7

MOSFETs in ICs—Scaling, Leakage, and Other Topics

CHAPTER OBJECTIVES

How the MOSFET gate length might continue to be reduced is the subject of this chapter. One important topic is the off-state current or the leakage current of the MOSFETs. This topic complements the discourse on the on-state current conducted in the previous chapter. The major topics covered here are the subthreshold leakage and its impact on device size reduction, the trade-off between I_{on} and I_{off} and the effects on circuit design. Special emphasis is placed on the understanding of the opportunities for future MOSFET scaling including mobility enhancement, high-*k* dielectric and metal gate, SOI, multigate MOSFET, metal source/drain, etc. Device simulation and MOSFET compact model for circuit simulation are also introduced.

Metal–oxide–semiconductor (MOS) integrated circuits (ICs) have met the world's growing needs for electronic devices for computing, communication, entertainment, automotive, and other applications with continual improvements in cost, speed, and power consumption. These improvements in turn stimulated and enabled new applications and greatly improved the quality of life and productivity worldwide.

7.1 ● TECHNOLOGY SCALING—FOR COST, SPEED, AND POWER CONSUMPTION ●

In the forty-five years since 1965, the price of one bit of semiconductor memory has dropped 100 million times. The cost of a logic gate has undergone a similarly dramatic drop. This rapid price drop has stimulated new applications and semiconductor technology has improved the ways people carry out just about all human endeavors. The primary engine that powered the proliferation of electronics is "miniaturization." By making the transistors and the interconnects smaller, more circuits can be fabricated on each silicon wafer and therefore each circuit becomes cheaper. Miniaturization has also been instrumental to the improvements in speed and power consumption of ICs.

Gordon Moore made an empirical observation in 1965 that the number of devices on a chip doubles every 18 to 24 months or so. This **Moore's Law** is a succinct description of the rapid and persistent trend of miniaturization. Each time the minimum line width is reduced, we say that a new **technology generation** or **technology node** is introduced. Examples of technology generations are 0.18 µm, 0.13 µm, 90 nm, 65 nm, 45 nm … generations. The numbers refer to the minimum metal line width. Poly-Si gate length may be even smaller. At each new node, all the features in the circuit layout, such as the contact holes, are reduced in size to 70% of the previous node. This practice of periodic size reduction is called **scaling**. Historically, a new technology node is introduced every two to three years.

The main reward for introducing a new technology node is the reduction of circuit size by half. (70% of previous line width means ~50% reduction in area, i.e., $0.7 \times 0.7 = 0.49$.) Since nearly twice as many circuits can be fabricated on each wafer with each new technology node, the cost per circuit is reduced significantly. That drives down the cost of ICs.

● Initial Reactions to the Concept of the IC ●

Anecdote contributed by Dr. Jack Kilby, January 22, 1991

"Today the acceptance of the integrated circuit concept is universal. It was not always so. When the integrated circuit was first announced in 1959, several objections were raised. They were:

1) Performance of transistors might be degraded by the compromises necessary to include other components such as resistors and capacitors.

2) Circuits of this type were not producible. The overall yield would be too low.

3) Designs would be expensive and difficult to change.

Debate of the issues provided the entertainment at technical meetings for the next five or six years."

In 1959, Jack Kilby of Texas Instruments and Robert Noyce of Fairchild Semiconductor independently invented technologies of interconnecting multiple devices on a single semiconductor chip to form an electronic circuit. Following a 10 year legal battle, both companies' patents were upheld and Noyce and Kilby were recognized as the co-inventors of the IC. Dr. Kilby received a Nobel Prize in Physics in 2000 for inventing the integrated circuit. Dr. Noyce, who is credited with the layer-by-layer planar approach of fabricating ICs, had died in 1990.

Besides the line width, some other parameters are also reduced with scaling such as the MOSFET gate oxide thickness and the power supply voltage. The reductions are chosen such that the transistor current density (I_{on}/W) increases with each new node. Also, the smaller transistors and shorter interconnects lead to smaller capacitances. Together, these changes cause the circuit delays to drop (Eq. 6.7.1). Historically, IC speed has increased roughly 30% at each new technology node. Higher speed enables new applications such as wide-band data transmission via RF mobile phones.

Scaling does another good thing. Eq. (6.7.6) shows that reducing capacitance and especially the power supply voltage is effective in lowering the power consumption. Thanks to the reduction in C and V_{dd}, power consumption per chip has increased only modestly per node in spite of the rise in switching frequency, f and the doubling of transistor count per chip at each technology node. If there had been no scaling, doing the job of a single PC microprocessor chip (operating a billion transistors at 2 GHz) using 1970 technology would require the power output of an electrical power generation plant.

In summary, scaling improves cost, speed, and power consumption per function with every new technology generation. All of these attributes have been improved by 10 to 100 million times in four decades—an engineering achievement unmatched in human history! When it comes to ICs, small is beautiful.

7.1.1 Innovations Enable Scaling

Semiconductor researchers around the world have been meeting several times a year for the purpose of generating consensus on the transistor and circuit performance that will be required to fulfill the projected market needs in the future. Their annually updated document: **International Technology Roadmap for Semiconductors (ITRS)** only sets out the goals and points out the challenging problems but does not provide the solutions [1]. It tells the vendors of manufacturing tools and materials and the research community the expected roadblocks. The list of show stoppers is always long and formidable but innovative engineers working together and separately have always risen to the challenge and done the seemingly impossible.

Table 7–1 is a compilation of some history and some ITRS technology projection. High-performance (HP) stands for high-performance computer processor technology. LSTP stands for the technology for low standby-power products such as mobile phones. The physical gate length, L_g, is actually smaller than the technology node. Take the 90 nm node, for example; although lithography technology can only print 90 nm photoresist lines, engineers transfer the pattern into oxide lines and then isotropically etch (see Section 3.4) the oxide in a dry isotropic-etching tool to reduce the width (and the thickness) of the oxide lines. Using the narrowed oxide lines as the new etch mask, they produce the gate patterns by etching. Innumerable innovations by engineers at each node have enabled the scaling of the IC technology.

7.1.2 Strained Silicon and Other Innovations

I_{on} in Table 7–1 rises rapidly. This is only possible because of the **strained silicon** technology introduced around the 90 nm node [2]. The electron and hole mobility can be raised (or lowered) by carefully engineered mechanical strains. The strain changes the lattice constant of the silicon crystal and therefore the E–k relationship through the Schrodinger's wave equation. The E–k relationship, in turn, determines the effective mass and the mobility.

For example, the hole surface mobility of a PFET can be raised when the channel is compressively stressed. The compressive strain may be created in several ways. We illustrate one way in Fig. 7–1. After the gate is defined, trenches are etched into the silicon adjacent to the gate. The trenches are refilled by

TABLE 7–1 • Scaling from 90 nm to 22 nm and innovations that enable the scaling.

Year of Shipment	**2003**	**2005**	**2007**	**2010**	**2013**
Technology Node (nm)	90	65	45	32	22
L_g (nm) (HP/LSTP)	37/65	26/45	22/37	16/25	13/20
EOT_e(nm) (HP/LSTP)	1.9/2.8	1.8/2.5	1.2/1.9	0.9/1.6	0.9/1.4
V_{DD} (V) (HP/LSTP)	1.2/1.2	1.1/1.1	1.0/1.1	1.0/1.0	0.9/0.9
I_{on}, HP (μA/μm)	1100	1210	1500	1820	2200
I_{off}, HP (μA/μm)	0.15	0.34	0.61	0.84	0.37
I_{on}, LSTP (μA/μm)	440	465	540	540	540
I_{off}, LSTP (μA/μm)	1E-5	1E-5	3E-5	3E-5	2E-5
Innovations	↦ Strained Silicon				
			↦ High-*k*/metal-gate		
				↦ Wet lithography	
					↦ New Structure

HP: High-Performance technology. LSTP: Low Standby Power technology for portable applications. EOT_e: Equivalent electrical Oxide Thickness, i.e., equivalent T_{oxe}. I_{on}: NFET I_{on}.

epitaxial growth (see Section 3.7.3) of SiGe—typically a 20% Ge and 80% Si mixture. Because Ge atoms are larger than Si atoms and in epitaxial growth the number of atoms in the trench is equal to the original number of Si atoms, it is as if a large hand is forced into a small glove. A force is created that pushes on the channel (as shown in Fig. 7–10) region and raises the hole mobility. It is also attractive to incorporate a thin film of Ge material in the channel itself because Ge has higher carrier mobilities than Si [3].

In Table 7–1, EOT_e or the **electrical equivalent oxide thickness** is the total thickness of the gate dielectric, poly-gate depletion (if any), and the inversion layer expressed in equivalent SiO_2 thickness. It is improved (reduced) at the 45 nm node by a larger factor over the previous node. The enabling innovations are metal gate and high-*k* dielectric, which will be presented in Section 7.4.

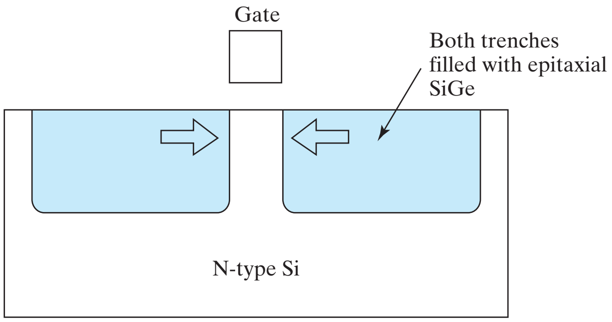


FIGURE 7–1 Example of strained-silicon MOSFET. Hole mobility can be raised with a compressive mechanical strain illustrated with the arrows pushing on the channel region.

At the 32 nm node, wet lithography (see Section 3.3.1) is used to print the fine patterns. At the 22 nm node, new transistor structures may be used to reverse the trend of increasing I_{off}, which is the source of a serious power consumption issue. Some new structures are presented in Section 7.8.

7.2 • SUBTHRESHOLD CURRENT—"OFF" IS NOT TOTALLY "OFF" •

Circuit speed improves with increasing I_{on}; therefore, it would be desirable to use a small V_t. Can we set V_t at an arbitrarily small value, say 10 mV? The answer is no.

At $V_{gs} < V_t$, an N-channel MOSFET is in the off state. However, a leakage current can still flow between the drain and the source. The MOSFET current observed at $V_{gs} < V_t$ is called the **subthreshold current**. This is the main contributor to the MOSFET **off-state current**, I_{off}. I_{off} is the I_d measured at $V_{gs} = 0$ and $V_{ds} = V_{dd}$. It is important to keep I_{off} very small in order to minimize the static power that a circuit consumes when it is in the standby mode. For example, if I_{off} is a modest 100 nA per transistor, a cell-phone chip containing one hundred million transistors would consume 10 A even in standby. The battery would be drained in minutes without receiving or transmitting any calls. A desktop PC processor would dissipate more power because it contains more transistors and face expensive problems of cooling the chip and the system.

Figure 7–2a shows a subthreshold current plot. It is plotted in a semi-log I_{ds} vs. V_{gs} graph. When V_{gs} is below V_t, I_{ds} is clearly a straight line, i.e., an exponential function of V_{gs}.

