Ch. 7 MOSFET Technology Scaling, Leakage Current, and Other Topics

MOS ICs have met the world's growing needs for electronic devices for computing, communication, entertainment, automotive, and other applications with steady improvements in cost, speed, and power consumption. Such steady improvements in turn stimulate and enable new applications and fuel the growth of IC sales. There is now an entrenched expectation that this trend of rapid improvements will continue. How the MOSFET might continue to meet this expectation is the subject of this chapter. One overarching topic introduced in this chapter is the off-state current or the leakage current of the MOSFETs. This topic compliments the discourse on the on-state current presented in the previous chapter.

7.1 Technology Scaling—Small is Beautiful

Since the 1960's the price of one bit of semiconductor memory has dropped 100 million times and the trend continues. The cost of a logic gate has undergone a similarly dramatic drop. This rapid price drop has stimulated new applications and semiconductor devices have improved the ways people carry out just about all human activities. The primary engine the powered the ascent 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 in the improvements in speed and power consumption.

Gordon Moore made an empirical observation in the 1960's that the number of devices on a chip doubles every 18 months or so. The "Moore's Law" is a succinct description of the persistent periodic increase in the level 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 0.18µm, 0.13µm, 90nm, 65nm, 45nm…generations. The numbers refer to the minimum metal line width. Poly-Si gate length may be smaller. At each new node, the various feature sizes of circuit layout, such as the size of contact holes, are 70% of the previous node. This practice of periodic size reduction is called scaling. Historically, a new technology node is introduced every three years or so.

The main reward for introducing a new technology node is the reduction of circuit size by 2. (70% of previous line width means ~50% reduction in area, i.e. 0.7 x 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 is the engine that drives down the cost of ICs.

Besides 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, integrated circuit speed has increased roughly 30% at each new technology node.

Scaling does another good thing. Eq. 6.7.6 shows that reducing capacitance and, especially, the power supply voltage is effective for 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 (gasp) the doubling of transistors per chip at each technology node. If there had been no scaling, doing the job of a single PC microprocessor chip-- running 500M transistors at 2GHz using 1970 technology would require the electrical power output of a medium-size power generation plant.

In summary, scaling improves cost, speed, and power 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. Table 7.1 shows that scaling is expected to continue. But, what are the barriers to further scaling? Can scaling go on forever?

Table 7.1: Excerpt of 2003 ITRS technology scaling from 90nm to 22nm. The International Technology Roadmap for Semiconductors presents the industry's annually updated projection of future technologies and challenges [1]. HP:High Performance technology. LSTP: Low Standby Power technology for portable applications. EOT: Equivalent Oxide Thickness.

Year of Production	**2004**	**2007**	**2010**	**2013**	**2016**
Technology Node (nm)	90	65	45	32	22
HP physical Lg (nm)	37	25	18	13	9
EOT(nm) (HP/LSTP)	1.2/2.1	0.9/1.6	0.7/1.3	0.6/1.1	0.5/1.0
Vdd (HP/LSTP)	1.2/1.2	1.1/1.1	1.1/1.0	1.0/0.9	0.9/0.8
Ion/W,HP (mA/mm)	1100	1510	1900	2050	2400
Ioff/W,HP (mA/mm)	0.05	0.07	0.1	0.3	0.5
Ion/W,LSTP (mA/mm)	440	510	760	880	860
Ioff/W,LSTP (mA/mm)	1e-5	1e-5	6e-5	8e-5	1e-4

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 10mV? The answer is no.

At $V_{gs}<V_t$, an N-channel MOSFET is in the off-state. However, an undesirable leakage current can 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 even when it is in the standby mode. For example, if I_{off} is a modest 100nA per transistor, a cell-phone chip containing one hundred million transistors would consume so much standby current (10A) that the battery would be drained in minutes without receiving or transmitting any calls. A desk-top PC chip may be able to tolerate this static power but not much more before facing expensive problems with cooling the chip and the system.

