << Q_B$, but it is important because it carries the subthreshold current of a MOSFET. As sketched in Fig. 2.5, the subthreshold inversion layer charge is observed to vary exponentially with gate voltage. Our objective is to derive an expression that describes this behavior. Instead of (2.10) we have

$$\varepsilon_{Si} E_S = -Q_B = qN_A W_D, \qquad (2.24a)$$

where W_D is the width of the surface depletion region. Equation (2.24a) can also be expressed as

$$E_S = qN_A W_D / \varepsilon_{Si} = q N_A / C_D, \qquad (2.24b)$$

where

$$C_D \equiv \varepsilon_{Si} / W_D \qquad (2.24c)$$

is the semiconductor depletion capacitance. With eqn. (2.24b), eqn. (2.7) becomes

$$Q_i(\psi_S) = -\frac{k_B T_L}{q} C_D \left(\frac{n_i}{N_A}\right)^2 e^{q\psi_S / k_B T_L} . \qquad (2.25)$$

Equation (2.25) shows that Q_i varies exponentially with surface potential, but we seek $Q_i(V_{GS})$.

If we were to use eqn. (2.12) to relate ψ_S to V_{GS}, the result would not be pretty. Alternatively, recall that eqn. (2.16b) implies the equivalent circuit of Fig. 2.7. Voltage division gives

$$\psi_S = \frac{C_{ox}}{C_{ox}+C_D} V_G' = \frac{V_G'}{m}, \tag{2.26}$$

where

$$m \equiv 1 + C_D / C_{ox}. \tag{2.27}$$

Equation (2.26) is not exactly correct, because C_D is non-linear, so the appropriate average depletion layer depth should be used. At the threshold voltage where $V_G' = V_T'$, $\psi_S = 2\psi_B$, we have

$$2\psi_B = \frac{k_B T_L}{q} \ln\left(\frac{N_A}{n_i}\right)^2 = \frac{V_T'}{m}. \tag{2.28}$$

Finally, using eqns. (2.26) - (2.28) in eqn. (2.25), we obtain

$$Q_i(V_{GS}) = -(m-1) C_{ox} \frac{k_B T_L}{q} e^{q(V_{GS}-V_T)/mk_B T_L}. \tag{2.29}$$

Equations (2.16a) and (2.29) explain the $Q_i(V_G)$ characteristic sketched in Fig. 2.5, which shows that Q_i varies exponentially with gate voltage below V_T and linearly with gate voltage above V_T. (This plot should be compared to Fig. 2.5a, which plots $Q_S(\psi_S)$.) Having established the key features of 1D MOS electrostatics, we must now consider the 2D effects that arise from the drain-to-source voltage.

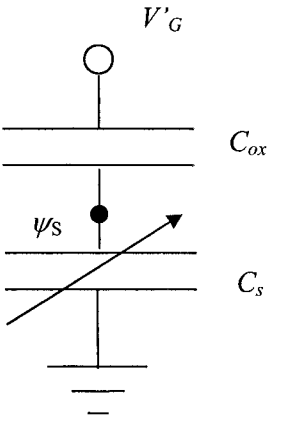


Fig. 2.7 Illustration of voltage division for the MOS capacitor. Below threshold, $C_S = C_D$, the depletion layer capacitance, and above threshold, $C_S = C_{inv}$, the inversion layer capacitance.

2.4 2D MOS Electrostatics

To understand a MOSFET, we need to include the effect of the drain potential and understand $Q_S(V_{GS},V_{DS})$. Figure 2.8 is a sketch of $E_C(x,0)$ vs. x from source to drain. The equilibrium case is shown in Fig. 2.8a, which should be compared to Fig. 2.4b. The pn junctions between the source/drain and channel have a built-in potential under flatband conditions of

$$V_{bi} = \frac{k_B T_L}{q} \ln \frac{N_A N_D}{n_i^2}. \tag{2.30}$$

As shown in Fig. 2.8a, a gate voltage can increase ψ_S and lower the barrier between the source and channel. At the threshold of inversion, $\psi_S = 2\psi_B$, and the barrier height is

$$E_b = k_B T_L \ln\left(\frac{N_D}{N_A}\right). \tag{2.31}$$

For typical values ($N_D \cong 10^{20} cm^{-3}$ and $N_A = 10^{18}$ cm^{-3}) $E_b \approx 0.1$ eV. Above threshold, the barrier is even smaller.

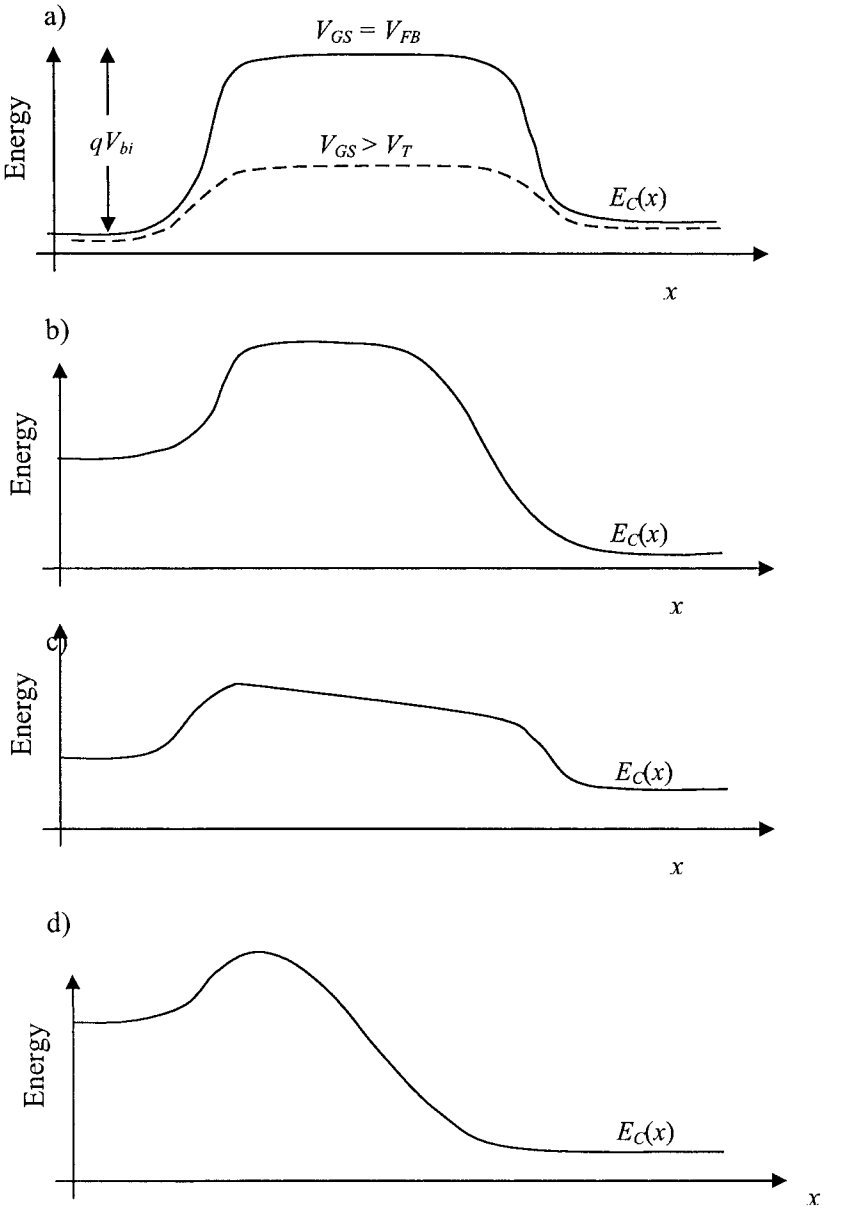


Fig. 2.8 Sketch of conduction band energy vs. position for four different bias conditions. a) equilibrium, $V_{GS} = V_{DS} = 0$; b) $0 < V_{GS} < V_T, V_{DS} >> 0$, c) $V_{GS} >> V_T, V_{DS} \approx 0$, d). $V_{GS} >> V_T, V_{DS} >> 0$.