Figure 7–2b–d explains the subthreshold current. At V_{gs} below V_t, the inversion electron concentration (n_s) is small but nonetheless can allow a small leakage current to flow between the source and the drain. In Fig. 7–2b, a larger V_{gs} would pull the E_c at the surface closer to E_F, causing n_s and I_{ds} to rise. From the equivalent circuit in Fig. 7–2c, one can observe that

$$\frac{d\varphi_s}{dV_{gs}} = \frac{C_{oxe}}{C_{oxe} + C_{dep}} \equiv \frac{1}{\eta} \tag{7.2.1}$$

$$\eta = 1 + \frac{C_{dep}}{C_{oxe}} \tag{7.2.2}$$

Integrating Eq. (7.2.1) yields

$$\varphi_s = \text{constant} + V_g/\eta \tag{7.2.3}$$

I_{ds} is proportional to n_s, therefore

$$I_{ds} \propto n_s \propto e^{q\varphi_s/kT} \propto e^{q(\text{constant}+V_g/\eta)/kT} \propto e^{qV_g/\eta kT} \tag{7.2.4}$$

A practical and common definition of V_t is the V_{gs} at which $I_{ds} = 100\text{ nA} \times W/L$ as shown in Fig. 6–12. (Some companies may use 200 nA instead of 100 nA.). Equation (7.2.4) may be rewritten as

$$I_{ds}(nA) = 100 \cdot \frac{W}{L} \cdot e^{q(V_{gs} - V_t)/\eta kT} \tag{7.2.5}$$

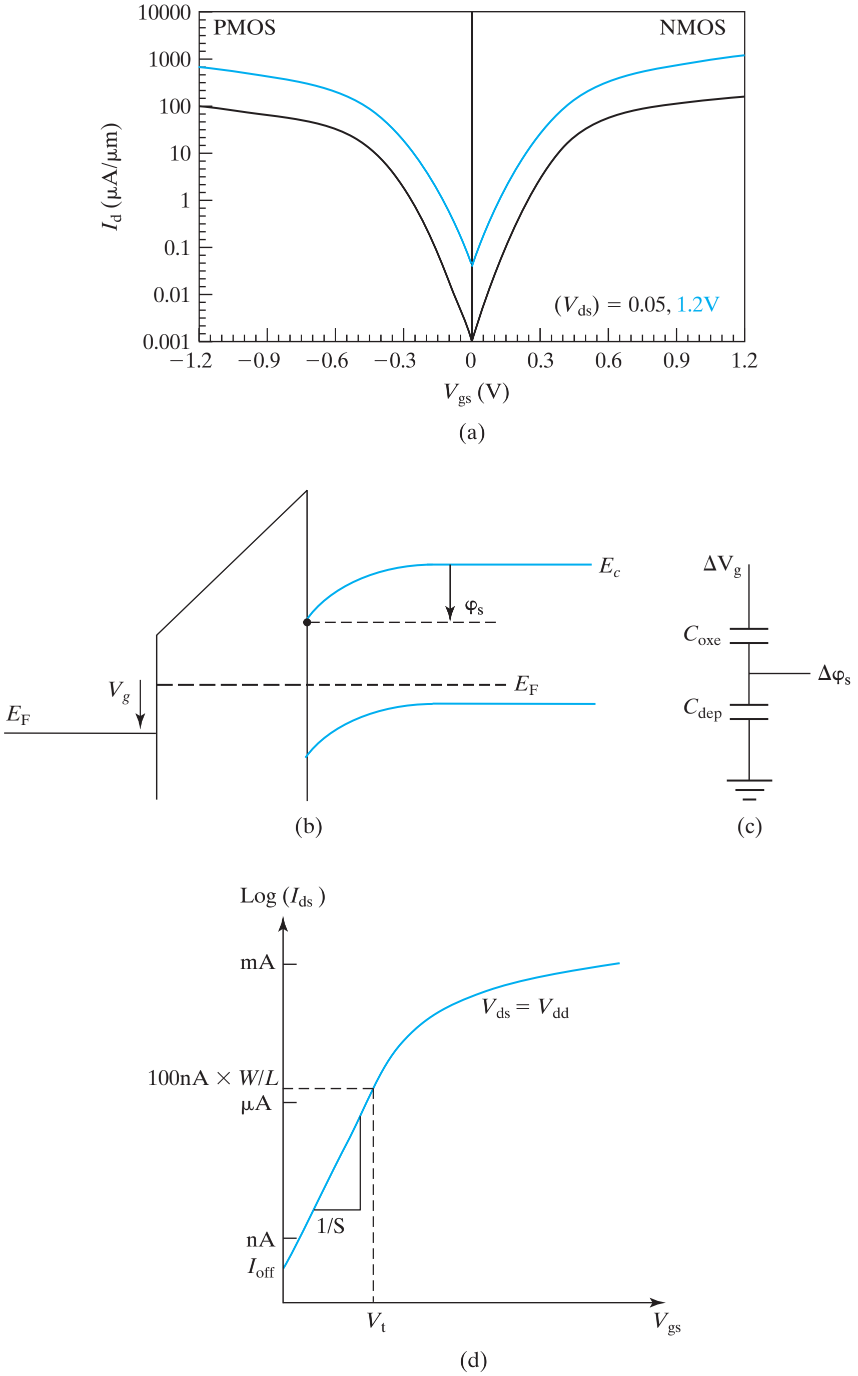


FIGURE 7–2 The current that flows at $V_{gs} < V_t$ is called the subthreshold current. V_t ~ 0.2 V. The lower/upper curves are for V_{ds} = 50 mV/1.2 V. After Ref. [2]. (b) When V_g is increased, E_c at the surface is pulled closer to E_F, causing n_s and I_{ds} to rise; (c) equivalent capacitance network; (d) subthreshold I-V with V_t and I_{off}. Swing, S, is the inverse of the slope in the subthreshold region.

Clearly, Eq. (7.2.5) agrees with the definition of V_t and Eq. (7.2.4). The simplicity of Eq. (7.2.5) is another reason for favoring the new V_t definition. At room temperature, the function $\exp(qV_{gs}/kT)$ changes by 10 for every 60 mV change in V_{gs}, therefore $\exp(qV_{gs}/\eta kT)$ changes by 10 for every $\eta \times 60$ mV. For example, if $\eta = 1.5$, Eq. (7.2.5) states that I_{ds} drops by ten times for every 90 mV of decrease in V_{gs} below V_t at room temperature. $\eta \times 60$ mV is called the **subthreshold swing** and represented by the symbol, S.

$$S(\mathrm{mV/decade}) = \eta \cdot 60\ \mathrm{mV} \cdot \frac{T}{300\mathrm{K}} \tag{7.2.6}$$

$$I_{ds}(\mathrm{nA}) = 100 \cdot \frac{W}{L} \cdot e^{q(V_{gs}-V_t)/\eta kT} = 100 \cdot \frac{W}{L} \cdot 10^{(V_{gs}-V_t)/S} \tag{7.2.7}$$

$$I_{off}(\mathrm{nA}) = 100 \cdot \frac{W}{L} \cdot e^{-qV_t/\eta kT} = 100 \cdot \frac{W}{L} \cdot 10^{-V_t/S} \tag{7.2.8}$$

For given W and L, there are two ways to minimize I_{off} illustrated in Fig. 7–2 (d). The first is to choose a large V_t. This is not desirable because a large V_t reduces I_{on} and therefore degrades the circuit speed (see Eq. (6.7.1)). The preferable way is to reduce the subthreshold swing. S can be reduced by reducing η. That can be done by increasing C_{oxe} (see Eq. 7.2.2), i.e., using a thinner T_{ox}, and by decreasing C_{dep}, i.e., increasing W_{dep}.[1] An additional way to reduce S, and therefore to reduce I_{off}, is to operate the transistors at significantly lower than the room temperature. This last approach is valid in principle but rarely used because cooling adds considerable cost.

Besides the subthreshold leakage, there is another leakage current component that has becomes significant. That is the tunnel leakage through very thin gate oxide that will be presented in Section 7.4. The drain to the body junction leakage is the third leakage component.

● The Effect of Interface States ●

The subthreshold swing is degraded when interface states are present (see Section 5.7). Figure 7–3 shows that when φ_S changes, some of the interface traps move from above the Fermi level to below it or vice versa. As a result, these interface traps change from being empty to being occupied by electrons. This change of charge in response to change of voltage (φ_S) has the effect of a capacitor. The effect of the interface states is to add a parallel capacitor to C_{dep} in Fig. 7–2c. The subthreshold swing is poor unless the semiconductor-dielectric interface has low density of interface states such as carefully prepared Si-SiO$_2$ interface. The subthreshold swing is often degraded after a MOSFET is electrically stressed (see sidebar in Section 5.7) and new interface states are generated.

[1] According to Eq. 6.5.2 and Eq. 7.2.2, η should be equal to m. In reality, η is larger than m because C_{oxe} is smaller at low V_{gs} (subthreshold condition) than in inversion due to a larger T_{inv} as shown in Fig. 5–25. Nonetheless, η and m are closely related.

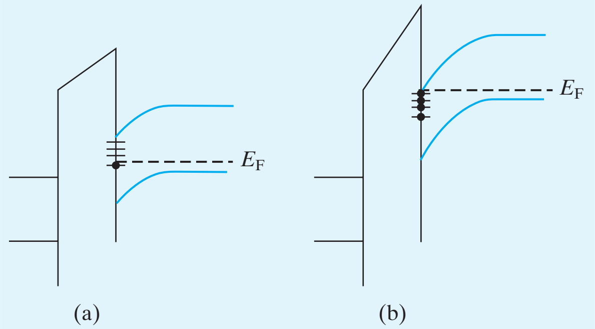


FIGURE 7–3 (a) Most of the interface states are empty because they are above E_F. (b) At another V_g, most of the interface states are filled with electrons. As a result, the interface charge density changes with V_g.

EXAMPLE 7–1 Subthreshold Leakage Current

An N-channel transistor has $V_t = 0.34$ V and S = 85 mV, W = 10 µm and L = 50 nm. (a) Estimate I_{off}. (b) Estimate I_{ds} at $V_g = 0.17$ V.

SOLUTION:

a. Use Eq. (7.2.6).

$$I_{off}(\mathrm{nA}) = 100 \cdot \frac{W}{L} \cdot 10^{-V_t/S} = 100 \cdot \frac{10}{0.05} \cdot 10^{-0.34/0.0085} = 2 \text{ nA}$$

b. Use Eq. (7.2.7).

$$I_{ds} = 100 \cdot \frac{W}{L} \cdot 10^{(V_g - V_t)/S} = 100 \cdot \frac{10}{0.05} \cdot 10^{(0.17-0.34)/0.085} = 200 \text{ nA}$$

7.3 • V_t ROLL-OFF—SHORT-CHANNEL MOSFETs LEAK MORE •

The previous section pointed out that V_t must not be set too low; otherwise, I_{off} would be too large. The present section extends that analysis to show that the channel length (L) must not be too short. The reason is this: V_t drops with decreasing L as illustrated in Fig. 7–4. When V_t drops too much, I_{off} becomes too large and that channel length is not acceptable.

• Gate Length (L_g) vs. Electrical Channel Length (L) •

Gate length is the physical length of the gate and can be accurately measured with a scanning electron microscope (SEM). It is carefully controlled in the fabrication plant. The channel length, in comparison, cannot be determined very accurately and easily due to the lateral diffusion of the source and drain junctions. L tracks L_g but the difference between the two just cannot be quantified precisely in spite of efforts such as described in Section 6.11. As a result, L_g is widely used in lieu of L in data presentations as is done in Fig. 7–4. L is still a useful concept and is used in theoretical equations even though L cannot be measured precisely for small transistors.