Fig. 7-1 shows a typical subthreshold current plot. It is almost always plotted in a semilog I_{ds} versus V_{gs} graph. When V_{gs} is below V_t, I_{ds} is an exponential function of V_{gs}.

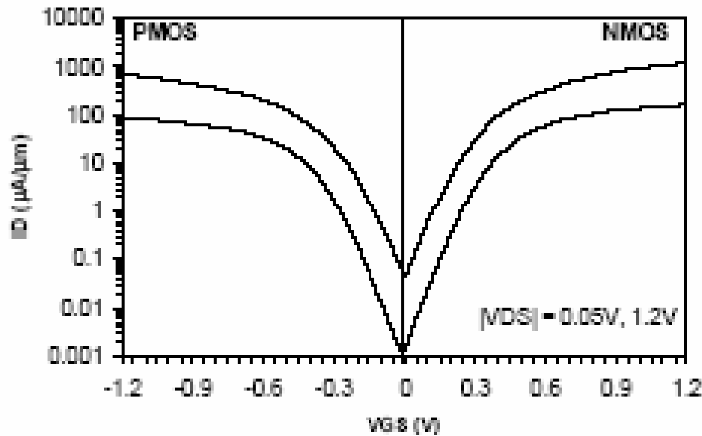


Figure 7-1 The current that flows at $V_{gs}<V_t$ is called the subthreshold current. V_t ~0.2V. The lower/upper curves are for Vds=50mV/1.2V. After Ref. [2].

Fig. 7-2 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. 7-2(a), a large 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. 7-2(b), 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

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

The practical definition of V_t in experimental studies is the V_{gs} at which I_{ds}=100nA $\times$W/L. (Some companies may use 200nA instead of 100nA.)[1]. Eq. (7.2.4) may be rewritten as

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

Clearly, Eq. (7.2.5) agrees with the definition of V_t and Eq. (7.2.4). Recall that 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$60mV. For example, if η=1.5, Eq. (7.2.5) states that I_{ds} drops by 10 times for every 90mV of decrease in V_{gs} below V_t. $\eta\times$60mV is called the **subthreshold swing** and represented by the symbol, S.