Figure 2.8b is a sketch of $E_C(x,0)$ vs. x for $V_{GS} < V_T$ and $V_{DS} >> 0$. In this case, the source to channel barrier is large, so the drain current is small. The case for $V_{GS} >> V_T$ and V_{DS} small is shown in Fig. 2.8c. In this case, the potential drop is linear in the channel, and the device behaves like a resistor. Note that charge at the top of the barrier ($x = 0$) is approximately the value for the MOS capacitor in equilibrium. Finally, $E_C(x,0)$ vs. x for $V_{GS} >> V_T$ and $V_{DS} >> 0$ is shown in Fig. 2.8d. In this case, the channel potential is non-linear, but the important point is that for an electrostatically well-designed MOSFET, $Q_S(x=0)$ *is still approximately what it was in equilibrium,* when $V_{DS} = 0$. The goal of MOSFET design is to manage the 2D electrostatics so that $Q_S(x = 0)$ is nearly independent of V_{DS} with an approximate value of $C_{GS}(V_{GS} - V_T)$.

Two Dimensional Electrostatics in MOSFETs

The electrostatic design of a MOSFET begins by solving the 2D Poisson equation,

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} = -\frac{\rho}{\varepsilon_{Si}} = +\frac{qN_A}{\varepsilon_{Si}}, \tag{2.32}$$

where the first term accounts for the effect of the drain and source potentials and the second describes the effect of the gate potential as in the 1D MOS capacitor. (See Fig. 2.4 for a definition of the x and y directions.) We assume subthreshold operation, so there is negligible mobile charge. Given a MOSFET structure, eqn. (2.32) can be solved numerically. We will describe a simple, phenomenological approach that gives insight into the nature of the numerical solutions.

The second term in eqn. (2.32) can be expressed phenomenologically as

$$\frac{\partial^2 \psi}{\partial y^2} = \frac{(V'_{GS} - \psi_S)}{\Lambda^2}, \tag{2.33}$$

where Λ is the so-called "geometric scaling length" [2.2]. For small Λ, the second term on the left hand side of eqn. (2.32) dominates, and (2.33) becomes

$$\frac{(V'_{GS} - \psi_S)}{\Lambda^2} = \frac{qN_A}{\varepsilon_{Si}} \tag{2.34a}$$

or

$$V'_{GS} = \psi_S + qN_A \Lambda^2 / \varepsilon_{Si} \ . \tag{2.34b}$$

This is the one-dimensional case, where $\partial^2\psi/\partial y^2$ dominates and $\partial^2\psi/\partial x^2$ can be ignored. For the 1D MOS capacitor, we already found that

$$V'_{GS} = \psi_S - Q_B / C_{ox} = \psi_S + qN_A W_D / C_{ox}, \tag{2.34c}$$

so to make eqn. (2.34b) consistent with 1D MOS theory, we must have

$$\Lambda = \sqrt{\frac{\varepsilon_{Si}}{\varepsilon_{ox}} W_D t_{ox}} \ . \tag{2.35}$$

Having specified Λ, we can use eqn. (2.33) in eqn. (2.32) to find

$$\frac{\partial^2 \psi_S}{\partial x^2} - \frac{(\psi_S - V'_{GS})}{\Lambda^2} = \frac{qN_A}{\varepsilon_{Si}} \ . \tag{2.36}$$

If we define

$$\phi = \psi_S - V'_{GS} + \frac{qN_A \Lambda^2}{\varepsilon_{Si}} \ , \tag{2.37}$$

then eqn. (2.36) becomes

$$\frac{d^2\phi}{dx^2} - \frac{\phi}{\Lambda^2} = 0, \tag{2.38}$$

which can be solved subject to the boundary conditions

$$\phi(0) = \phi_S \tag{2.39a}$$

$$\phi(L) = \phi_D \tag{2.39b}$$

to find

$$\phi(x) = \mathrm{A}e^{-x/\Lambda} + \mathrm{B}e^{x/\Lambda}. \tag{2.40}$$

The final solution,

$$\psi_S(x) = (V'_{GS} - qN_A\Lambda^2/\varepsilon_{Si}) + \phi(x) \tag{2.41}$$

is sketched in Fig. 2.9. The first term on the RHS of eq. (2.41) describes the effect of the gate potential; it tries to hold ψ_S constant at a value determined by V_{GS}. The second term describes the lowering of $E_C(x)$ due to the drain and source potentials, and, if the channel length is too short compared to Λ can lead to so-called drain induced barrier lowering (DIBL), as shown in Fig. 2.9.

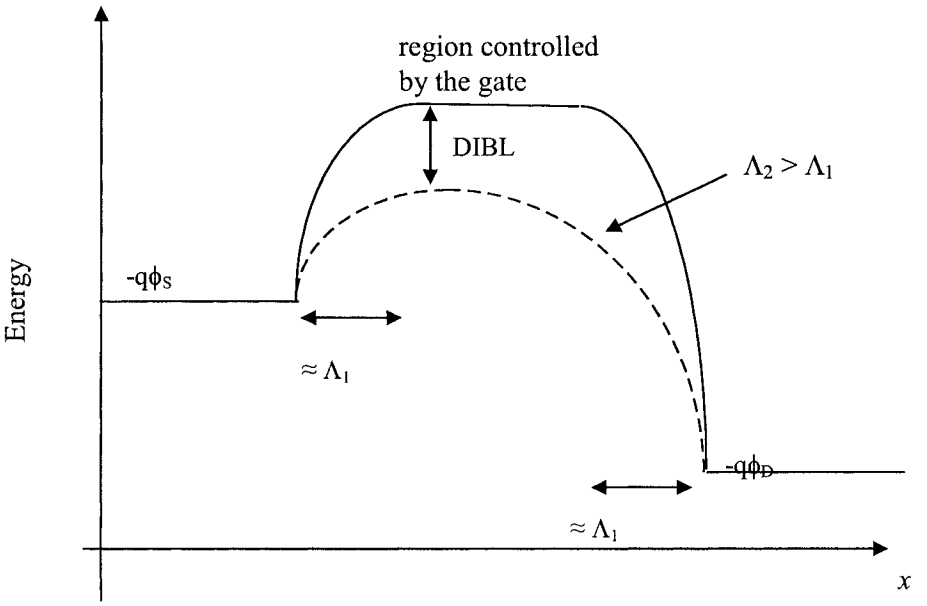


Fig. 2.9 Sketch of the electron energy vs. position showing the role of the geometric scaling length, Λ. For a given L, a short Λ (Λ_1 above) results in a region of the channel that is strongly controlled by the gate, but for a longer Λ (Λ_2 above) DIBL occurs and the gate does not have total control over the potential barrier.

To produce a portion of the channel where ψ_S is controlled by V_{GS} and is independent of V_{DS}, we require

$$L >> \Lambda, \tag{2.42}$$

which is the criterion for an electrostatically well-designed MOSFET. An electrostatically well-designed MOSFET is one for which the channel is controlled by the gate voltage, not by the drain to source voltage. A small Λ permits the smallest L while maintaining acceptable DIBL and S.

Equation (2.35) shows that the oxide and silicon body thickness should be small for good gate control. Similar considerations apply to a single gate SOI MOSFET, with W_D replaced by the thickness of the silicon body, T_{Si}. Double gate MOSFETs have a smaller Λ [2.9] and the cylindrical gate MOSFET, an even smaller one [2.10]. A careful derivation based on a series solution to Poisson's equation gives Λ from a transcendental equation [2.10]. Figure 2.10 shows why Λ is called a geometric scaling length; it depends on the geometry of the MOSFET. In a bulk MOSFET, field lines from the drain can reach through and lower the barrier near the source. By surrounding the channel with the gate, the field lines from the drain terminate on the gate, which screens the drain potential so that it does not influence the channel potential near the source.

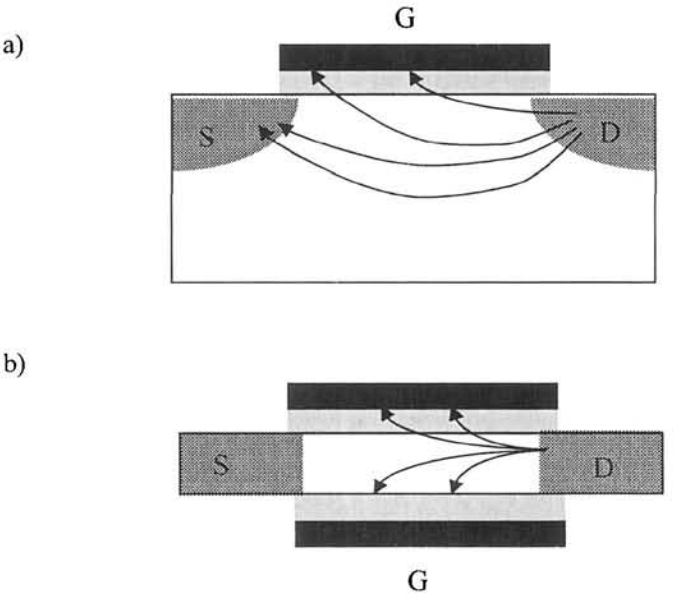


Fig. 2.10 Illustration of the drain electric field lines in a) a bulk and b) a double gate MOSFET.