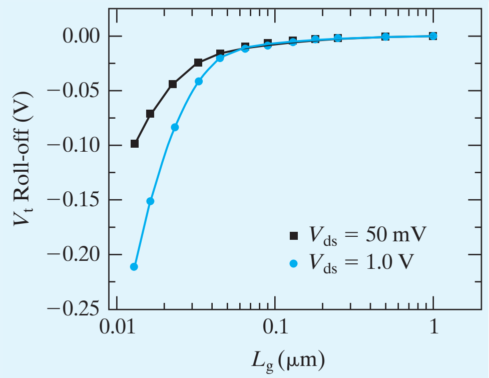


FIGURE 7–4 $|V_t|$ decreases at very small L_g. This phenomenon is called V_t roll-off. It determines the minimum acceptable L_g because I_{off} is too large when V_t becomes too low or too sensitive to L_g.

At a certain L_g, V_t becomes so low that I_{off} becomes unacceptable [see Eq. (7.2.8)]. Doping the bodies of the short-channel devices more heavily than the long-channel devices would raise their V_t. Still, at a certain L_g, V_t is so sensitive to the manufacturing caused variation in L that the worst case I_{off} becomes unacceptable. Device development engineers must design the device such that the V_t roll-off does not prevent the use of the targeted minimum L_g, e.g., those listed in the second row of Table 7–1.

Why does V_t decrease with decreasing L? Figure 7–5 illustrates a model for understating this effect. Figure 7–5a shows the energy band diagram along the semiconductor–insulator interface of a long channel device at $V_{gs} = 0$. Figure 7–5b shows the case at $V_{gs} = V_t$. In the case of (b), E_c in the channel is pulled lower than

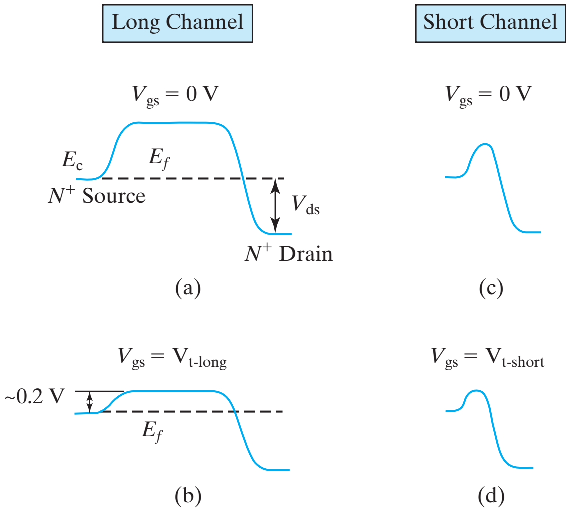


FIGURE 7–5 a–d: Energy band diagram from source to drain when $V_{gs} = 0$ V and $V_{gs} = V_t$. a–b long channel; c–d short channel.

in case (a) and therefore is closer to the E_c in the source. When the channel E_c is only ~0.2 eV higher than the E_c in the source (which is also ~E_{Fn}), n_s in the channel reaches ~10^{17} cm^3 and inversion threshold condition (I_{ds} = 100nA × W/L) is reached. We may say that a 0.2 eV potential barrier is low enough to allow the electrons in the N^+ source to flow into the channel to form the inversion layer. The following analogy may be helpful for understanding the concept of the energy barrier height. The source is a reservoir of water; the potential barrier is a dam; and V_{gs} controls the height of the dam. When V_{gs} is high enough, the dam is sufficiently low for the water to flow into the channel and the drain. That defines V_t.

Figure 7–5c shows the case of a short-channel device at $V_{gs} = 0$. If the channel is short enough, E_c will not be able to reach the same peak value as in Fig. 7–5a. As a result, a smaller V_{gs} is needed in Fig. 7–5d than in Fig. 7–5b to pull the barrier down to 0.2 eV. In other words, V_t is lower in the short channel device than the long channel device. This explains the V_t roll-off shown in Fig. 7–4.

We can understand V_t roll-off from another approach. Figure 7–6 shows a capacitor between the gate and the channel. It also shows a second capacitor, C_d, between the drain and the channel terminating at around the middle of the channel, where E_c peaks in Fig. 7–5d. As the channel length is reduced, the drain to source and the drain to "channel" distances are reduced; therefore, C_d increases. Do not be concerned with the exact definition or value of C_d. Instead, focus on the concept that C_d represents the capacitive coupling between the drain and the channel barrier point.

From this two-capacitor equivalent circuit, it is evident that the drain voltage has a similar effect on the channel potential as the gate voltage. V_{gs} and V_{ds}, together, determine the channel potential barrier height shown in Fig. 7–5. When V_{ds} is present, less V_{gs} is needed to pull the barrier down to 0.2 eV; therefore, V_t is lower by definition. This understanding gives us a simple equation for V_t roll-off,

$$V_t = V_{\text{t-long}} - V_{ds} \cdot \frac{C_d}{C_{oxe}} \tag{7.3.1}$$

where $V_{\text{t-long}}$ is the threshold voltage of a long-channel transistor, for which $C_d = 0$. More accurately, V_{ds} should be supplemented with a constant that represents the combined effects of the 0.2 V built-in potentials between the N^- inversion layer and both the N^+ drain and source at the threshold condition [4].

$$V_t = V_{\text{t-long}} - (V_{ds} + 0.4\text{ V}) \cdot \frac{C_d}{C_{oxe}} \tag{7.3.2}$$

Using Fig. 7–6, one can intuitively see that as L decreases, C_d increases. Recall that the capacitance increases when the two electrodes are closer to each other. That intuition is correct for the two-dimensional geometry of Fig. 7–6, too. However, solution of the Poisson's equation (Section 4.1.3) indicates that C_d is an exponential function of L in this two-dimensional structure [5]. Therefore,

$$V_t = V_{\text{t-long}} - (V_{ds} + 0.4\text{ V}) \cdot e^{-L/l_d} \tag{7.3.3}$$

where $$l_d \propto \sqrt[3]{T_{oxe} W_{dep} X_j} \tag{7.3.4}$$

X_j is the drain junction depth. Equation (7.3.3) provides a semi-quantitative model of the roll-off of V_t as a function of L and V_{ds}. It can serve as a guide for designing

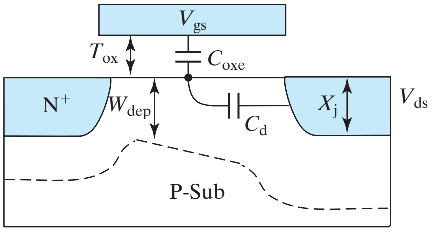


FIGURE 7–6 Schematic two-capacitor network in MOSFET. C_d models the electrostatic coupling between the channel and the drain. As the channel length is reduced, drain to "channel" distance is reduced; therefore, C_d increases.

small MOSFET and understanding new transistor structures. At a very large L, V_t is equal to $V_{t\text{-long}}$ as expected. The roll-off is an exponential function of L. The roll-off is also larger at larger V_{ds}, which can be as large as V_{dd}. The acceptable I_{off} determines the acceptable V_t through Eq. (7.2.8). This in turn determines the acceptable minimum L through Eq. (7.3.3). *The acceptable minimum L is several times of l_d.* The concept that the drain can lower the source–channel barrier and reduce V_t is called **drain-induced barrier lowering** or **DIBL**. l_d may be called the **DIBL characteristic length**. In order to support the reduction of L at each new technology node, l_d must be reduced in proportion to L. This means that we must reduce T_{ox}, W_{dep}, and/or X_j. In reality, all three are reduced at each node to achieve the desired reduction in l_d. Reducing T_{ox} increases the gate control or C_{oxe}. Reducing X_j decreases C_d by reducing the size of the drain electrode. Reducing W_{dep} also reduces C_d by introducing a ground plane (the neutral region of the substrate or the bottom of the depletion region) that tends to electrostatically shield the channel from the drain.

The basic message in Eq. (7.3.4) is that *the vertical dimensions in a MOSFET (T_{ox}, W_{dep}, X_j) must be reduced in order to support the reduction of the gate length.* As an example, Fig. 7–7 shows that the oxide thickness has been scaled roughly in proportion to the line width (gate length).

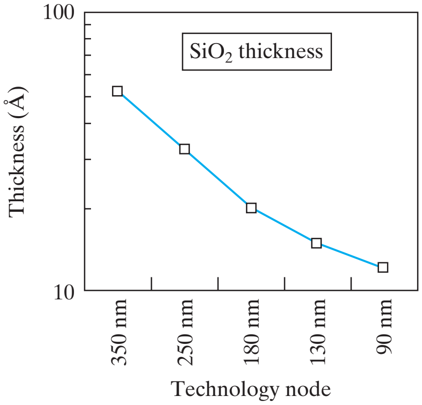


FIGURE 7–7 In the past, the gate oxide thickness has been scaled roughly in proportion to the line width.

7.4 • REDUCING GATE-INSULATOR ELECTRICAL THICKNESS AND TUNNELING LEAKAGE •

SiO_2 has been the preferred gate insulator since silicon MOSFET's beginning. The oxide thickness has been reduced over the years from 300 nm for the 10 μm technology to only 1.2 nm for the 65 nm technology. There are two reasons for the relentless drive to reduce the oxide thickness. First, a thinner oxide, i.e., a larger C_{ox} raises I_{on} and a large I_{on} raises the circuit speed [see Eq. (6.7.1)]. The second reason is to control V_t roll-off (and therefore the subthreshold leakage) in the presence of a shrinking L according to Eqs. (7.3.3) and (7.3.4). One must not underestimate the importance of the second reason. Figure 7–7 shows that the oxide thickness has been scaled roughly in proportion to the line width.

Thinner oxide is desirable. What, then, prevents engineers from using arbitrarily thin gate oxide films? Manufacturing thin oxide is not easy, but as Fig. 6–5 illustrates, it is possible to grow very thin and uniform gate oxide films with high yield. Oxide breakdown is another limiting factor. If the oxide is too thin, the electric field in the oxide can be so high as to cause destructive breakdown. (See the sidebar, "SiO_2 Breakdown Electric Field.") Yet another limiting factor is that long-term operation at high field, especially at elevated chip operating temperatures, breaks the weaker chemical bonds at the Si–SiO_2 interface thus creating oxide charge and V_t shift (see Section 5.7). V_t shifts cause circuit behaviors to change and raise reliability concerns.

For SiO_2 films thinner than 1.5 nm, tunneling leakage current becomes the most serious limiting factor. Figure 7–8a illustrates the phenomenon of gate leakage by tunneling (see Section 4.20). Electrons arrive at the gate oxide barrier at thermal velocity and emerge on the side of the gate with a probability given by Eq. (4.20.1). This is the cause of the gate leakage current. Figure 7–8b shows that the exponential rise of the SiO_2 leakage current with decreasing thickness agrees with the tunneling model prediction [6]. At 1.2 nm, SiO_2 leaks 10^3 A/cm^2. If an IC chip contains

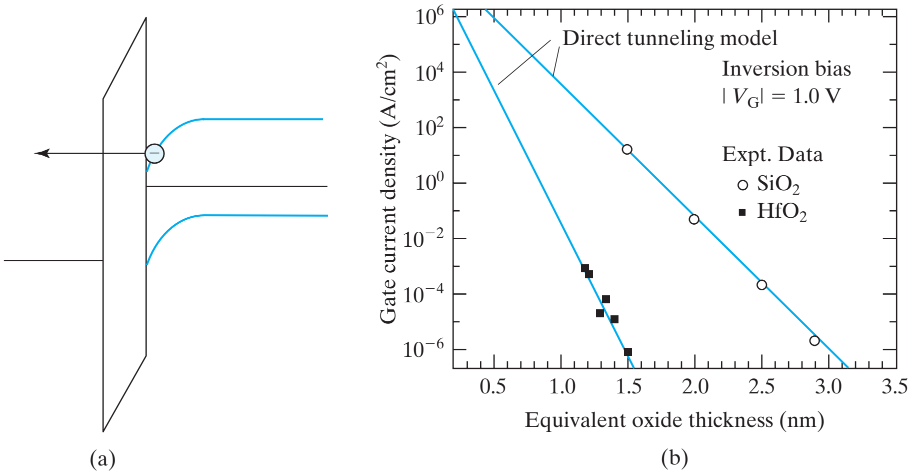


FIGURE 7–8 (a) Energy band diagram in inversion showing electron tunneling path through the gate oxide; (b) 1.2 nm SiO_2 conducts 10^3 A/cm^2 of leakage current. High-k dielectric such as HfO_2 allows several orders lower leakage current to pass. (After [6]. © 2003 IEEE.)