$$S = \eta \cdot 60mV \cdot \frac{T}{300} \tag{7.2.6}$$

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

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

1. The alternative shown in How to Measure the V_t of a MOSFET in Sec 6.4 is not applicable at large V_{ds}.

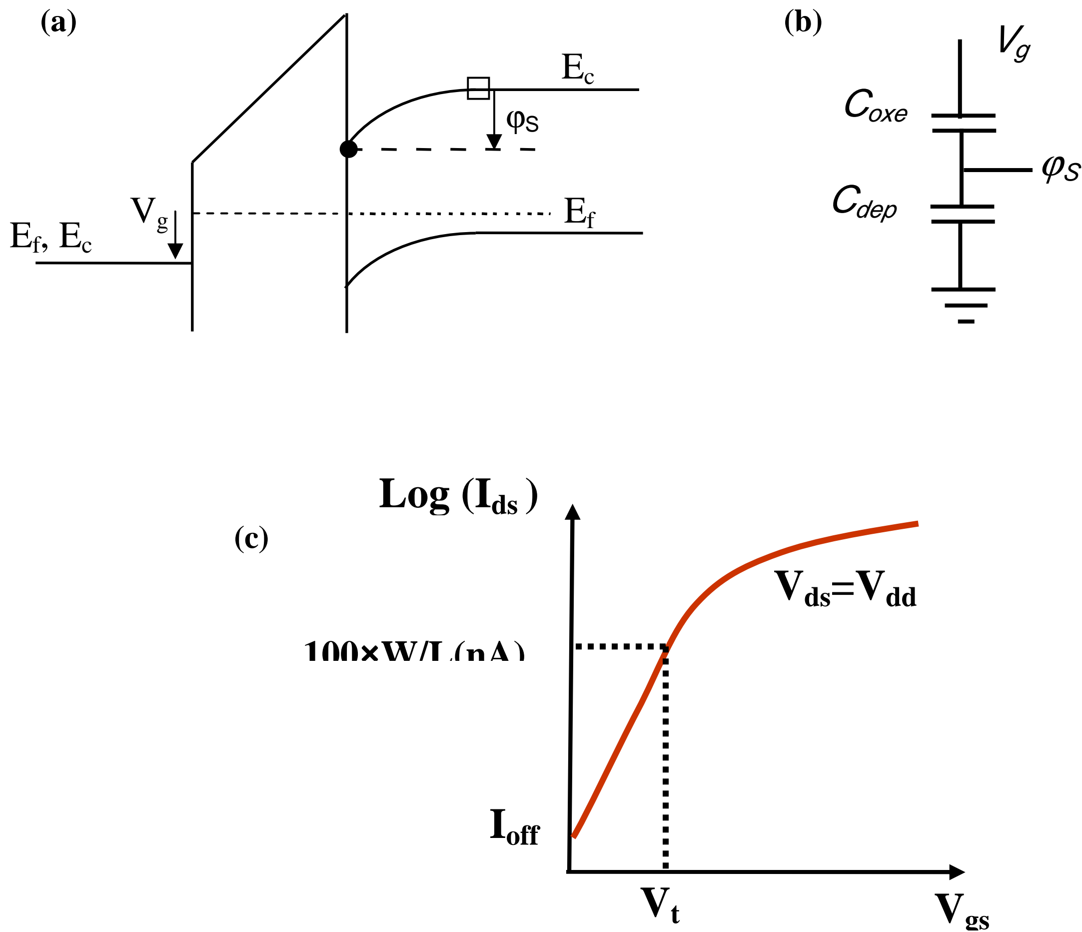


Figure 7-2: (a) When V_g is increased, E_c at the surface is pulled closer to E_f, causing n_s and I_{ds} to rise; (b) equivalent capacitance network; (c) Subthreshold IV with V_t and I_{off}.

For given W and L, there are two ways to minimize I_{off} illustrated in Fig. 7-2 (c). The first is to choose a large V_t. This is not desirable because a large V_t reduces I_{on} and therefore increases the gate delays (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}.[2] An additional way to reduce S, and therefore to reduce I_{off}, is to operate the transistors at a lower temperature. This last approach is valid in principal but rarely used because cooling adds considerable cost.

2. 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, ηand m are closely related.

Example: Subthreshold Leakage Current

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

Answer:

A.) Use Eq. 7.2.6.

$$I_{off}(nA) = 100 \cdot \frac{W}{L} \cdot 10^{-V_t/S} = 100 \cdot \frac{10}{0.05} \cdot 10^{-0.34/0.085} = 2nA$$

B.) Use Eq. 7.2.7.

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

7.3 V_t Roll-off --- Short-channel MOSFETs are Hard to Turn Off

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. 7-3. When V_t drops too much, I_{off} becomes too large and that channel length is not acceptable.

Sidebar: Gate Length (L_g) versus Channel Length (L) and Experimental Data versus Equations

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 (called **fab** in short). The channel length, in comparison, can not be determined accurately due to the lateral diffusion of the source and drain junctions. L tracks L_g well but the difference between the two just can not be quantified precisely. As a result, L_g is widely used in lieu of L in data collection and presentations such as in Fig. 7-3. L is used in theoretical equations but it is understood that L can not be known precisely for small real transistors. Thus we rely on measured data and complex computer simulations of

devices for precise device development and circuit design. On the other hand, we rely on the theoretical equations to guide the interpretation of the data, the design of new experiments, and the search for new innovative ideas.