A Capacitor Model for 2D Electrostatics

The MOSFET's gate electrode induces charge in the channel of the transistor, but the source and drain electrodes can also induce charge. As sketched in Fig. 2.11a, our interest is in the potential and charge at the top of the barrier (which defines the beginning of the channel). The three-capacitor model of Fig. 2.11b describes the modulation of the potential and charge at the top of the barrier by the three terminals. This simple model provides an alternative, very physical, picture of 2D MOSFET electrostatics.

Before using the three capacitor model, recall the 1D results of eqn. (2.12a), which relates the charge and surface potential to the gate voltage. Solving for the surface potential, we find

$$\psi_S = V'_{GS} + \frac{Q}{C_{ox}}. \tag{2.43}$$

When the charge, Q, is zero, the Laplace solution to the electrostatics shows that the surface potential is simply the gate voltage. When the charge is present and the gate voltage is zero, the surface potential is the charging potential of Q/C_{ox}.

Returning to the three capacitor model, we can solve the problem shown in Fig. 2.11b by superposition. First assume that the charge at the top of the barrier is zero, and solve the Laplace problem for the potential at the top of the barrier, then set the voltages to zero and compute the charging potential. The result is

$$\psi_S = \left(\frac{C_G}{C_\Sigma} V_G + \frac{C_D}{C_\Sigma} V_D + \frac{C_S}{C_\Sigma} V_S \right) + \frac{Q}{C_\Sigma}, \tag{2.44}$$

where

$$C_\Sigma = C_G + C_D + C_S. \tag{2.45}$$

(Note that C_G is what we usually call C_{ox} for a MOSFET.)

A well-designed transistor is one for which the gate capacitance dominates so that the source and drain voltages have little effect on the surface potential at the top of the barrier. Equation (2.44) explains the drain-induced barrier lowering (DIBL) sketched in Fig. 2.3c and the classic short channel effects of a MOSFET as sketched in Fig. 2.3d. As the drain voltage increases, the source to channel barrier height decreases, which increases the subthreshold current (DIBL). As the channel length decreases, the drain capacitance increases, which lowers the barrier. A lower gate voltage,

therefore, is needed to reduce the barrier height to the value given by eqn. (2.31), where the MOSFET turns on. The result is a decrease of V_T with decreasing channel length (threshold voltage roll-off).

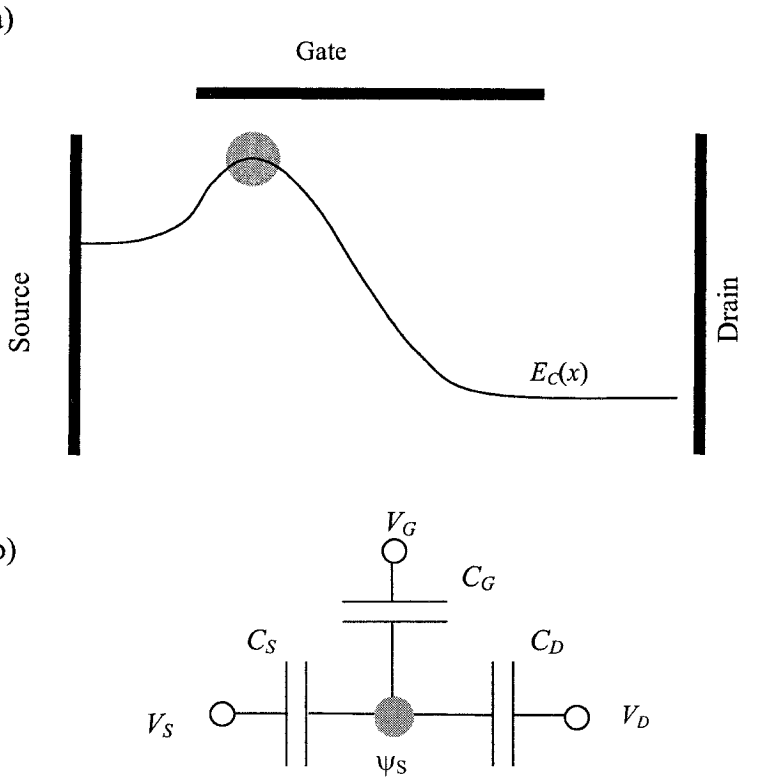


Fig. 2.11 Illustration of the capacitor model of 2D electrostatics. (a) The conduction band profile under high gate and drain bias with the inversion layer charge at the top of the barrier identified. (b) A three-capacitor model that describes the control of the potential at the top of the barrier.

2.5 MOSFET Current vs. Voltage Characteristics

Having understood the $Q_i(V_{GS},V_{DS})$ characteristic, it is relatively easy to establish the essential features of the MOSFET $I_D(V_{GS},V_{DS})$ characteristic. As shown in Fig. 2.12, for low V_{DS}, the MOSFET behaves like a resistor while for high V_{DS}, it behaves more like a current source. To minimize the

mathematics, we will develop an expression for the linear (low V_{DS}) and saturated (high V_{DS}) regions only.

The drain current is the product of charge times velocity,

$$I_D = -W\,Q_i(x)\langle \upsilon(x)\rangle. \tag{2.46a}$$

Because current is continuous, we choose to evaluate eqn. (2.46a) at $x = 0$, where we know $Q_i(x=0)$ from eqn. (2.16). The result is

$$I_D = -W\,Q_i(0)\langle \upsilon(0)\rangle \tag{2.46b}$$

or

$$I_D = W\,C_{GS}\left(V_{GS} - V_T\right)\langle \upsilon(0)\rangle. \tag{2.46c}$$

(Recall that the gate-source capacitance, C_{GS}, is less than C_{ox}, because it is in series with the semiconductor capacitance, C_S.)

In the linear region of the $I_D - V_{DS}$ characteristic, the potential drop in the channel is linear, the electric field approximately constant (recall Fig. 2.8c), so

$$I_D = W\,C_{ox}\left(V_{GS} - V_T\right)\mu_{eff}\mathrm{E}_x, \tag{2.47}$$

where μ_{eff} is the effective mobility of the inversion layer electrons and E_x is the electric field in the channel. In the linear region,

$$\mathrm{E}_x \approx \frac{V_{DS}}{L}, \tag{2.48}$$

so

$$I_D = \frac{W}{L}\mu_{eff}C_{ox}\left(V_{GS} - V_T\right)V_{DS} = \frac{V_{DS}}{R_{ch}}. \tag{2.49}$$

The channel resistance,

$$R_{ch} = \frac{1}{\mu_{eff}C_{ox}\left(V_{GS} - V_T\right)}\frac{L}{W}, \tag{2.50}$$

is proportional to the length of the channel, as expected.

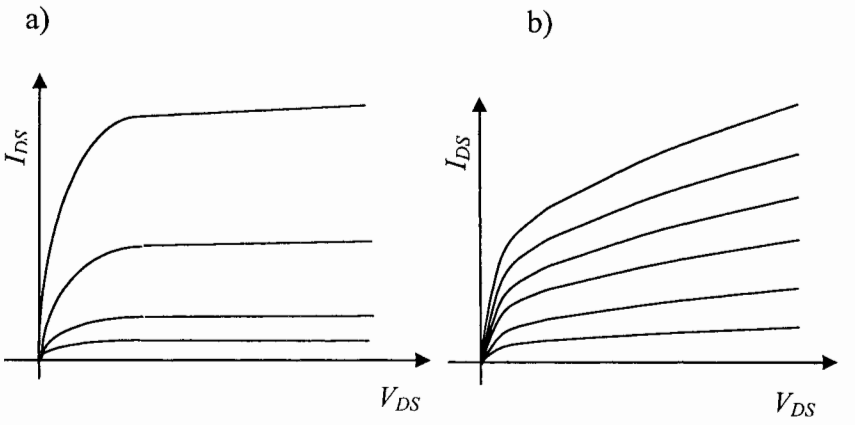


Fig. 2.12 MOSFET I_D vs. V_{DS} characteristic a) long channel (on-current varies as the square of the gate voltage) and b) short channel (on-current varies linearly with the gate voltage).