1 mm^2 total area of this thin dielectric, the chip oxide leakage current would be 10 A. This large leakage would drain the battery of a cell phone in minutes. The leakage current can be reduced by about 10 × with the addition of nitrogen into SiO_2.

Engineers have developed high-k dielectric technology to replace SiO_2. For example, HfO_2 has a relative dielectric constant (k) of ~24, six times larger than that of SiO_2. A 6 nm thick HfO_2 film is equivalent to 1 nm thick SiO_2 in the sense that both films produce the same C_{ox}. We say that this HfO_2 film has an **equivalent oxide thickness** or **EOT** of 1 nm. However, the HfO_2 film presents a much thicker (albeit lower) tunneling barrier to the electrons and holes. The consequence is that the leakage current through HfO_2 is several orders of magnitude smaller than that through SiO_2 as shown in Fig. 7–8b. Other attractive high-k dielectrics include ZrO_2 and Al_2O_3. The difficulties of adopting high-k dielectrics in IC manufacturing are chemical reactions between them and the silicon substrate, lower surface mobility than the Si–SiO_2 system, and more oxide charge. These problems are minimized by inserting a thin SiO_2 interfacial layer between the silicon substrate and the high-k dielectric.

Note that Eq. (7.3.4) contains the electrical oxide thickness, T_{oxe}, defined in Eq. (5.9.2). Besides T_{ox} or EOT, the poly-Si gate depletion layer thickness also needs to be minimized. Metal is a much better gate material in this respect. NFET and PFET gates may require two different metals (with metal work functions close to those of N^+ and P^+ poly-Si) in order to achieve the optimal V_ts [7].

In addition, T_{inv} is also part of T_{oxe} and needs to be minimized. The material parameters that determine T_{inv} is the electron or hole effective mass. A larger effective mass leads to a thinner T_{inv}. Unfortunately, a larger effective mass leads to a lower mobility, too (see Eq. (2.2.4)). Fortunately, the effective mass is a function of the spatial direction in a crystal. The effective mass in the direction normal to the oxide interface determines T_{inv}, while the effective mass in the direction of the current flow determines the surface mobility. It may be possible to build a transistor with a wafer orientation (see Fig. 1–2) that offers larger m_n and m_p normal to the oxide interface but smaller m_n and m_p in the direction of the current flow.

● SiO_2 Breakdown Electric Field ●

What is the breakdown field of SiO_2? There is no one simple answer because the breakdown field is a function of the test time. If a one second (1s) voltage pulse is applied to a 10 nm SiO_2 film, 15 V is needed to breakdown the film for a breakdown field of 15 MV/cm. The breakdown field is significantly lower if the same oxide is tested for one hour. The field is lower still if it is tested for a month. This phenomenon is called **time-dependent dielectric breakdown**. Most IC applications require a device lifetime of several years to over 10 years. Clearly, manufacturers cannot afford the time to actually measure the 10 year breakdown voltage for new oxide technologies. Instead, engineers predict the 10 year breakdown voltage based on hours- to month-long tests in combination with theoretical models of the physics of oxide breakdown. A wide range of breakdown field was predicted for SiO_2 by different models. In retrospect, the most optimistic of the predictions, 7 MV/cm for a 10 year operation, was basically right.

This breakdown model considers a sequence of events[8]. Carrier tunneling through the oxide at high field breaks up the weaker Si–O bonds in SiO_2, thus creating oxide defects. This process progresses more rapidly at those spots in the oxide sample where the densities of the weaker bonds happen to be statistically high. When the generated defects reach a critical density at any one spot, breakdown occurs. In a longer-term stress test, the breakdown field is lower because a lower rate of defect generation is sufficient to build up the critical defect density over the longer test time. A fortuitous fact is that the breakdown field increases in very thin oxide. The charge carriers gain less energy while traversing through a very thin oxide than a thick oxide film at a given electric field and are less able to create oxide defects.

7.5 • HOW TO REDUCE W_{dep} •

Equation (7.3.4) suggests that a small W_{dep} helps to control V_t roll-off and enable the use of a shorter L. W_{dep} can be reduced by increasing the substrate doping concentration, N_{sub}, because W_{dep} is proportional to $1/\sqrt{N_{sub}}$. However, Eq. (5.4.3), repeated here,

$$V_t = V_{fb} + \phi_{st} + \frac{\sqrt{qN_{sub}2\varepsilon_s\phi_{st}}}{C_{ox}} \tag{7.5.1}$$

dictates that, if V_t is not to increase, N_{sub} must not be increased unless C_{ox} is increased, i.e., T_{ox} is reduced. Equation (7.5.1) can be rewritten as Eq (7.5.2) by eliminating N_{sub} with Eq. (5.5.1). Clearly, W_{dep} can only be reduced in proportion to T_{ox}.

$$V_t = V_{fb} + \phi_{st}\left(1 + \frac{2\varepsilon_s T_{ox}}{\varepsilon_{ox} W_{dep}}\right) \tag{7.5.2}$$

This fact establishes T_{ox} as the main enabler of L reduction according to Eq. (7.3.4).

There is another way of reducing W_{dep}—adopt the steep retrograde doping profile illustrated in Fig. 6–12. In that case, W_{dep} is determined by the thickness of the lightly doped surface layer. It can be shown (see sidebar) that V_t of a MOSFET with ideal retrograde doping is

$$V_t = V_{fb} + \phi_{st}\left(1 + \frac{\varepsilon_s T_{ox}}{\varepsilon_{ox} T_{rg}}\right) \tag{7.5.3}$$

where T_{rg} is the thickness of the lightly doped thin layer. Again, T_{rg} in Eq. (7.5.3) can only be scaled in proportion to T_{ox} if V_t is to be kept constant. However, T_{rg}, the W_{dep} of an ideal retrograde device, can be about half the W_{dep} of a uniformly doped device [see Eq. (7.5.2)] and yield the same V_t. That is an advantage of the retrograde doping. Another advantage of retrograde doping is that ionized impurity scattering (see Section 2.2.2) in the inversion layer is reduced and the surface mobility can be higher. To produce a sharp retrograde profile with a very thin lightly doped layer, i.e., a very small W_{dep}, care must be taken to prevent dopant diffusion.

● Derivation of Eq. (7.5.3) ●

The energy diagram at the threshold condition is shown in Fig. 7–9.

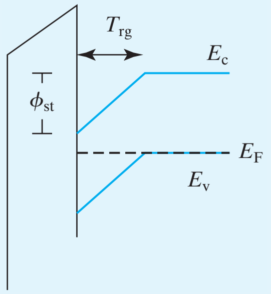


FIGURE 7–9 Energy diagram of a steep-retrograde doped MOSFET at the threshold condition.

The band bending, ϕ_{st}, is dropped uniformly over T_{rg}, the thickness of the lightly doped depletion layer, creating an electric field, $\mathscr{E}_s = \phi_{st}/T_{rg}$. Because of the continuity of the electric flux, the oxide field is $\mathscr{E}_{ox} = \mathscr{E}_s \cdot \varepsilon_s/\varepsilon_{ox}$. Therefore,

$$V_{ox} = T_{ox}\mathscr{E}_{ox} = \phi_{st}\frac{\varepsilon_s T_{ox}}{\varepsilon_{ox} T_{rg}} \tag{7.5.4}$$

From Eqs. (5.2.2), (7.5.4)

$$V_t = V_{fb} + \phi_{st}\left(1 + \frac{\varepsilon_s T_{ox}}{\varepsilon_{ox} T_{rg}}\right) \tag{7.5.5}$$

Here is an intriguing note about reducing W_{dep} further. A higher N_{sub} in Eq. (7.5.1) (and therefore a smaller W_{dep}) or a smaller T_{rg} in Eq. (7.5.3) can be used although it produces a large V_t than desired if this larger V_t is brought back down with a body (or well) to source bias voltage, V_{bs} (see Section 6.4). The required V_{bs} is a forward bias across the body–source junction. A forward bias is acceptable, i.e., the forward bias current is small, if V_{bs} is kept below 0.6 V.

● Predicting the Ultimate Low Limit of Channel Length—A Retrospective ●

When the channel length is too small, a MOSFET would have too large an I_{off} and it ceases to be a usable transistor for practical purposes. Assuming that lithography and etching technologies can produce as small features as one desires, what is the ultimate low limit of MOSFET channel length?

In the 1970s, the consensus in the semiconductor industry was that the ultimate lower limit of channel length is 500 nm. In the 1980s, the consensus was 250 nm. In the 1990s, it was 100 nm. Now it is much smaller. What made the experts underestimate the channel length scaling potential?

A review of the historical literature reveals that the researchers were mistaken about how thin the engineers can make the gate oxide in mass production. In the 1970s, it was thought that ~15 nm would be the limit. In the 1980s, it was 8 nm, and so on. Since the T_{ox} estimate was off, the estimates of the minimum acceptable W_{dep} and therefore the minimum L would be off according to Eq. (7.3.4).

7.6 • SHALLOW JUNCTION AND METAL SOURCE/DRAIN MOSFET •

Figure 7–10, first introduced as Fig. 6–24b, shows the cross-sectional view of a typical drain (and source) junction. Extra process steps are taken to produce the **shallow junction extension** between the deep N^+ junction and the channel. This shallow junction is needed because the drain junction depth must be kept small according to Eq. (7.3.4). In order to keep this junction shallow, only very short annealing at the lowest necessary temperature is used to activate the dopants and anneal out the implantation damages in the crystal in 0.1S (flash annealing) or 1μS (laser annealing) (see Section 3.6). To further reduce dopant diffusion, the doping concentration in the **shallow junction extension** is kept much lower than the N^+ doping density. Shallow junction and light doping combine to produce an undesirable parasitic resistance that reduces the precious I_{on}. That is a price to pay for suppressing V_t roll-off and the subthreshold leakage current. Farther away from the channel, as shown in Fig. 7–10, a deeper N^+ junction is used to minimize total parasitic resistance. The width of the dielectric spacer in Fig. 7–10 should be as small as possible to minimize the resistance.