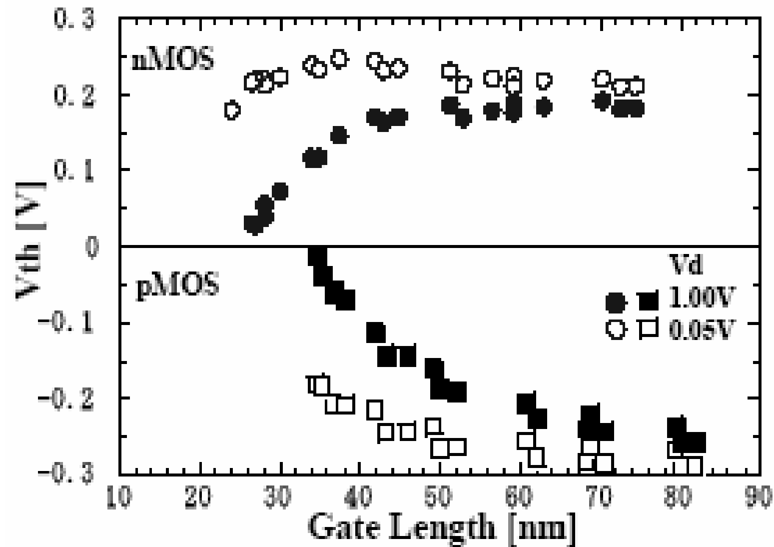


Figure 7-3 V_t decreases with decreasing L_g. This phenomenon is called V_t roll-off. It determines the minimum acceptable L_g because I_{off} is too large when V_t falls too low. After Ref. [3].

At certain L_g, V_t becomes so low that I_{off} becomes unacceptable (see Eq. 7.2.8). Device development engineers must design the device so that the V_t roll-off does not prevent the use of the targeted minimum-L_g, for example those listed in the third row of Table 7-1. Of course, lithography resolution must able to support the L_g targets, too.

Why does V_t decrease with L? Fig. 7-4 provides the answer. Fig. 7-4(a) shows the energy-band diagram along the semiconductor/insulator interface of a long channel device at V_{gs}=0. Fig. 7-4(b) shows the case at $V_{gs}=V_t$. In the case of (b), E_c in the channel is pulled lower than in case (a) and therefore is closer to the E_c in the source. When the channel Ec is only ~0.2eV 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 is reached. We may say that a 0.2eV potential barrier is low enough to allow the electrons in the N^+ source to flow into the channel and then into the drain. 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.

Fig. 7-4(c) 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.7-4(a). As a result, a smaller V_{gs} is needed in Fig. 7-4(d) than in Fig. 7-4(b) to pull the barrier down to 0.2eV. In other words, V_t is lower in case (d), the short channel device than in case (b), the long channel device. This explains the V_t roll-off shown in Fig. 7-3.

We can understand V_t roll-off from another approach. Fig. 7-5 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 the location where E_c peaks in Fig. 7-4(d). As the channel length is reduced, the drain to source and drain to "channel" distance is reduced; therefore C_d increases. Do not be concerned with the exact definition or value of C_d. Just remember that it represents the strength of capacitive coupling in the complex two-dimensional structure of the drain and the channel.