Saturation Region: Long Channel:

When $V_{DS} > 0$, the potential along the channel varies with position. The simplest way to treat this two-dimensional problem is to modify the one-dimensional result, eqn. (2.16), to

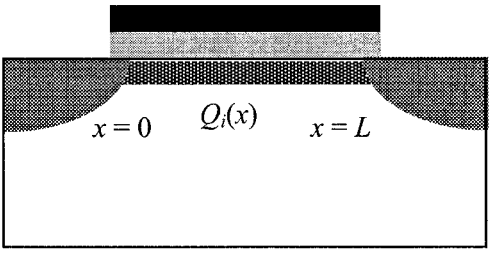
$$Q_i(x) = -C_{ox}\left(V_{GS} - V_T - V(x)\right), \tag{2.51}$$

where $V_S < V(x) < V_D$ is the channel potential. Equation (2.51) is the well known "gradual channel approximation" of MOS theory [2.7, 2.8]. Under low drain bias, $V(x)$ is small, and the inversion layer is uniform as sketched in Fig. 2.13a. When the drain bias increases, however, the potential difference between the gate and substrate is reduced near the drain, so the inversion layer density decreases. When

$$V(x = L) = V_{Dsat} = \left(V_{GS} - V_T\right), \tag{2.52}$$

then eqn. (2.51) predicts that $Q_i(x = L) = 0$. As shown in Fig. 2.13b, the channel is said to be "pinched-off" at $x = L$. (There is, of course, a small, positive Q_i to carry the drain current. The value is determined by the velocity of carriers in the pinched-off region.)

(a)



(b)

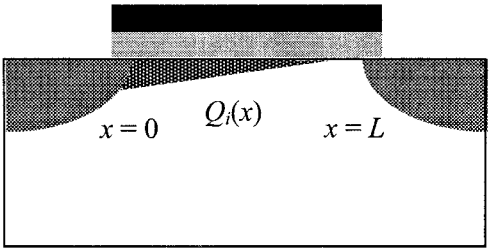


Fig. 2.13 Illustration of the channel for $V_{GS} > V_T$ and (a) low and (b) high V_{DS} conditions.

When $V_{DS} > V_{Dsat}$, the potential drop across the inverted portion of the channel is $\approx (V_{GS} - V_T)$, so we can estimate the electric field as

$$\mathrm{E}_x(0) \approx \frac{(V_{GS} - V_T)}{2L} \tag{2.53}$$

(the factor of 2 comes from a proper treatment of the non-uniform electric field within the channel). Finally, from eqns. (2.47) and (2.53), we find the saturation current as

$$I_{Dsat} = \frac{W}{2L} \mu_{eff} C_{ox} (V_{GS} - V_T)^2 . \tag{2.54}$$

The square law behavior of the long channel MOSFET was sketched in Fig. 2.12a.

Saturation Region: Short Channel

For long channel MOSFETs we can assume that $\langle \upsilon \rangle = \mu E_x$, but at high electric fields, transport becomes nonlinear as was sketched in Fig. 1.11. When the field is higher than about 10^4 V/cm, the velocity of electrons in silicon saturates at about 10^7 cm/s. Such fields occur for nanoscale MOSFETs with $L \approx$ 100nm and $V_{DS} \approx$ 1.0V. For a short channel MOSFET, therefore, eqn. (2.46c) gives the saturated drain current as

$$I_D = W C_{ox} \upsilon_{sat} (V_{GS} - V_T). \tag{2.55}$$

Figure 2.12 compared a long channel MOSFET for which $I_D \propto (V_{GS} - V_T)^2$ to a short channel, velocity saturated MOSFET for which $I_D \propto (V_{GS} - V_T)$. In practice, $I_D \propto (V_{GS} - V_T)^\alpha$, where $1 < \alpha < 2$.

Finally, we should mention that transport across a short, high-field region (the channel of a nanoscale MOSFET under high bias) is one of the more difficult problems in transport theory. Figure 1.12 sketched the steady state $\langle \upsilon \rangle$ vs. x profile for a high field step. The velocity saturates at $\approx 10^7$ cm/s, but initially it overshoots. The near-equilibrium carriers injected into the high-field region initially have high mobility. As they gain energy from the field, they scatter more, and the mobility decreases. Velocity overshoot occurs because the mobility and electric field are both high near the beginning of the step. The spatial extent of velocity overshoot is roughly 100nm, so for a nanoscale MOSFET, velocity overshoot may occur throughout the entire channel.

Subthreshold Conduction:

Under subthreshold conditions, the source to channel barrier is large, Q_i is small, and the electric field in the channel is also small. If diffusion dominates,

$$\langle \upsilon(0) \rangle = D_{eff} / L, \tag{2.56}$$

so, using eqn. (2.29) for $Q_i(V_{GS})$, eqn. (2.46b) gives

$$I_D = \frac{W}{L}(m-1)\mu_{eff} C_{ox} \left(\frac{k_B T_L}{q} \right)^2 e^{q(V_{GS}-V_T)/mk_B T_L}. \tag{2.57}$$

Figure 2.14 sketches the $I_D(V_{GS})$ characteristic below and above V_T. In the subthreshold region, I_D varies exponentially with $(V_{GS} - V_T)$, but above

threshold, it varies as $(V_{GS}-V_T)^{\alpha}$. The subthreshold swing is readily evaluated from eqn. (2.57) to give

$$S = 2.3m\frac{k_B T_L}{q} \qquad \text{(V/decade)} \tag{2.58}$$

Since $m > 1$ (recall that $m = 1 + C_D/C_{ox}$), we find that S > 60 mV/decade. Finally, note that

$$I_{off} = I_D(V_{GS}=0, V_{DS}=V_{DD}) \propto e^{-qV_T/mk_BT_L} \tag{2.59a}$$

and

$$I_{on} = I_D(V_{GS}=V_{DS}=V_{DD}) \propto (V_{DD}-V_T)^{\alpha} \tag{2.59b}$$

so

$$I_{off} \propto e^{\beta I_{on}^{1/\alpha}}, \tag{2.59c}$$

where β is a constant. The result is that if we lower V_T to get a little more on-current, we get exponentially more off-current. In fact, technologies are often characterized by a $\log_{10} I_{off}$ vs. I_{on} plot.

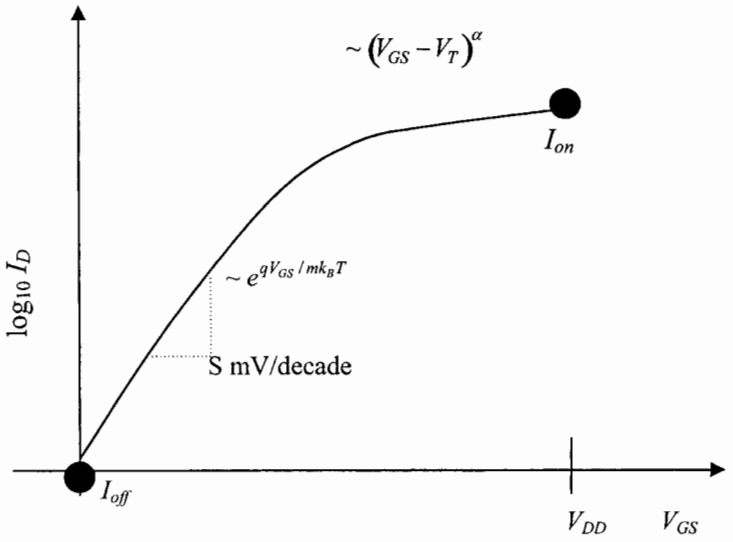


Fig. 2.14 The $\log_{10} I_D$ vs. V_{GS} characteristic illustrating the important parameters below and above threshold.

2.6 The Bipolar Transistor

Figure 2.15 compares the idealized structures and energy band diagrams for MOS and bipolar transistors. For the bipolar transistor, the base-emitter voltage lowers the emitter-base energy barrier, so that $n(0)$ electrons are injected into the base. The injected carriers diffuse across the base and are collected by the collector. Since the density of electrons is low at the base-collector junction, we find

$$J_C = qD_n \frac{dn}{dx} = qn(0)\frac{D_n}{W_B} = \left[q\frac{n_i^2}{N_A}\frac{D_n}{W_B} \right] e^{qV_{BE}/k_B T_L}, \tag{2.60}$$

where the second expression comes from the well-known "Law of the Junction" for $n(0)$ [2.8]. The collector current follows directly to write

$$I_C = qA_E \frac{n_i^2}{N_A} \langle \upsilon(0) \rangle e^{qV_{BE}/k_B T_L}, \tag{2.61}$$

where $\langle \upsilon(0) \rangle$ is the diffusion velocity and A_E the emitter area. Equation (2.61) describes the bipolar transistor in the normal, active mode of operation where the emitter-base junction is forward-biased and the base-collector junction reverse biased.