7.6.1 MOSFET with Metal Source/Drain

A **metal source/drain MOSFET** or **Schottky source/drain MOSFET** shown in Fig. 7–11a can have very shallow junctions (good for the short-channel effect) and low series-resistance because the silicide is ten times more conductive than N+ or

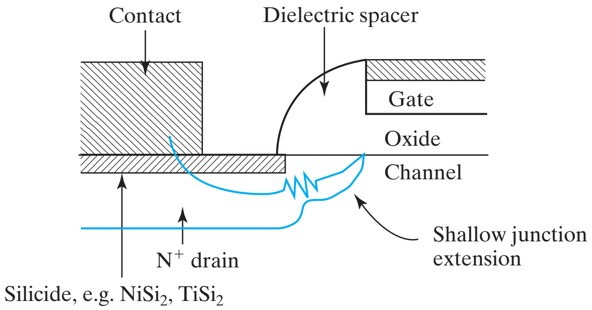
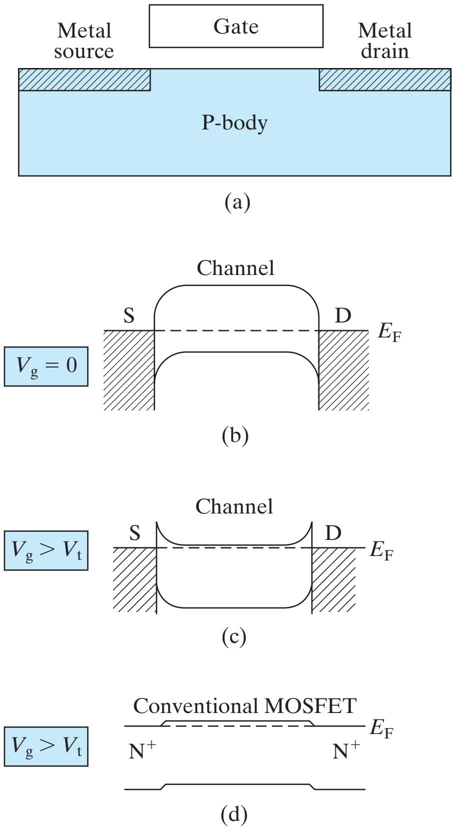


FIGURE 7–10 Cross-sectional view of a MOSFET drain junction. The shallow junction extension next to the channel helps to suppress the V_t roll-off.



(a)

(b)

(c)

(d)

FIGURE 7–11 (a) Metal source/drain is the ultimate way to reduce the increasingly important parasitic resistance; (b) energy band diagrams in the off state; (c) in the on state there may be energy barriers impeding current flow. These barriers do not exist in the conventional MOSFET (d) and must be minimized.

P+ Si. The only problem is that the Schottky-S/D MOSFET would have a lower I_d than the regular MOSFET if ϕ_B is too large to allow easy flow of carriers (electrons for NFET) from the source into the channel.

Figure 7–11b shows the energy band diagram drawn from the source along the channel interface to the drain. V_{ds} is set to zero for simplicity. The energy diagram is similar to that of a conventional MOSFET at $V_g = 0$ in that a potential barrier stops the electrons in the source from entering the channel and the transistor is off. In the on state, Fig. 7–11c, channel E_c is pulled down by the gate voltage, but not at the source/drain edge, where the barrier height is fixed at ϕ_B (see Section 4.16). This barrier does not exist in a conventional MOSFET as shown in Fig. 7–11d, and they can degrade I_d of the metal S/D MOSFET.

To unleash the full potentials of Schottky S/D MOSFET, a very low-ϕ_B Schottky junction technology should be used (for NFETs). A thin N^+ region can be added between the metal and the channel. This minimizes the effect of the barriers on current flow as shown in Fig. 4–46. Attention must be paid to reduce the large reverse leakage current of a low-ϕ_{Bn} Schottky drain to body junction [9].

7.7 ● TRADE-OFF BETWEEN I_{on} AND I_{off} AND DESIGN FOR MANUFACTURING ●

Subthreshold I_{off} would not be a problem if V_t is set at a very high value. That is not acceptable because a high V_t would reduce I_{on} and therefore reduce circuit speed. Using a larger V_{dd} can raise I_{on}, but that is not acceptable either because it would raise the power consumption, which is already too large for comfort. Decreasing L can raise I_{on} but would also reduce V_t and raise I_{off}.

QUESTION • *Which, if any, of the following changes lead to both subthreshold leakage reduction and I_{on} enhancement? A larger V_t. A larger L. A smaller V_{dd}.*

Figure 7–12 shows a plot of log I_{off} vs. I_{on} of a large number of transistors [2]. The trade-off between the two is clear. Higher I_{on} goes hand-in-hand with larger I_{off}. The spread in I_{on} (and I_{off}) is due to a combination of unintentional manufacturing variances in L_g and V_t and intentional difference in the gate length.

Techniques have been developed to address the strong trade-off between I_{on} and I_{off}, i.e., between speed and standby power consumption.

One technique gives circuit designers two or three (or even more) V_ts to choose from. A large circuit may be designed with only the high-V_t devices first. Circuit timing simulations are performed to identify those signal paths and circuits where speed must be tuned up. Intermediate-V_t devices are substituted into them. Finally, low-V_t devices are substituted into those few circuits that need even more help with speed. A similar strategy provides multiple V_{dd}. A higher V_{dd} is provided to a small number of circuits that need speed while a lower V_{dd} is used in the other circuits. The larger V_{dd} provides higher speed and/or allows a larger V_t to be used (to suppress leakage). Yet the dynamic power consumption (see Eq. (6.7.6)) can be kept low because most of the circuits operate at the lower V_{dd}.

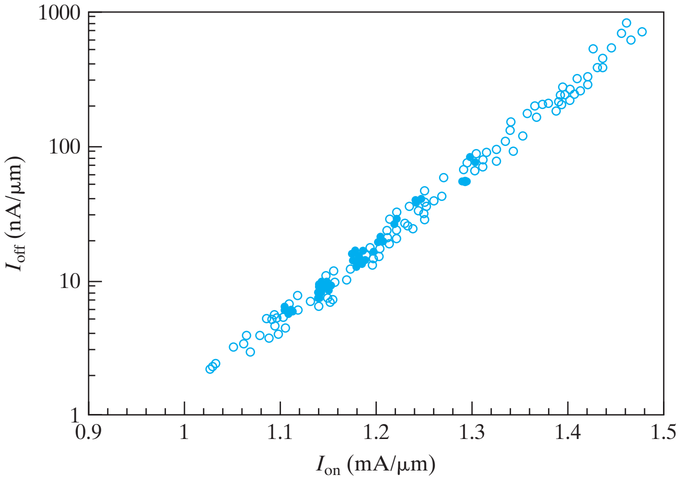


FIGURE 7–12 Log I_{off} vs. linear I_{on}. The spread in I_{on} (and I_{off}) is due to the presence of several slightly different drawn L_gs and unintentional manufacturing variations in L_g and V_t. (After [2]. © 2003 IEEE.)

In a large circuit such as a microprocessor, only some circuit blocks need to operate at high speed at a given time and other circuit blocks operate at lower speed or are idle. V_t can be set relatively low to produce large I_{on} so that circuits that need to operate at high speed can do so. A well bias voltage, V_{sb} in Eq. (6.4.6), is applied to the other circuit blocks to raise the V_t and suppress the subthreshold leakage. This technique requires intelligent control circuits to apply V_{sb} where and when needed.

This well bias technique also provides a way to compensate for the chip-to-chip and block-to-block variations in V_t that results from nonuniformity among devices due to inevitable variations in manufacturing equipment and process. Many techniques at the border between manufacturing and circuit design can help to ease the problem of manufacturing variations. These techniques are collectively known as **design for manufacturing** or **DFM**. A major cause of the device variations is the imperfect control of L_g in the lithography process. Some of the variation is more or less **random variation** in nature. The other part is more or less predictable, called **systematic variation**. One example of the systematic variations is the distortion in photolithography due to the interference of neighboring patterns of light and darkness. Elaborate mathematical optical proximity correction or OPC (see Section 3.3) reshapes each pattern in the photomask to compensate for the effect of the neighboring patterns. Another example is that the carrier mobility and therefore the current of a MOSFET is changed by the mechanical **stress effect** (see Section 7.1.1) created by nearby structures, e.g., shallow trench isolation or other MOSFETs. Sophisticated simulation tools can analyze the mechanical strain and predict the I_{on} based on the neighboring structures and feed the I_{on} information to circuit simulators to obtain more accurate simulation results. An example of random variation is the **gate edge roughness** or waviness caused by the graininess of the photoresist and the poly-crystalline Si. Yet another example of random variation is the **random dopant fluctuation** phenomenon. The statistical variation of the number of dopant atoms and their location in small size MOSFET creates significant variations in the threshold voltage. It requires complex design methodologies to include the intra-chip and inter-chip random variations in circuit design.

7.8 ● ULTRA-THIN-BODY SOI AND MULTIGATE MOSFETs ●

There are alternative MOSFET structures that are less susceptible to V_t roll-off and allow gate length scaling beyond the limit of conventional MOSFET. Figure 7–6 gives a simple description of the competition between the gate and the drain over the control of the channel barrier height shown in Fig. 7–5. We want to maximize the gate-to-channel capacitance and minimize the drain-to-channel capacitance. To do the former, we reduce T_{ox} as much as possible. To accomplish the latter, we reduce W_{dep} and X_j as much as possible. It is increasingly difficult to make these dimensions smaller. The real situation is even worse. In the subthreshold region, T_{ox} may be a small part T_{oxe} in Eq. (7.3.4) because the inversion-layer thickness, T_{inv} in Sec. 5.9, is large. Imagine that T_{ox} could be made infinitesimally small. This would give the gate a perfect control over the potential barrier height—but only right at the Si surface. The drain could still have more control than the gate along other leakage current paths that are some distance below the Si surface as shown in Fig. 7–13. At this submerged location, the gate is far away and the gate control is weak. The drain voltage can pull the potential

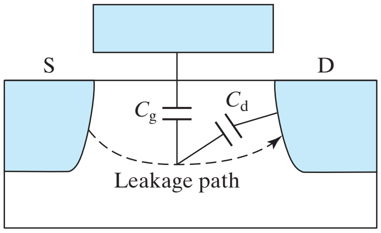


FIGURE 7–13 The drain could still have more control than the gate along another leakage current path that is some distance below the Si surface.

barrier down and allow leakage current to flow along this submerged path. There are two transistor structures that can eliminate the leakage paths that are far away from the gate [10]. One is called the **ultra-thin-body MOSFET** or **UTB MOSFET**. The other is **multigate MOSFET**. They are presented next.

7.8.1 Ultra-Thin-Body MOSFET and SOI

There are two ways to eliminate these submerged leakage paths. One is to use an ultra-thin-body structure as shown in Fig. 7–14 [11]. This MOSFET is built in a thin Si film on an insulator (SiO_2). Since the Si film is very thin, perhaps less than 10 nm, no leakage path is very far from the gate. (The worst-case leakage path is along the bottom of the Si film.) Therefore, the gate can effectively suppress the leakage. Figure 7–15 shows that the subthreshold leakage is reduced as the Si film is made thinner. It can be shown that the thin Si thickness should take the places of W_{dep} and X_j in Eq. (7.3.4) such that L_g can be scaled roughly in proportion to T_{Si}, the Si thickness. T_{Si} should be thinner than about one half of the gate length in order to reap the benefit of the UTB MOSFET concept to sustain scaling. UTB MOSFETs, as the multigate MOSFETs of the next section, offer additional device benefits. Because small l_d (Eq. (7.3.4)) can be obtained without heavy channel doping, carrier mobility is improved. The body effect that is detrimental to circuit speed (see Section 6.4) is eliminated because the body is **fully depleted** and floating and has no fixed voltage. One challenge posed by UTB MOSFETs is the large source/drain resistance due to their thinness. The solution is to thicken the source and drain with epitaxial deposition. These **raised source/drains** are visible in Figs. 7–14 and 7–15.