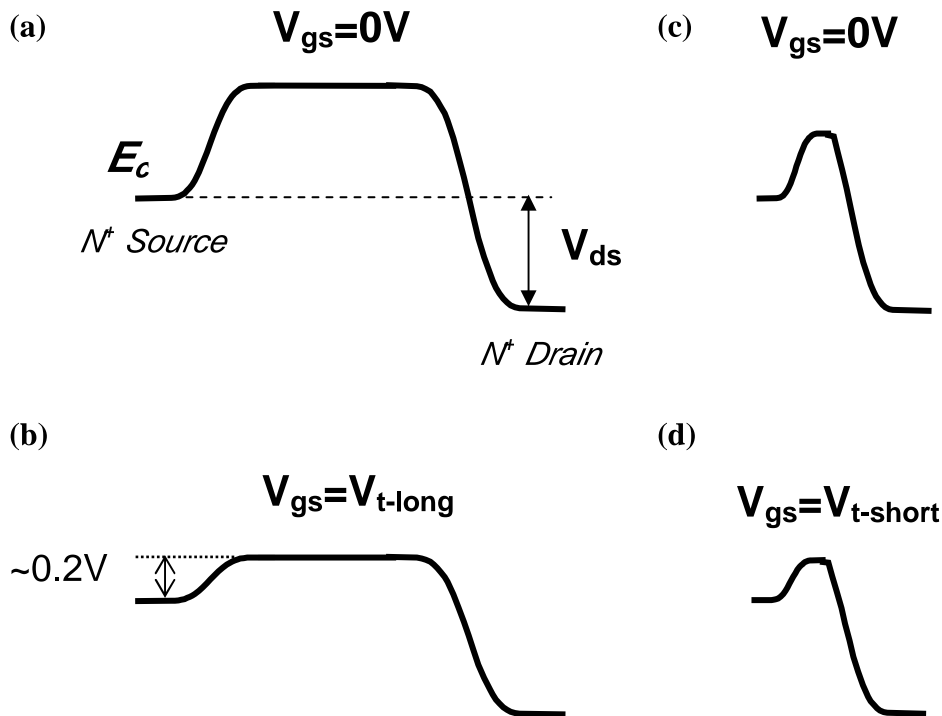


Figure 7-4 (a)-(d): Energy-band diagram from source to drain when $V_{gs}=0V$ and $V_{gs}=V_t$. (a)-(b) long channel; (c)-(d) short channel.

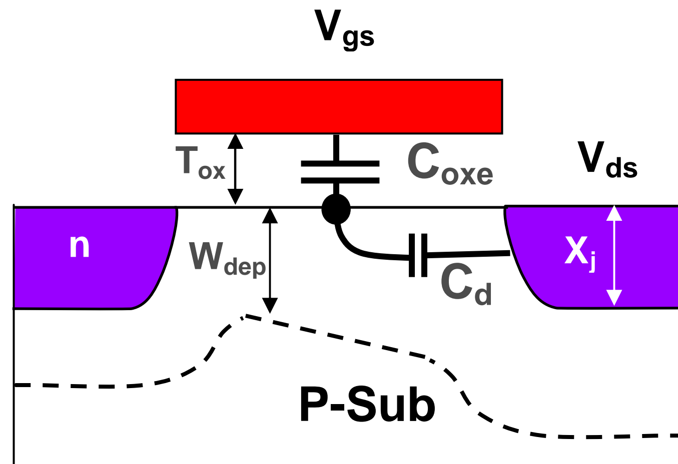


Figure 7-5 Schematic two-capacitor network in MOSFET. C_d models the electrostatic coupling between the channel and the drain. As the channel length is reduced, drain to "channel" distance is reduced; therefore C_d increases.

From this two-capacitor equivalent circuit, one immediately sees that the drain voltage has a similar effect as the gate voltage on the channel potential. V_{gs} and V_{ds}, together, determine the channel potential barrier height shown in Fig. 7-4. When V_{ds} is present, less V_{gs} is needed to pull the barrier down to 0.2eV, therefore V_t is lower by definition. This understanding gives us a simple equation for V_t roll-off,

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

where V_t-long is the threshold voltage of a long-channel transistor, for which C_d=0. More exactly, V_{ds} should be supplemented with a constant that represent the effect of the built-in potentials between the N^- channel and the N^+ drain and source, about 0.4V [4].

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

Using Fig. 7-5, 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 has been confirmed with 2-dimensioal computer simulations and analytical solutions of the Poisson equation. These analyses further indicate that C_d is an exponential function of L in this two-dimensional structure [5]. Therefore,

$$V_t = V_{t-long} - (V_{ds} + 0.4) \cdot e^{-L/l_d} \quad , \qquad (7.3.3)$$

where $l_d \propto \sqrt[3]{T_{ox} W_{dep} X_j}$ (7.3.4)

X_j is the drain junction depth. Eq. 7.3.3 provides a semi-quantitative model of the roll-off of V_t as a function of L and V_{ds}. At a very large L, V_t is equal to V_t-long as expected. The roll-off is an exponential function of L. The roll-off is also larger at larger V_{ds}, and the worst case is $V_{ds}=V_{dd}$. I_{off} becomes unacceptable when V_t is too small. This condition determines the minimum acceptable L. **The minimum acceptable L is several times of *l_d*.** 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 shields the channel from the drain.