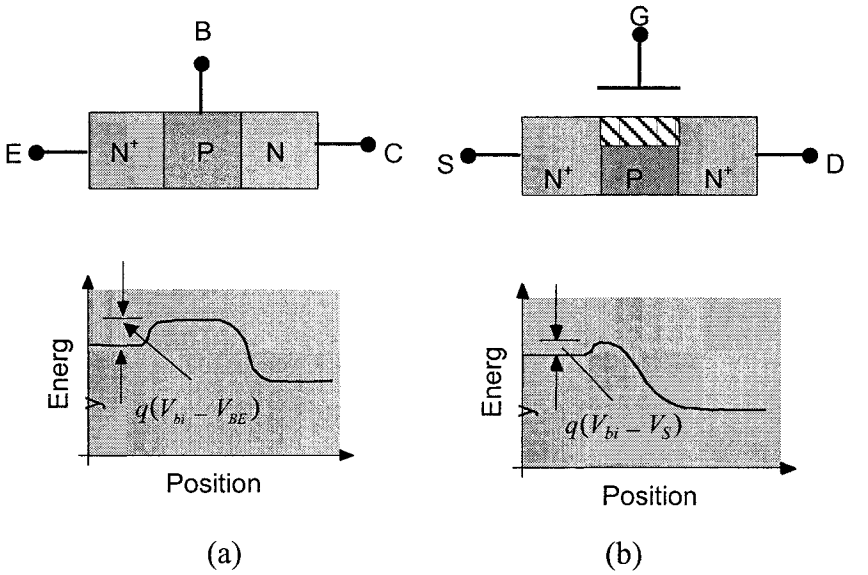


Fig. 2.15 Idealized device structures and energy band diagrams for: (a) bipolar transistor and (b) a MOSFET.

Because of the similarity of the bipolar and MOS transistor energy band diagrams as displayed in Fig. 2.15, we should expect that they operate similarly. Let's see if we can derive the MOSFET characteristic from eqn. (2.61), which describes the bipolar device.

By recognizing that ψ_S plays the role of V_{BE}, eqn. (2.61) becomes

$$I_C = qWt_{inv}\frac{n_i^2}{N_A}\langle \upsilon(0)\rangle e^{q\psi_S/k_BT_L}, \tag{2.62}$$

where Wt_{inv} is the cross-sectional area for current flow. Recall that

$$n(0) = \left(\frac{n_i^2}{N_A}\right)e^{q\psi_S/k_BT_L} \tag{2.63}$$

and that $n_S = t_{inv}n(0)$, so

$$I_D = WQ_i(0)\langle \upsilon(0)\rangle, \tag{2.64}$$

which is precisely eqn. (2.46b) from which we derived the MOSFET $I_D(V_{GS},V_{DS})$ characteristic.

One often hears the statement "below threshold, a MOSFET operates like a bipolar transistor," by which it is meant that the current varies exponentially with the input voltage. This is easy to see by using $\psi_S = V'_{GS}/m$ from eqn. (2.26) in eqn. (2.62) to obtain the subthreshold characteristic, eqn. (2.57). It is not as well known that above threshold, the MOSFET still operates like a bipolar transistor [2.11].

According to eqn. (2.11), in strong inversion,

$$\psi_S = \frac{2k_BT_L}{q}\ln(Q_i) \approx \frac{2k_BT_L}{q}\ln\left[C_{ox}(V_{GS}-V_T)\right]. \tag{2.65}$$

The drain current of a MOSFET varies exponentially with ψ_S both above ***and*** below threshold, but above threshold, ψ_S varies logarithmically with $(V_{GS} - V_T)$ so the result is that I_D varies linearly with $(V_{GS} - V_T)$ above threshold. The reduced control of the gate occurs because the source to channel barrier is modulated indirectly by V_G. Above threshold, it is difficult to modulate ψ_S by the gate voltage because the inversion layer charge is strong and it screens out the charge on the gate.

2.7 CMOS Technology

The important device performance metrics are derived from the requirements of CMOS circuits. The basic element of a CMOS digital system is the inverter shown in Fig. 2.16. It consists of two normally off (so-called enhancement mode) MOSFETs in series. The one on the top is a p-channel MOSFET ($V_T < 0$) and the one on the bottom is an n-channel MOSFET ($V_T > 0$). The PMOS transistor is referred to as the "pull-up" transistor, because when it is on, it pulls the output up to the power supply voltage. Similarly, the NMOS transistor is referred to as the "pull down" transistor, because when it is on, it pulls the output voltage down to ground. If MOSFETs were ideal switches that closed when the gate voltage exceeded V_T (or is less than V_T for the PMOS) then the transfer characteristic (the output voltage vs. input voltage) would be the dashed line in Fig. 2.16b. When the input voltage is low, the PMOS transistor turns on (NMOS off) and connects the output node to V_{DD}; when the input voltage is high, the NMOS transistor turns on (PMOS off) and connects the output node to ground. If a low voltage represents a logical 0 and a high voltage a logical 1, then this circuit operates as a digital inverter. More complicated logical functions can be realized by placing transistors in series and parallel.

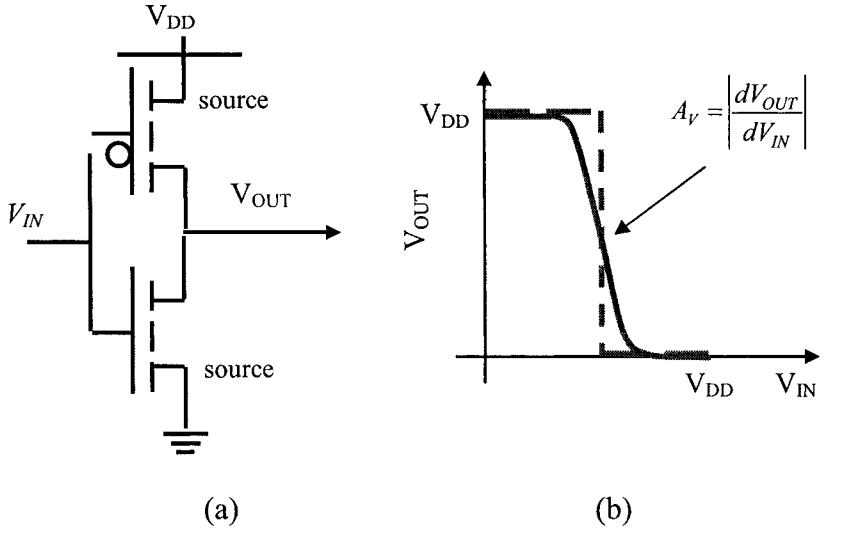


Fig. 2.16 A CMOS inverter. (a) circuit schematic and (b) the transfer characteristic. (The dashed line is the transfer characteristic if the CMOS transistors were ideal switches.)

The solid line in Fig. 2.16b is the transfer characteristic of a well-designed inverter. The switching point occurs at $V_{DD}/2$, which is achieved by matching the currents and threshold voltages of the two transistors. (PMOS transistors have lower mobility than NMOS transistors, so the PMOS transistor is typically 2-3 times wider than the NMOS transistor.) The slope of the transfer characteristic at the switching point is the voltage gain, A_V. (By biasing a transistor at this point, the inverter can be used as an analog amplifier.) Note that away from the switching point, the output voltage is insensitive to the input voltage, which provides the inverter with a *noise margin*. Even if a logical zero is not exactly 0 volts (because of noise on the input line), the output voltage will be restored to a logical 1. Similarly, a logical 1 may be less than V_{DD} volts, and the output voltage is restored exactly to a logical 0. Noise margins are what make digital systems possible, otherwise noise would degrade the signals after propagating through a few stages. In order to have noise margins at the low and high end (that is regions of the transfer characteristic that are flat at low and high input voltages), we require

$$|A_V| > 1. \tag{2.66}$$

Gain provides signal restoration, which is essential for any device to be used in a digital system.