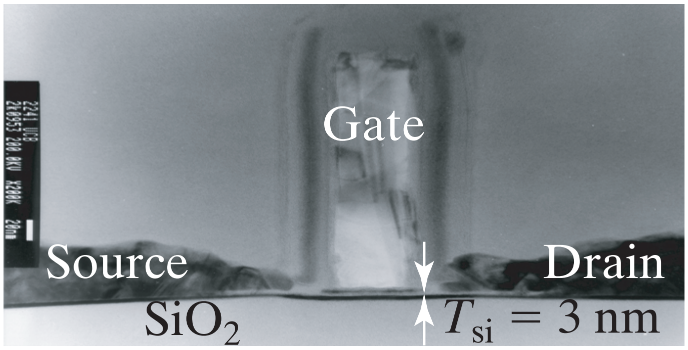


FIGURE 7–14 The SEM cross section of UTB device. (After [11]. © 2000 IEEE.)

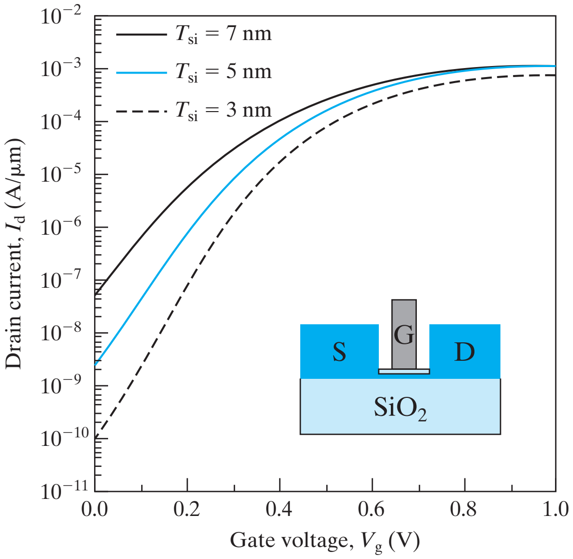


FIGURE 7–15 The subthreshold leakage is reduced as the Si film (transistor body) is made thinner. L_g = 15 nm. (After [11]. © 2000 IEEE.)

● SOI-Silicon on Insulator ●

Figure 7–16 shows the steps of making an SOI or **silicon-on-Insulator** wafer [12]. (The conventional wafer is sometimes called **bulk silicon** wafer for clarity.) Step 1 is to implant hydrogen into a silicon wafer that has a thin SiO_2 film at the surface. The hydrogen concentration peaks at a distance D below the surface. Step 2 is to place the first wafer, upside down, over a second plain wafer. The two wafers adhere to each other by the atomic bonding force. A low temperature annealing causes the two wafers to fuse together. Step 3 is to apply another annealing step that causes the implanted hydrogen to coalesce and form a large number of tiny hydrogen bubbles at depth D. This creates sufficient mechanical stress to break the wafer at that plane. The final step, Step 4, is to polish the surface. Now the SOI wafer is ready for use.

The Si film is of high quality and suitable for IC manufacturing. Even without using an ultra-thin body, SOI provides a speed advantage because the source/drain to body junction capacitance is practically eliminated as the source and drain diffusion regions extend vertically to the buried oxide. The cost of an SOI wafer is higher than an ordinary Si wafer and increases the cost of IC chips. For these reasons, only some microprocessors, which command high prices and compete on speed, have employed this technology so far. Figure 7–17 shows the cross-sectional SEMs of an SOI product. SOI also finds other compelling applications because it offers extra flexibility for making novel structures such as the ultra-thin-body MOSFET and some multigate MOSFET structures that can be scaled to smaller gate length beyond the capability of bulk MOSFETs.

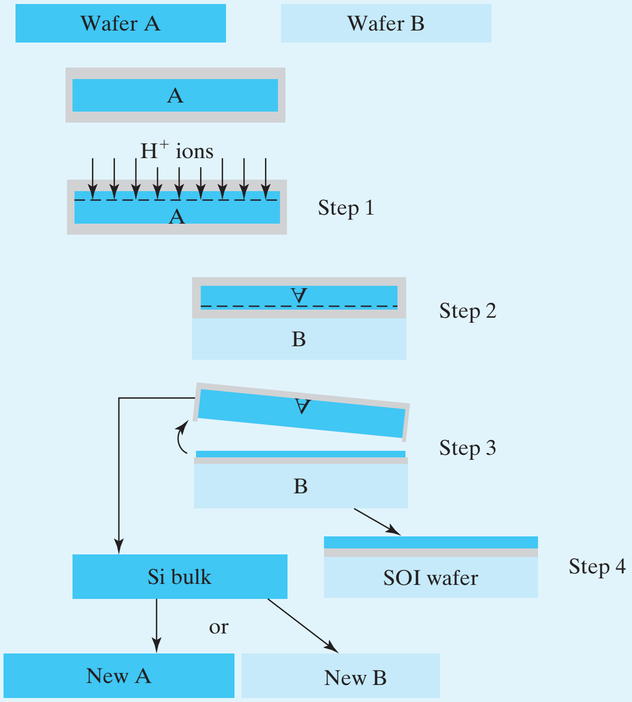


FIGURE 7–16 Steps of making an SOI wafer. (After [12].)

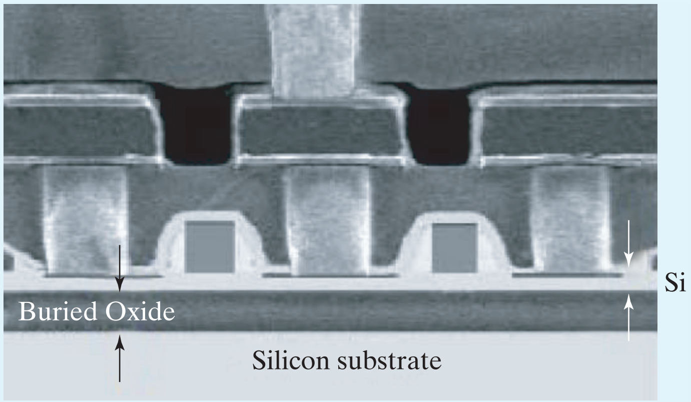


FIGURE 7–17 The cross-sectional electron micrograph of an SOI integrated circuit. The lower level structures are transistors and contacts. The upper two levels are the vias and the interconnects, which employ multiple layers of materials to achieve better reliability and etch stops.

7.8.2 FinFET - Multigate MOSFET

The second way of eliminating deep submerged leakage paths is to provide gate control from more than one side of the channel as shown in Fig. 7–18. The Si film is

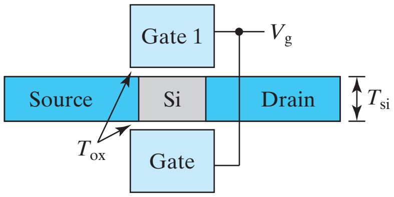


FIGURE 7–18 A schematic sketch of a double-gate MOSFET with gates connected.

very thin so that no leakage path is far from one of the gates. (The worst-case path is along the center of the Si film.) Therefore, the gate(s) can suppress leakage current more effectively than the conventional MOSFET. Because there are more than one gate, the structure may be called **multigate MOSFET**. The structure shown in Fig. 7–18 is a **double-gate MOSFET**. Shrinking T_{Si} automatically reduces W_{dep} and X_j in Eq. (7.3.4) and V_t roll-off can be suppressed to allow L_g to shrink to as small as a few nm. Because the top and bottom gates are at the same voltage and the Si film is fully depleted, the Si surface potential moves up and down with V_g mV for mV in the subthreshold region. The voltage divider effect illustrated in Fig. 7–1c does not exist and η in Eq. (7.2.4) is the desired unity and I_{off} is very low. There is no need for heavy doping in the channel to reduce W_{dep}. This leads to low vertical field and less impurity scattering; as a result the mobility is higher (see Section 6.3). Finally, there are two channels (top and bottom) to conduct the transistor current. For these reasons, a multigate MOSFET can have shorter L_g, lower I_{off}, and larger I_{on} than a single-gate MOSFET. But, there is one problem—how to fabricate the multigate MOSFET structure.

There is a multigate structure that is attractive for its simplicity of fabrication and it is illustrated in Fig. 7–19. Consider the center structure in Fig. 7–19. The process starts with an SOI wafer or a bulk Si wafer. A thin fin of Si is created by lithography and etching. Gate oxide is grown over the exposed surfaces of the fin. Poly-Si gate material is deposited over the fin and the gate is patterned by lithography and etching. Finally, source/drain implantation is

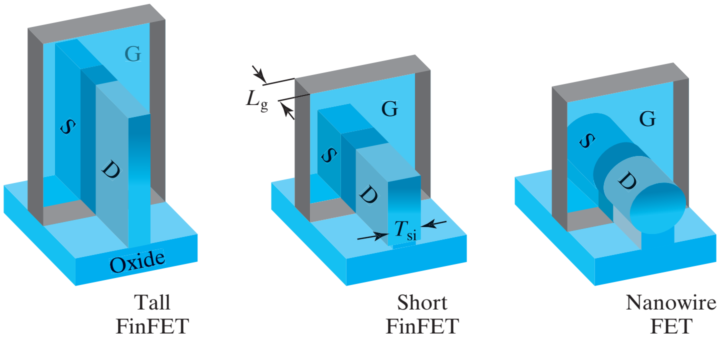


FIGURE 7–19 Variations of FinFET. Tall FinFET has the advantage of providing a large W and therefore large I_{on} while occupying a small footprint. Short FinFET has the advantage of less challenging lithography and etching. Nanowire FET gives the gate even more control over the transistor body by surrounding it. FinFETs can also be fabricated on bulk Si substrates.

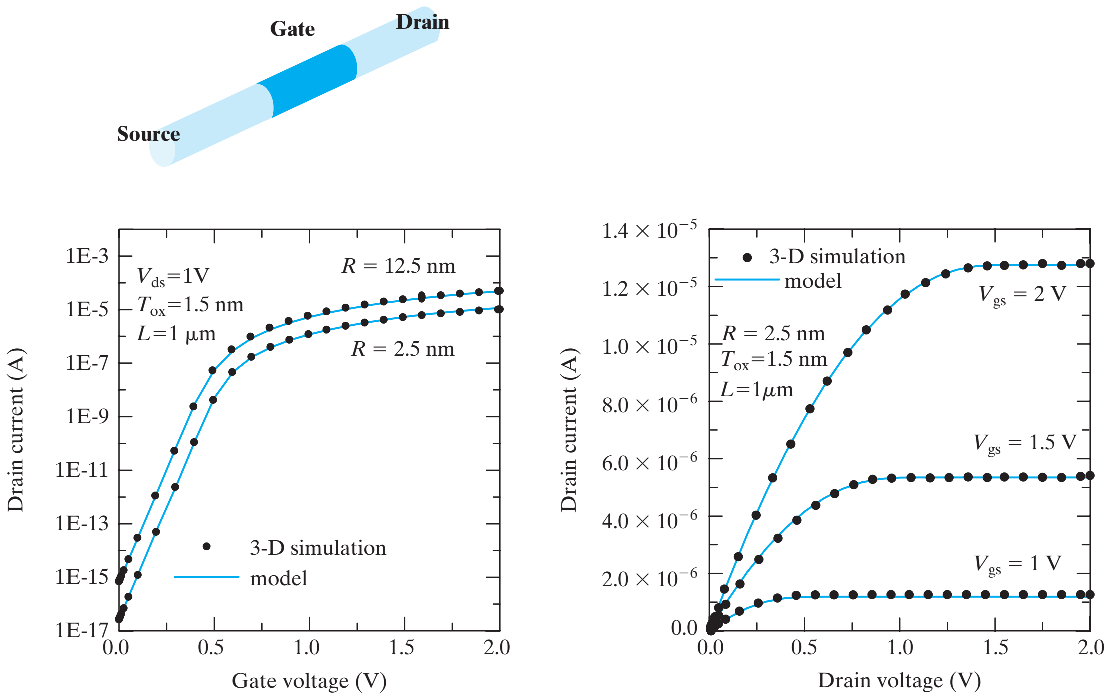


FIGURE 7–20 Simulated I–V curves of a nanowire MOSFET. R is the nanowire radius. (After [16].)

performed. The final structure in Fig. 7–19 is basically the multigate structure in Fig.7–18 turned on its side. This structure is called the **FinFET** because its Si body resembles the back fin of a fish [13]. The channel consists of the two vertical surfaces and the top surface of the fin. The channel width, W, is the sum of twice the fin height and the width of the fin.