One way to summarize the message of Eq. 7.3.4 is that **vertical dimensions in a MOSFET (T_{ox}, W_{dep}, X_j) must be reduced in order to support the reduction of gate length**.

7.4 Reducing the Gate Insulator Thickness and T_{oxe}

SiO_2 has been the preferred gate insulator for silicon MOSFET since its very beginning in the 1960's and the oxide thickness has been reduced over the years from 300nm for 10μm technology to 1.2nm for 65nm 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}. A large I_{on} is desirable for maximizing 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 falling L according to Eqs. 7.3.3 and 7.3.4. One must not underestimate the importance of the second reason. Fig. 7-6 shows that the oxide thickness has been scaled roughly in proportion to the line width.

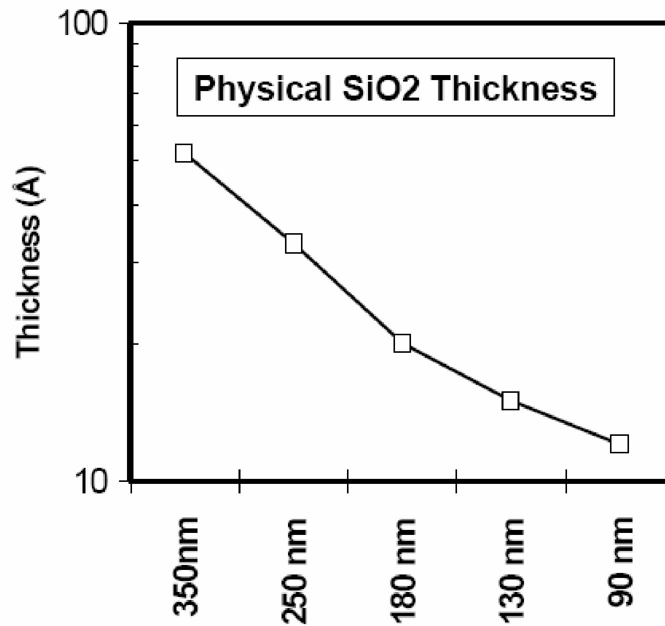


Figure 7-6: Oxide thickness has been scales roughly in proportion to the line width.

So, 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 atomic bonds at the Si/SiO_2 interface thus creating oxide charge and V_t shift (see Sec. 5.7). V_t shifts cause circuit behaviors to change and raise reliability concerns.

For SiO_2 films thinner than 1.5nm, tunneling leakage current becomes the most serious limiting factor. Fig. 7-7(a) illustrates the phenomenon of tunneling. Fig. 7-7(b) shows that the very rapid rise of the SiO_2 leakage current with decreasing thickness agrees with the tunneling model prediction [6]. At 1.2nm, SiO_2 leaks 10^3 A/cm^2. If an IC chip contains 1mm^2 total area of this thin dielectric, the chip oxide leakage current would be 10A. This large leakage would drain the battery of a cell phone in minutes.

Researchers are developing high-k dielectrics to replace SiO_2. For example, HfO_2 has a relative dielectric constant (k) of ~24, six times large than that of SiO_2. A 6nm thick HfO_2 film is equivalent to 1nm 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 1nm. However, the HfO_2 film presents a much thicker (albeit a lower) tunneling barrier to the electrons and holes. The consequence is a leakage current that is several orders of magnitude smaller than that in SiO_2 as shown in Fig. 7-7(b) and (c). A metal gate is used

to reduce the poly-Si gate depletion and EOT in 7-7(c) [7]. Other candidates of high-k gate dielectric include ZrO_2 and Al_2O_3. The difficulties of adopting high-k dielectrics in IC manufacture insclude chemical reactions between them and the silicon substrate and gate, lower surface mobility than the Si/SiO_2 system. These problems can be minimized by inserting a thin SiO_2 interfacial layer between the silicon substrate and the high-k dielectric and using a metal gate instead of a poly-Si gate.