Figure 2.17 shows the pull down portion of a CMOS gate. The capacitor, C, represents the capacitance of the gates and interconnects that are connected to the output node. During the pull-up phase, the capacitor is charged to V_{DD} – a logic one. A clock drives the gate, and when the clock voltage is high, the NMOS pull-down transistor turns on and discharges the capacitor. The power dissipation is

$$P = \frac{1/2\, C V_{DD}^2}{T_{CL}/2} = f C V_{DD}^2, \tag{2.67}$$

where T_{CL} is the period of the clock. As the number of transistors per chip and clock frequencies continue to increase, power management is becoming a crucial issue. Equation (2.67) explains why the supply voltage must be reduced with each technology generation.

Equation (2.67) describes the dynamic (or switching) power of the gate. The key advantage of CMOS, which led to its adoption over NMOS, was the fact that essentially no power was dissipated unless a gate switched. But this is changing as leakage currents are increasing. Consider the circuit of Fig.

2.17 when the NMOS pull down is off and C is charged. The static power dissipated is

$$P_S = I_{off} V_{DD}. \tag{2.68}$$

To minimize the static power, a low I_{off} (high V_T) is required, but a high V_T reduces the on-current and the speed suffers as discussed next.

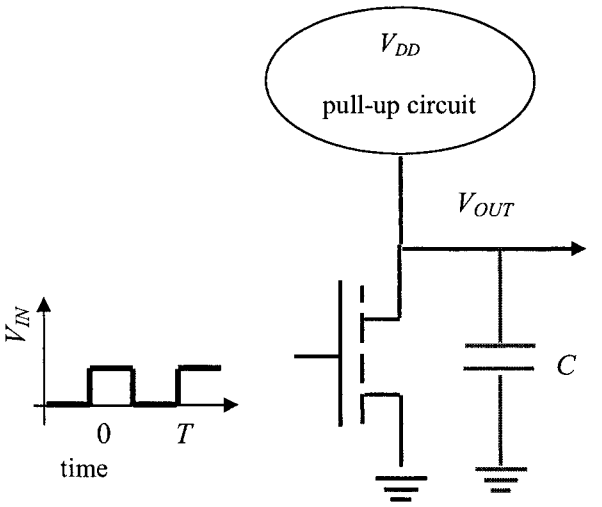


Fig. 2.17 Illustration of the pull-down portion of a CMOS gate.

The gate switching delay determines the maximum clock frequency. If we measure the gate delay by the time, τ, to remove the capacitor's charge, we find

$$\tau = \frac{CV_{DD}}{I_{on}}, \tag{2.69}$$

which explains why a high on-current is important. From eqn. (2.46c), with we can evaluate the device delay metric as

$$\tau = \frac{C_{GS} W L V_{DD}}{W C_{GS} (V_{GS} - V_T) \langle \upsilon(0) \rangle} \approx \frac{L}{\langle \upsilon(0) \rangle} \tag{2.70}$$

so the device delay metric is closely related, but not identical to, the transit time of carriers across the channel. The delay metric for current-day technology (the 90nm technology node) is about 1 ps. This represents the intrinsic switching speed of a transistor and would correspond to a clock frequency of several hundred GHz. Integrated circuits operate much slower because of the need to charge and discharge large capacitances because the output node is connected to several gates (fan-out) and the interconnecting wires also add capacitance.

A typical integrated circuit contains many layers of wiring to interconnect the circuit, and the need to charge and discharge interconnects limits speed and increases power. Figure 2.18 is a cross sectional sketch of an interconnect. The wire is characterized by a capacitance per unit length and a resistance per unit length and can be viewed as a distributed RC transmission line [2.7]. The mathematics of signal propagation on an RC transmission line is identical to the mathematics of particles diffusing in a semiconductor. The "diffusion coefficient" for the RC transmission line is

$$D_{RC} = 1/R_w C_w, \tag{2.71a}$$

where R_w is the resistance per unit length and C_w the capacitance per unit length of the wire. Recall that the delay time for minority carriers diffusing across the base of a bipolar transistor is $W^2/2D$, where W is the width of the base. By analogy, therefore, the time for a signal to propagate across an RC transmission line is

$$\tau_{RC} = 0.5 R_w C_w L_w^2, \tag{2.71b}$$

where L_w is the length of the interconnect. The resistance per unit length depends on the resistivity of the metal interconnect, which explains why copper metallization has replaced aluminum metallization. The capacitance per unit length depends on the geometry of the interconnect and the dielectric constant of the medium, which explains why so-called low-κ insulators are replacing SiO_2 in the interconnect layers. The critical point, however, is that the delay depends on the *square* of the length of the interconnect. (Note, however, that for short interconnects, eqn. (2.71b) could imply that signals

propagate faster than the speed of light). For such cases, a more physical transmission line model must be used.

An integrated circuit contains a distribution of interconnect lengths, most of which are short, local interconnects. A few long interconnects, however, are essential, and these long interconnects limit the speed of the system. Originally, device speed determined the speed of the circuit, but now interconnect delays dominate, and a central task of the chip designer is to manage these delays.

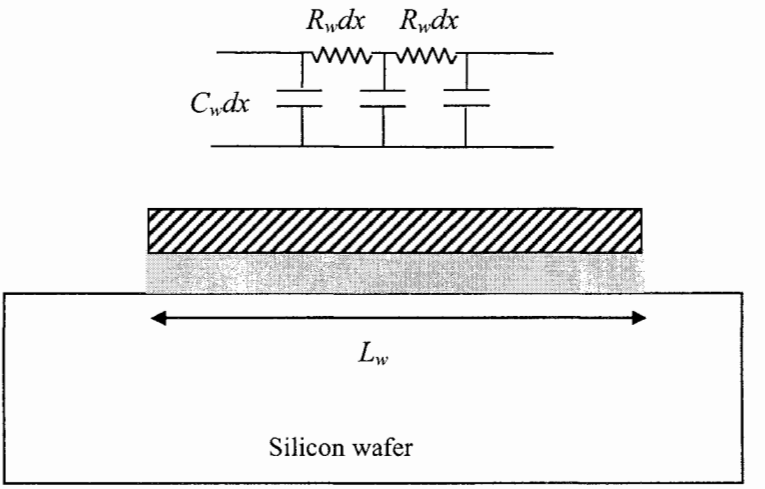


Fig. 2.18 Cross-sectional sketch of an interconnect wire showing how it is modeled as a distributed RC transmission line.

MOSFET Scaling:

The objective of device scaling is to shrink transistor dimensions so that more transistors can be placed on a chip. Typically, the scaling factor, k, is $\approx\sqrt{2}$, so that the area of a transistor shrinks by one-half and the number per area increases by a factor of two. The challenge in scaling is to maintain suitable electrical characteristics.

Consider scaling the channel length,

$$L \rightarrow L/k. \tag{2.72}$$

To control two-dimensional effects such as DIBL and V_T roll-off, we require that $L >> \Lambda$, where Λ is the geometric scaling length discussed in Sec. 2.4. In practice

$$L > 1.5\Lambda \tag{2.73}$$

provides acceptable control of two-dimensional electrostatics. It is necessary, therefore, that we also scale Λ,

$$\Lambda \rightarrow \Lambda / k, \tag{2.74}$$

which, according to eqns. (2.35) can be accomplished by reducing t_{ox} and W_D (by increasing the channel doping). The source/drain junction depth, an effect not included in eqns. (2.35) should also be reduced.

Because scaling increases the number of transistors per chip, the power dissipation per chip increases unacceptably unless the power supply voltage is also scaled,

$$V_{DD} \rightarrow V_{DD} / k. \tag{2.75}$$

The on-current per unit width, however, must be maintained so that circuit speed does not suffer. Therefore,

$$I_{ON} / W \rightarrow I_{ON} / W. \tag{2.76}$$

Because V_{DD} is reduced, we must also reduce V_T to maintain on-current,

$$V_T \rightarrow V_T / k, \tag{2.77}$$

but eqn. (2.59a) shows that the off-current, and therefore the standby power, increases exponentially as V_T is scaled down. The on-current/off-current trade-off is an increasingly difficult challenge to manage. Note also that as V_T is scaled down, variations in V_T increase for small devices, so device-to-device variations are becoming an important issue.

It is also interesting to see how channel resistance of the MOSFET scales. We find,

$$R_{ch} \equiv \frac{V_{DD}}{I_{ON}} \rightarrow R_{ch} / k. \tag{2.78}$$

The intrinsic resistance of the device is scaling down, but the parasitic resistances, which depend on metal-semiconductor contact resistance and the junction depth and doping, are increasing. The result is that device performance will be increasingly degraded by series resistance as channel lengths push into the nanoscale regime.