Several variations of FinFET are shown in Fig. 7–19 [14,15]. A tall FinFET has the advantage of providing a large W and therefore large I_{on} while occupying a small footprint. A short FinFET has the advantage of less challenging etching. In this case, the top surface of the fin contributes significantly to the suppression of V_t roll-off and to leakage control. This structure is also known as a **triple-gate MOSFET**. The third variation gives the gate even more control over the Si wire by surrounding it. It may be called a nanowire FET and its behaviors shown in Fig. 7–20 can be modeled with the same methods and concepts used to model the basic MOSFETs. FinFETs with L_g as small as 3 nm have been experimentally demonstrated. It will allow transistor scaling beyond the scaling limit of the conventional planar transistor.

7.9 ● OUTPUT CONDUCTANCE ●

Output conductance limits the transistor voltage gain. It has been introduced in Section 6.13. However, its cause and theory are intimately related to those of V_t roll-off. Therefore, the present chapter is a fitting place to explain it.

What device design parameters determine the output conductance? Let us start with Eq. (6.13.1),

$$g_{\mathrm{ds}} \equiv \frac{\mathrm{d}I_{\mathrm{dsat}}}{\mathrm{d}V_{\mathrm{ds}}} = \frac{\mathrm{d}I_{\mathrm{dsat}}}{\mathrm{d}V_{\mathrm{t}}} \cdot \frac{\mathrm{d}V_{\mathrm{t}}}{\mathrm{d}V_{\mathrm{ds}}} \tag{7.9.1}$$

Since I_{ds} is a function of $V_{\mathrm{gs}} - V_{\mathrm{t}}$ [see Eq. (6.9.11)], it is obvious that

$$\frac{\mathrm{d}I_{\mathrm{dsat}}}{\mathrm{d}V_{\mathrm{t}}} = \frac{-\mathrm{d}I_{\mathrm{dsat}}}{\mathrm{d}V_{\mathrm{gs}}} = -g_{\mathrm{msat}} \tag{7.9.2}$$

The last step is the definition of g_{msat} given in Eq. (6.6.8). Now, Eq. (7.9.1) can be evaluated with the help of Eq. (7.3.3).

$$g_{\mathrm{ds}} = g_{\mathrm{msat}} \times \mathrm{e}^{-L/l_{\mathrm{d}}} \tag{7.9.3}$$

$$\text{Instrinsic voltage gain} = \frac{g_{\mathrm{msat}}}{g_{\mathrm{ds}}} = \mathrm{e}^{-L/l_{\mathrm{d}}} \tag{7.9.4}$$

Intrinsic voltage gain was introduced in Eq. (6.13.5). Equation (7.3.3) states that increasing V_{ds} would reduce V_{t}. That is why I_{ds} continues to increase without saturation. *The output conductance is caused by the drain/channel capacitive coupling, the same mechanism that is responsible for V_{t} roll-off.* This is why g_{ds} is larger in a MOSFET with shorter L. To reduce g_{ds} or to increase the intrinsic voltage gain, we can use a large L and/or reduce l_{d}. Circuit designers routinely use much larger L than the minimum value allowed for a given technology node when the circuits require large voltage gains. Reducing l_{d} is the job of device designers and Eq. (7.3.4) is their guide. Every design change that improves the suppression of V_{t} roll-off also suppresses g_{ds} and improves the voltage gain.

V_{t} dependence on V_{ds} is the main cause of output conductance in very short MOSFETs. For larger L and V_{ds} close to V_{dsat}, another mechanism may be the dominant contributor to g_{ds}—**channel length modulation**. A voltage, V_{ds}–V_{dsat}, is dissipated over a finite (non-zero) distance next to the drain. This distance increases with increasing V_{ds}. As a result, the effective channel length decreases with increasing V_{ds}. I_{ds}, which is inversely proportional to L, thus increases without true saturation. It can be shown that g_{ds}, due to the channel length modulation, is approximately

$$g_{\mathrm{ds}} = \frac{l_{\mathrm{d}} \cdot I_{\mathrm{dsat}}}{L(V_{\mathrm{ds}} - V_{\mathrm{dsat}})} \tag{7.9.5}$$

where l_{d} is given in Eq. (7.3.4). This component of g_{ds} can also be suppressed with larger L and smaller T_{ox}, X_{j}, and W_{dep}.

7.10 ● DEVICE AND PROCESS SIMULATION ●

There are commercially available computer simulation suites [17] that solve all the equations presented in this book with few or no approximations (e.g., Fermi–Dirac statistics is used rather than Boltzmann approximation). Most of these equations are solved simultaneously, e.g., Fermi–Dirac probability, incomplete ionization of dopants, drift and diffusion currents, current continuity equation, and Poisson equation. Device simulation is an important tool that provides the engineers with quick feedback about device behaviors. This narrows down the number of variables that need to be checked with expensive and time-consuming experiments. Examples of simulation results are shown in Figs. 7–15 and 7–20. Each of the figures takes from minutes to several hours of simulation time to generate.

Related to device simulation is process simulation. The input that a user provides to the process simulation program are the lithography mask pattern, implantation dose and energy, temperatures and times for oxide growth and annealing steps, etc. The process simulator then generates a two- or three-dimensional structure with all the deposited or grown and etched thin films and doped regions. This output may be fed into a device simulator together with the applied voltages and the operating temperature as the input to the device simulator.

7.11 • MOSFET COMPACT MODEL FOR CIRCUIT SIMULATION •

Circuit designers can simulate the operation of circuits containing up to hundreds of thousands or even more MOSFETs accurately, efficiently, and robustly. Accuracy must be delivered for DC as well as RF operations, analog as well digital circuits, memory as well as processor ICs. In circuit simulations, MOSFETs are modeled with analytical equations much like the ones introduced in this and the previous two chapters. More details are included in the model equations than this textbook can introduce. These models are called **compact models** to highlight their computational efficiency in contrast with the device simulators described in Section 7.10.

It could be said that the compact model (and the layout design rules) is the link between two halves of the semiconductor industry—technology/manufacturing on the one side and design/product on the other. A compact model must capture all the subtle behaviors of the MOSFET over wide ranges of voltage, L, W, and temperature and present them to the circuit designers in the form of equations. Some circuit-design methodologies, such as analog circuit design, use circuit simulations directly. Other design methodologies use **cell libraries**. A cell library is a collection of hundreds of small building blocks of circuits that have been carefully designed and characterized beforehand using circuit simulations.

At one time, nearly every company developed its own compact models. In 1997, an industry standard setting group selected **BSIM** [18] as the first industry standard model. If the I_{ds} equation of BSIM is printed out on paper, it will fill several pages.

Figure 7–21 shows selected comparisons of a compact model and measured device data to illustrate the accuracy of the compact model [19]. It is also important for the compact model to accurately model the transistor behaviors for any L and W that a circuit designer may specify. Figure 7–22 illustrates this capability. Finally, a good compact model should provide fast simulation times by using simple model equations. In addition to the IV of N-channel and P-channel transistors, the model also includes capacitance models, gate dielectric leakage current model, and source

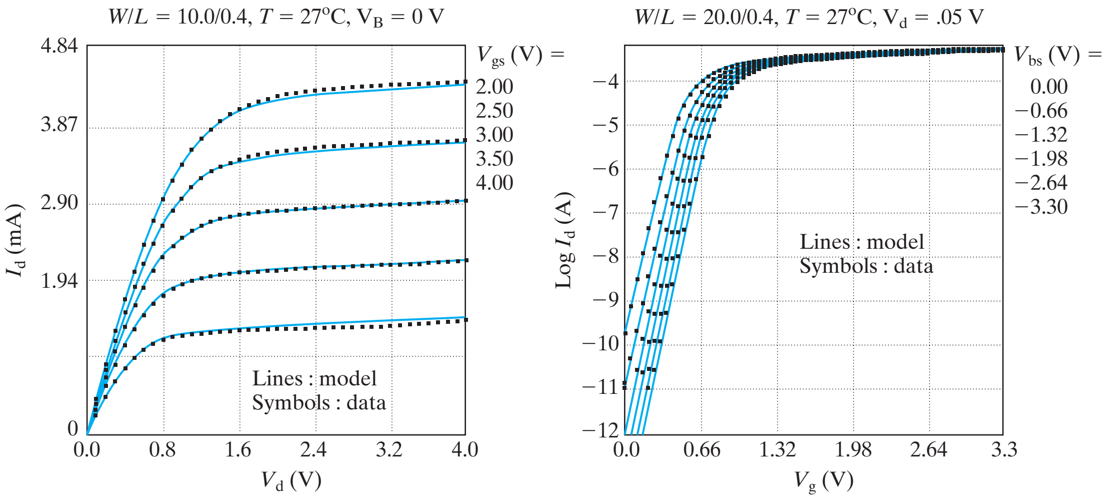


FIGURE 7–21 Selected comparisons of BSIM and measured device data to illustrate the accuracy of a compact model. (After [18].)

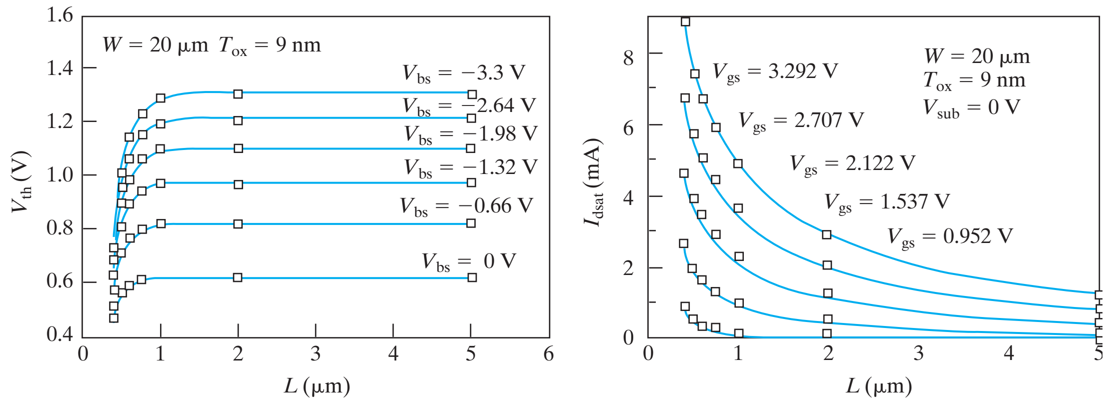


FIGURE 7–22 A compact model needs to accurately model the transistor behaviors for any L and W that circuit designers may specify. (After [19]. © 1997 IEEE.)

and drain junction diode model. Noise and high-frequency models are usually provided, too.