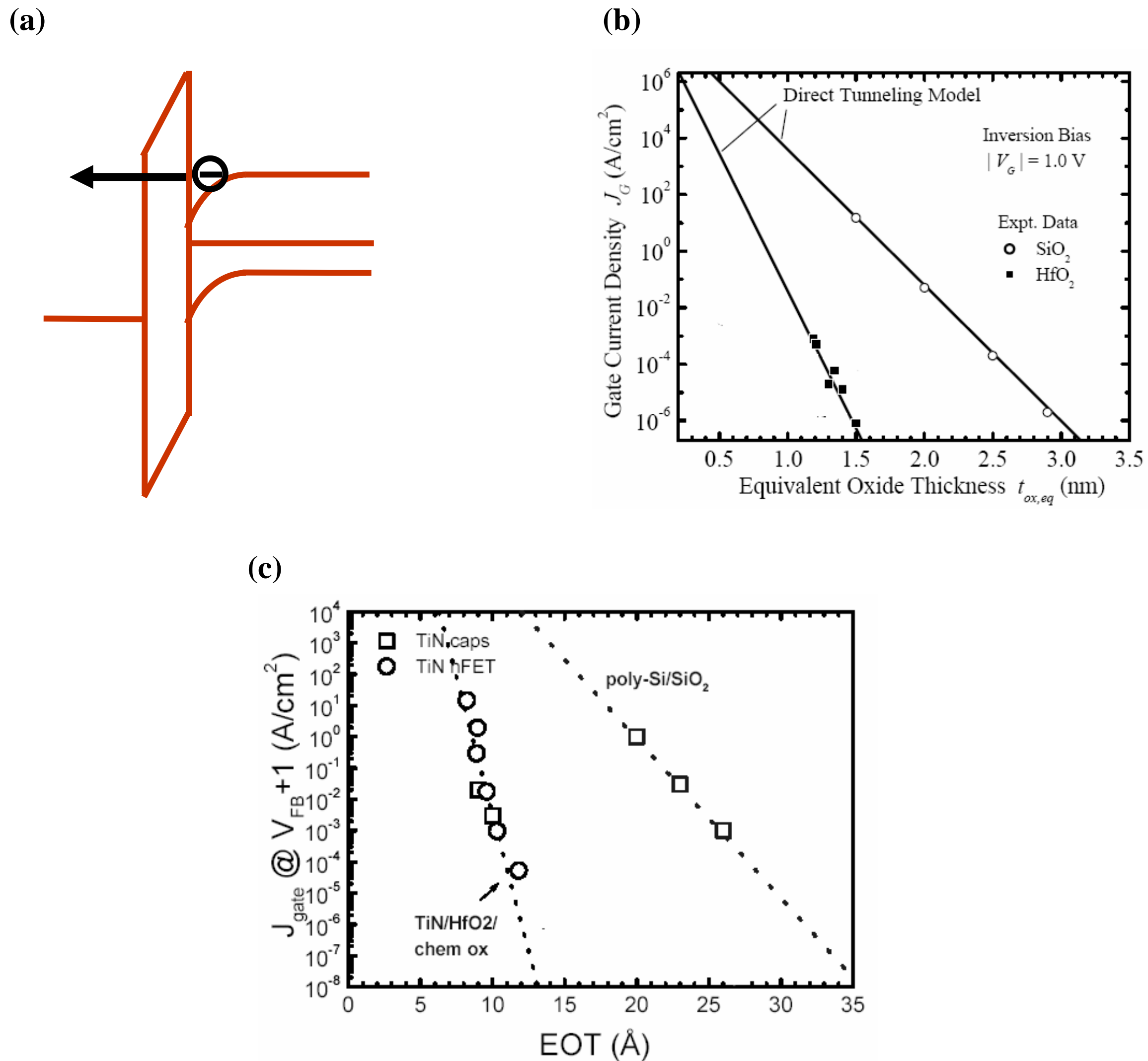


Figure 7-7 (a)-(c): (a) Energy band diagram in inversion showing electron tunneling path through the gate oxide. 1.2 nm SiO_2 conducts 10^3 A/cm^2 of leakage current. High-k dielectric such as HfO_2 has several orders lower leakage; After Ref. [6] (c) HfO_2 is used with a TiN metal gate and a thin SiO_2 at the Si interface produced by wet chemistry. After Ref. [7]