Moore's Law states that the number of transistors per chip doubles each technology generation [2.12]. (Currently, a technology generation is about two years). The doubling of transistor density is a result of three factors: 1) device scaling, 2) improvements in layout, which increase transistor packing density, and 3) increased die (chip) size. Because the scaling factor, k, is somewhat greater than $1/\sqrt{2}$, factors 2) and 3) play an important role in Moore's Law. The International Technology Roadmap for Semiconductors is a statement of the technology characteristics needed to maintain Moore's Law in the future [2.13]. Current technologies have channel lengths below 100nm, and if scaling continues for another decade or so, channel lengths will be less than 10nm. Designers are increasingly challenged by off-state leakage (from the source to drain and through the ultra-thin gate insulator), low on-currents, increasing variation of device parameters across a chip, interconnect delays, and power dissipation.

2.8 Ultimate Limits

We conclude this chapter with a look at the ultimate limits for transistor-based digital computation. In particular, we seek to establish the:

1) Minimum energy dissipation per logic transition, E_{min} (J)
2) Minimum device size, L, in nm (or maximum device density, n_S, per cm^2)
3) Minimum device delay, t_{min} (ps)
4) Power dissipation, P (W/cm^2).

The topic of dissipation in computing is a deep problem [2.14], [2.15] with a rich history [2.16]. Some of these issues are still being debated. Our goal in this section is modest; we seek to establish the ultimate limits for any transistor that operates in a conventional circuit by modulating the flow of current across an electrostatic barrier. We follow an approach that is similar to [2.17] and [2.18].

To begin, let's consider the switching energy, the energy dissipated when we convert a logical one to a logical zero,

$$E_S = \frac{1}{2}CV_{DD}^2 = \frac{1}{2}QV_{DD} = \frac{q}{2}NV_{DD}, \tag{2.79}$$

where Q is the charge stored on the capacitor, and N is the number of electrons stored. The minimum switching energy occurs when $N = 1$, so we need to establish the minimum power supply voltage, V_{min}.

According to eqn. (2.66), a CMOS inverter must have a gain greater than one in order to have a noise margin. We compute the gain by equating I_D(NMOS) to I_D(PMOS), solving for V_{OUT}, and differentiating it with respect to V_{IN}. If we assume subthreshold conduction, ideal devices ($m = 1$ in eqn. (2.57)), and symmetrical N and P-channel devices, then we find that [?]

$$V_{\min} = 2\ln(2)k_B T_L / q. \tag{2.80}$$

When eqn. (2.80) is inserted into eqn. (2.79), we find that the minimum energy per logical transition is [2.17]

$$E_S = \ln(2)k_B T_L \quad (T_L = 300\text{K}). \tag{2.81}$$

Equation (2.81) is a specific example of a much more general result. Whenever a logical transition occurs and the bit is erased, then there is an inevitable energy dissipation that must be at least as great as the result of eqn. (2.81) [2.14]. If the capacitors are charged and discharged adiabatically, however, or if information is preserved (so-called reversible computing), then the energy dissipation per logic transition can be arbitrarily small [2.15]. The interested reader is referred to the literature for these topics. Our focus is on the limits of conventional CMOS logic.

Having addressed question 1), we now turn to questions 2) – 4) and use an approach that is similar to that of Zhirnov, *et al.* [2.18]). Consider the "transistor" sketched in Fig. 2.19, which consists of two thermal equilibrium reservoirs (source and drain) separated by an energy barrier whose height is modulated by a gate voltage. We assert that the results of this analysis apply in general to any transistor that controls a current by modulating a potential energy barrier. Since our focus is on limits, we assume a ballistic transistor. No scattering occurs in the channel – only in the source/drain regions where strong scattering maintains thermal equilibrium.

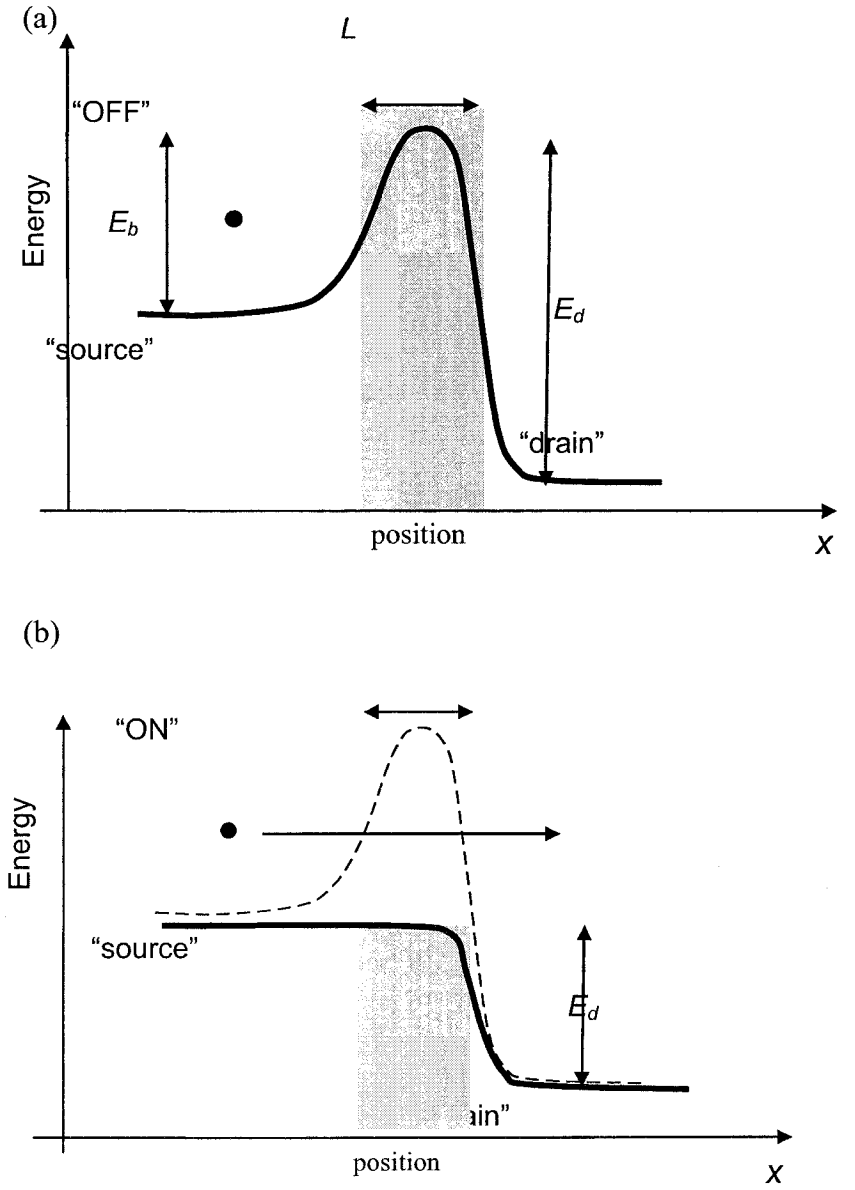


Fig. 2.19 Illustration of a hypothetical, digital electronic device. (a) the off -state and (b) the on-state.

Consider the switching energy for the model transistor. In the off-state, an electron injected from the source must have less than a 50:50 chance of propagating to the drain, and in the on-state, an electron injected into the drain dissipates its energy in the drain and thermalizes. There must be a minimum barrier (drain voltage) so that a thermal equilibrium electron in the drain has less than a 50:50 chance of returning to the source (this assures us that we can distinguish the two states). We conclude that

$$e^{-E_d/k_BT_L} < \frac{1}{2}, \tag{2.82}$$

which leads to the same minimum switching energy as eqn. (2.81).

Next, we address the question of the minimum size of a transistor. In order for the off-state to be distinguishable from the on-state, we require that the tunneling probability through the barrier be less than 0.5. Using a WKB approximation for the tunneling probability, we find

$$P(\text{WKB}) = \exp\left(-\frac{2\sqrt{2mE}}{\mathsf{h}}L\right) < \frac{1}{2}. \tag{2.83}$$

If we assume a thermal equilibrium electron in the source, $E = k_BT = E_{\min}/\ln(2)$, then eqn. (2.83) leads to

$$L > \frac{[\ln(2)]^{3/2}}{2}\left(\frac{\hbar}{\sqrt{2mE_S}}\right) \approx \left(\frac{\hbar}{\sqrt{2mE_S}}\right) = 1.5 \text{ nm } (T_L = 300\text{K}). \tag{2.84}$$

The last form is what we could have obtained directly from the Uncertainty Principle, $\Delta x\,\Delta p \geq \hbar$. According to eqn. (2.84), the minimum size of a transistor is about 1 nm at room temperature. MOSFETs with gate lengths only a few times larger than this have already been demonstrated [2.19]. (Note, however, that the minimum size of the overall MOSFET is typically 10-15 times larger than the length of the gate.)