7.12 • CHAPTER SUMMARY •

To reduce cost and improve speed in order to open up new applications, transistors and interconnects are downsized periodically. Very small MOSFETs are prone to have excessive leakage current called I_{off}. The basic component of I_{off} is the **subthreshold current**

$$I_{\text{off}}(\text{nA}) = 100 \cdot \frac{W}{L} \cdot e^{-qV_t/\eta kT} = 100 \cdot \frac{W}{L} \cdot 10^{-V_t/S} \tag{7.2.8}$$

S is the **subthreshold swing.** To keep I_{off} below a given level, there is a minimum acceptable V_t. Unfortunately, a larger V_t is deleterious to I_{on} and speed. Therefore, it is important to reduce S by reducing the ratio $T_{\text{oxe}}/W_{\text{dep}}$. Furthermore, V_t decreases with L, a fact known as **V_t roll-off**, caused by DIBL.

$$V_t = V_{\text{t-long}} - (V_{ds} + 0.4\text{V}) \cdot e^{-L/l_d} \tag{7.3.3}$$

where $$l_d \propto \sqrt[3]{T_{\text{oxe}} W_{\text{dep}} X_j} \tag{7.3.4}$$

Since V_t is a sensitive function of L, even the small (a few nm) manufacturing variations in L can cause problematic variations in V_t, I_{off}, and I_{on}. To allow L reduction, Eq. (7.3.3) states that l_d must be reduced, i.e., $\boldsymbol{T_{\text{oxe}}}$, $\boldsymbol{W_{\text{dep}}}$, and/or $\boldsymbol{X_j}$ must be reduced.

T_{ox} reduction is limited mostly by **gate tunneling leakage**, which can be suppressed by replacing SiO_2 with a **high-*k* dielectric** such as HfO_2. Metal gate can reduce T_{oxe} by eliminating the poly-Si gate depletion effect.

W_{dep} can be reduced with retrograde body doping. X_j can be reduced with mS flash annealing or the metal source–drain MOSFET structure. X_j and W_{dep} can also be reduced with the ultra-thin-body SOI device structure or the multigate MOSFET structure. More importantly, these new structures eliminate the more vulnerable leakage paths, which are the farthest from the gate.

Equation (7.3.3) also provides a theory for output conductance of the short channel transistors.

$$g_{ds} = g_{\text{msat}} \times e^{-L/l_d} \tag{7.9.3}$$

● PROBLEMS ●

● Subthreshold Leakage Current ●

7.1 Assume that the gate oxide between an n+ poly-Si gate and the p-substrate is 11 Å thick and $N_a = 1\text{E}18\ \text{cm}^{-3}$.

(a) What is the V_t of this device?

(b) What is the subthreshold swing, S?

(c) What is the maximum leakage current if $W = 1\ \mu\text{m}$, $L = 18\ \text{nm}$? (Assume $I_{ds} = 100\ W/L$ (nA) at $V_g = V_t$.)

● Field Oxide Leakage ●

7.2 Assume the field oxide between an n+ poly-Si wire and the p-substrate is 0.3 µm thick and that $N_a = 5\text{E}17\ \text{cm}^{-3}$.

(a) What is the V_t of this field oxide device?

(b) What is the subthreshold swing, S?

(c) What is the maximum field leakage current if $W = 10\ \mu\text{m}$, $L = 0.3\ \mu\text{m}$, and $V_{dd} = 2.0\ \text{V}$?

● V_t Roll-off ●

7.3 Qualitatively sketch log(I_{ds}) vs. V_g (assume $V_{ds} = V_{dd}$) for the following:

(a) $L = 0.2\ \mu\text{m}$, $N_a = 1\text{E}15\ \text{cm}^{-3}$.

(b) $L = 0.2\ \mu\text{m}$, $N_a = 1\text{E}17\ \text{cm}^{-3}$.

(c) $L = 1\ \mu\text{m}$, $N_a = 1\text{E}15\ \text{cm}^{-3}$.

(d) $L = 1\ \mu\text{m}$, $N_a = 1\text{E}17\ \text{cm}^{-3}$.

Please pay attention to the positions of the curves relative to each other and label all curves.

● Trade-off between I_{off} and I_{on} ●

7.4 Does each of the following changes increase or decrease I_{off} and I_{on}? A larger V_t. A larger L. A shallower junction. A smaller V_{dd}. A smaller T_{ox}. Which of these changes contribute to leakage reduction without reducing the precious I_{on}?

7.5 There is a lot of concern that we will soon be unable to extend Moore's Law. In your own words, explain this concern and the difficulties of achieving high I_{on} and low I_{off}.

(a) Answer this question in one paragraph of less than 50 words.

(b) Support your description in (a) with three hand-drawn sketches of your choice.

(c) Why is it not possible to maximize I_{on} and minimize I_{off} by simply picking the right values of T_{ox}, X_j, and W_{dep}? Please explain in your own words.

(d) Provide three equations that help to quantify the issues discussed in (c).

7.6 **(a)** Rewrite Eq. (7.3.4) in a form that does not contain W_{dep} but contains V_t. Do so by using Eqs. (5.5.1) and (5.4.3) assuming that V_t is given.

(b) Based on the answer to (a), state what actions can be taken to reduce the minimum acceptable channel length.

7.7 **(a)** What is the advantage of having a small W_{dep}?

(b) For given L and V_t, what is the impact of reducing W_{dep} on I_{dsat} and gate? (Hint: consider the "m" in Chapter 6)

Discussion: Overall, smaller W_{dep} is desirable because it is more important to be able to suppress V_t roll-off so that L can be scaled.

● MOSFET with Ideal Retrograde Doping Profile ●

7.8 Assume an N-channel MOSFET with an N$^+$ poly gate and a substrate with an idealized retrograde substrate doping profile as shown in Fig. 7–23.

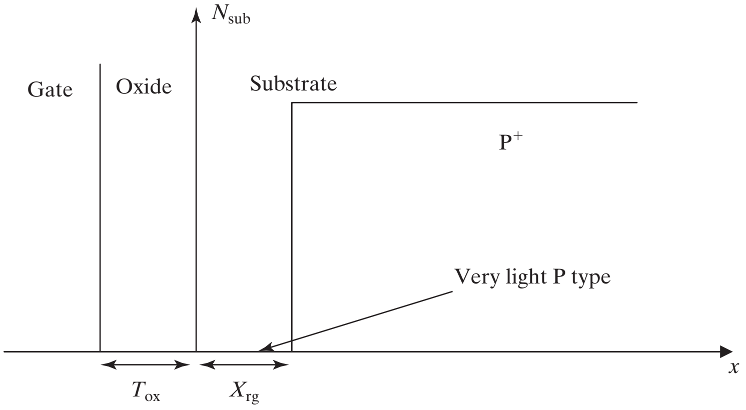


FIGURE 7–23

(a) Draw the energy band diagram of the MOSFET along the x direction from the gate through the oxide and the substrate, when the gate is biased at threshold voltage. (Hint: Since the P region is very lightly doped you may assume that the field in this region is constant or dε/dx = 0). Assume that the Fermi level in the P$^+$ region coincides with E_v and the Fermi level in the N$^+$ gate coincides with E_c. Remember to label E_c, E_v, and E_F.

(b) Find an expression for V_t of this ideal retrograde device in terms of V_{ox}. Assume V_{ox} is known. (Hint: Use the diagram from (a) and remember that V_t is the difference between the Fermi levels in the gate and in the substrate. At threshold, E_c of Si coincides with the Fermi level at the Si–SiO$_2$ interface).

(c) Now write an expression for V_t in terms of X_{rg}, T_{ox}, ε_{ox}, ε_{si} and any other common parameters you see fit, but not in terms of V_{ox}. Hint: Remember N_{sub} in the lightly doped region is almost 0, so if your answer is in terms of N_{sub}, you might want to rethink your strategy. Maybe $\varepsilon_{ox}\varepsilon_{ox} = \varepsilon_{si}\varepsilon_{si}$ could be a starting point.

(d) Show that the depletion layer width, W_{dep} in an ideal retrograde MOSFET can be about half the X_{dep} of a uniformly doped device and still yield the same V_t.

(e) What is the advantage of having a small W_{dep}?

(f) For given L and V_t, what is the impact of reducing W_{dep} on I_{dsat} and inverter delay?

● REFERENCES ●

1. International Technology Roadmap for Semiconductors (http://public.itrs.net/)

2. Ghani, T., et al. "A 90 Nm High Volume Manufacturing Logic Technology Featuring Novel 45 nm Gate Length Strained Silicon CMOS Transistors," *IEDM Technical Digest*. 2003, 978–980.

3. Yeo, Y-C., et al. "Enhanced Performance in Sub-100nm CMOSFETs Using Strained Epitaxial Si-Ge." *IEDM Technical Digest*. 2000, 753–756.

4. Liu, Z. H., et al. "Threshold Voltage Model for Deep-Submicrometer MOSFETs." *IEEE Trans. on Electron Devices*. 40, 1 (January 1993), 86–95.

5. Wann, C. H., et al. "A Comparative Study of Advanced MOSFET Concepts." *IEEE Transactions on Electron Devices*. 43, 10 (October 1996), 1742–1753.

6. Yeo, Yee-Chia, et al. "MOSFET Gate Leakage Modeling and Selection Guide for Alternative Gate Dielectrics Based on Leakage Considerations." *IEEE Transactions on Electron Devices*. 50, 4 (April 2003), 1027–1035.

7. Lu, Q., et al. "Dual-Metal Gate Technology for Deep-Submicron CMOS Transistor," Symp. on VLSI Technology Digest of Technical Papers, 2000, 72–73.

8. Chen, I. C., et al. "Electrical Breakdown in Thin Gate and Tunneling Oxides." *IEEE Trans. on Electron Devices*. ED-32 (February 1985), 413–422.

9. Kedzierski, J., et al. "Complementary Silicide Source/Drain Thin-Body MOSFETs for the 20 nm Gate Length Regime." *IEDM Technical Digest*, 2000, 57–60.

10. Hu, C. "Scaling CMOS Devices Through Alternative Structures," *Science in China (Series F)*. February 2001, 44 (1) 1–7.

11. Choi, Y-K., et al. "Ultrathin-body SOI MOSFET for Deep-sub-tenth Micron Era," *IEEE Electron Device Letters*. 21, 5 (May 2000), 254–255.

12. Celler, George, and Michael Wolf. "Smart Cut™ A Guide to the Technology, the Process, the Products," *SOITEC*. July 2003.

13. Huang, X., et al. “Sub 50-nm FinFET: PMOS.” *IEDM Technical Digest*, (1999), 67–70.

14. Yang, F-L, et al. “25 nm CMOS Omega FETs.” *IEDM Technical Digest*. (1999), 255–258.

15. Yang, F-L, et al. “5 nm-Gate Nanowire FinFET.” VLSI Technology, 2004. Digest of Technical Papers, 196–197.

16. Lin, C-H., et al. “Corner Effect Model for Compact Modeling of Multi-Gate MOSFETs.” 2005 SRC TECHCON.

17. Taurus Process, Synoposys TCAD Manual, Synoposys Inc., Mountain View, CA.

18. http://www-device.eecs.berkeley.edu/~bsim3/bsim4.html

19. Cheng, Y., et al. “A Physical and Scalable I-V Model in BSIM3v3 for Analog/Digital Circuit Simulation.” *IEEE Trans. on Electron Devices*. 44, 2, (February 1997), 277–287.

● GENERAL REFERENCES ●

1. Taur, Y., and T. H. Ning. *Fundamentals of Modern VLSI Devices*. Cambridge, UK: Cambridge University Press, 1998.

2. Wolf, S. *VLSI Devices*. Sunset Beach, CA: Lattice Press, 1999.