Having determined the minimum size of a device, we can determine the maximum density of devices as

$$n_{\max} = \frac{1}{L_{\min}^2} \approx 4.7\times10^{13} \quad \text{devices / cm}^2, \tag{2.85}$$

which is an enormous number – four to five orders of magnitude larger than present day CMOS. Two orders of magnitude come from the fact that a device is at least 10 times larger than its minimum feature. But more importantly, as we shall see later, the density of devices is not limited by our ability to make small devices; it is limited by our ability to dissipate the power generated by the devices.

Consider next the switching speed of the transistor, which is the transit time across the control region,

$$t_S = \frac{L}{\upsilon} = \frac{L}{\sqrt{2E/m}}. \tag{2.86}$$

Using $L = L_{\min}$ and $E = k_B T_L = E_{\min}/\ln(2)$, we find

$$t_{\min} = \left(\frac{\ln(2)}{2}\right)^{5/2} \frac{\hbar}{E_{\min}} \approx \frac{\hbar}{E_{\min}} = 40 \text{ fs}, \tag{2.87}$$

which is only a few times smaller than CMOS transistors are expected to achieve.

Finally, let us estimate the power dissipation per cm^2 for a chip operating at the density and speed limits. We have

$$P = \frac{\alpha n_S E_S}{t_S}, \tag{2.88}$$

where α is the switching activity factor (the average fraction of clock cycles that an average transistor switches). If we assume that α = 1 and use the minimum switching energy and maximum switching speed we find

$$P_{\max} = \frac{n_{\max} E_{\min}}{t_{\min}} = 3.7 \times 10^6 \text{ W/cm}^2. \tag{2.89}$$

Equation (2.89) is almost three orders of magnitude higher than the energy flux at the surface of the sun! Silicon technologists hope to be able to develop affordable heat sinking techniques to dissipate 100 W/cm^2, but that is more than four orders of magnitude smaller than the density-limited power dissipation given by eqn. (2.89). Power dissipation is already a critical issue for designers. Present day technology operates at four to five orders of magnitude above $k_B T \ln(2)$. Operating speeds are well below fundamental limits, but the switching energy is much larger than $k_B T \ln(2)$. Operation

at relatively a relatively high voltage (~ 1V) is necessary to minimize errors due to spontaneous thermal transitions, to accommodate device-to-device variations, and to provide sufficient speed. Our ability to remove thermal energy from the chip, not our ability to make devices small, is what limits the device density of a chip. It is likely that CMOS technology will be capable of placing more transistors on a chip than can be tolerated from a power dissipation point of view.

We have established rather optimistic upper limits for transistors. For example, we assumed an on-off ratio of 2. Realistic designs require much higher ratios. We also assumed an operation voltage of $\sim k_B T$, but such a low voltage would result in too many spontaneous errors. New transistor materials and structures make allow CMOS technology to operate closer to these fundamental limits, but no barrier-modulation transistor can be fundamentally better than the silicon MOSFET.

2.9 Summary

This chapter summarized the conventional theory of MOS devices and circuits. It provides some background for examining the MOSFET from a new perspective. We seek a new understanding of small MOSFETs, one that explains the key results for submicron MOSFETs that we have just reviewed, but that also applies all the way to the scaling limit. We also seek an approach that applies to MOSFETs, as well as to the unconventional devices that are being explored to complement or even replace the MOSFET. One important point that will arise again is the central importance of self-consistent electrostatics. The device performance metrics and circuits and systems aspects of speed and power will also apply to any device intended for use in the conventional digital systems that we are familiar with.

Chapter 2 References

[2.1] J. D. Meindl, Q. Chen, and J. A. Davis, "Limits on Silicon Nanoelectronics for Terascale Integration," *Science*, **293**, pp. 2044-2049, 2001.

[2.2] D. Frank, et al., "Device Scaling Limits of Si MOSFETs and Their Application Dependencies," *Proc. IEEE*, **89**, pp. 259-288, 2001.

[2.3] M. Ieong, B. Doris, J. Kedzierski, K. Rim, and M. Yang, "Silicon Scaling to the Sub- 10-nm Regime," *Science*, **306**, pp. 2057-2060, 2004.

[2.4] X. Huang, et al., "Sub-50 nm P-channel FinFET," *IEEE Trans. Electron. Dev.*, **48**, pp. 880-886, 2001.

[2.5] B.S. Doyle, S. Datta, M. Doczy, S. Hareland, B. Jin, J. Kavalieros, T. Linton, A. Murthy, R. Rios, and R. Chau, "High Performance Fully-Depleted Tri-Gate CMOS Transistors," *IEEE Electron Dev. Lett.*, **24**, pp. 263-265, 2003.

[2.6] B. Gobel, et al., "Fully Depleted Surrounding Gate Transistor (SGT) for 70nm DRAM and Beyond," *Tech. Digest, International Electron Devices Meeting,* San Francisco, CA, Dec. 9-11, 2002.

[2.7] Y. Taur and T. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge Univ. Press, Cambridge, U.K., 1998.

[2.8] R.F. Pierret, *Fundamentals of Semiconductor Devices*, Addison-Wesley, Reading, MA, 1996.

[2.9] K. Suzuki, T. Tanaka, Y. Tosaka, H. Horie, and Y. Arimoto, "Scaling Theory of Double-Gate SOI MOSFETs," *IEEE Trans. Electron Dev.*, **40**, pp. 2326—2329, 1993.

[2.10] C.P. Auth and J.D. Plummer, "Scaling Theory for Cylindrical, Fully-Depleted, Surrounding-Gate MOSFET's," *IEEE Electron Dev. Lett.*, **18**, pp. 74-77, 1997.

[2.11] E. O. Johnson, "The Insulated-Gate Field-Effect Transistor - A Bipolar Transistor in Disguise," *RCA Review*, **34,** pp. 80-94, 1973.

[2.12] G. E. Moore, "Cramming more components onto integrated circuits," Electronics, 38, pp. 114-117, 1965

[2.13] http://public.itrs.net

[2.14] R. Landauer, "Irreversibility and Heat Generation in the Computing Process," *IBM J. Res. and Dev.*, **5**, pp. 183-191 (1961).

[2.15] C.H. Bennet, "Logical Reversability of Computation," *IBM J. Res. and Dev.*, **17**, pp. 525-532 (1973).

[2.16] S. Datta, Quantum Transport Atom to Transistor, 1st Ed., Cambridge University Press, Cambridge, UK. 2005.

[2.17] J. D. Meindl and J.A. Davis, "The Fundamental Limit on Binary Switching Energy for Terascale Integration (TSI)," *IEEE J. Solid-State Circuits,* **35**, pp. 1515-1516, 2000.

[2.18] V. V. Zhirnov, R. K. Cavin, J. A. Hutchby, G. I. Bourianoff, "Limits to Binary Logic Switch Scaling – A Gedanken Model," *Proc. of the IEEE*, Special Issue on Nanoelectronics and Nanoscale Processing, Bing Sheu, Peter Wu & Simon Sze, Guest Editors, Nov., 2003.

[2.19] B. Doris, M. Ieong, T. Kanarsky, Y. Zhang, R.A. Roy, O. Dokumaci, Z. Ren, F-F Jamin, L. Shi, W. Natzle, H-J Huang, J. Mezzapelle, A. Mocuta, S. Womack, M. Gribelyuk, E. C. Jones, R. J. Miller, H-S P. Wong, and W. Haensch, "Extreme Scaling with Ultra-Thin Si Channel MOSFETs," Tech. Dig., IEEE Electron Devices Mtg., pp. 267-270, Washington, Dec. 2002.

http://www.springer.com/978-0-387-28002-8

Nanoscale Transistors
Device Physics, Modeling and Simulation
Lundstrom, M.; Guo, J.
2006, VIII, 218 p. 106 illus., Hardcover
ISBN: 978-0-387-28002-8