Note that Eq. 7.3.4 contains the electrical oxide thickness, T_{oxe}, defined in Eq. 5.9.2. Besides T_{ox} or EOT in the case of high-k dielectric, poly-Si gate depletion layer also needs to be minimized. A metal gate would be the ultimate gate material in this respect. The challenge there is to find metals that have work functions close to those of N^+ and P^+ poly-Si.

In addition, T_{inv} needs to be minimized. The material parameters that determine T_{inv} is the electron and hole effective masses. A larger effective mass leads to a thinner T_{inv}. Unfortunately, a larger effective mass leads to a lower mobility. Fortunately, the effective mass is a function of the spatial direction of carrier motion in a crystal. The effective mass in the direction normal to the channel determines T_{inv}, while the effective mass in the plane of the channel determines the surface mobility, μ_s. It may be possible to choose a semiconductor and a wafer surface orientation (see Fig. 1-2) that together produce large m_n and m_p normal to the channel and small m_n and m_p in the plane of the channel.

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 stress time. If a one second (1s) voltage pulse is applied to a 10nm SiO_2 film, 15V is needed to breakdown the film for a breakdown field of 15MV/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 one month. This phenomenon is called **time-dependent dielectric breakdown**. Many IC applications require a device lifetime of 10 years. Clearly, manufacturers can not afford the time to actually measure the 10 yr breakdown fields for new oxide technologies. Instead, researchers have predicted the 10 yr breakdown fields based on short-term tests in combination with theoretical models of the physics of oxide breakdown. In retrospect, the most optimistic of the predictions, 7MV/cm for 10year operation, was basically right, and SiO_2 thickness has been scaled further than the other models predicted possible [8].

This breakdown model suggests that carrier tunneling at high field introduces holes into SiO_2. Holes cause the break-up of the weaker Si-O bonds in amorphous SiO_2 thus creating oxide defects. This process progresses more rapidly at random 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 stress time. A fortuitous fact is that the breakdown field increases somewhat with decreasing oxide thickness. The reason is that a larger fraction of holes may escape the thinner film without generating defects; therefore a higher field can be tolerated.

7.5 How to Reduce W_{dep}

Eq. 7.3.4 suggests that a small W_{dep} helps to control Vt 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,

$$V_t = V_{fb} + 2\phi_B + \frac{\sqrt{qN_{sub}2\varepsilon_s 2\phi_B}}{C_{ox}} \qquad 7.5.1$$

tells us 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. It can be shown that W_{dep} can only be reduced in proportion to T_{ox}.

$$V_t = V_{fb} + 2\phi_B + \frac{2\varepsilon_s 2\phi_B}{C_{ox}W_{dep}} \qquad 7.5.2$$

This fact further highlights the importance of reducing 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 this case, W_{dep} is determined by the thickness of the lightly doped surface layer. It can be shown that V_t of an MOSFET with ideal retrograde doping is

$$V_t = \frac{E_g}{q} - 0.1 + \left(\frac{E_g}{q} - 0.1\right)\frac{\varepsilon_{si}T_{ox}}{\varepsilon_{oxi}L_{rg}} \qquad 7.5.3$$

where L_{rg} is the thickness of the lightly doped thin layer. The derivation of Eq. 7.5.3 is left as an exercise for the interested readers in the Problems at the end of the chapter. Again, W_{dep} (=L_{rg}) can only be scaled in proportion to T_{ox} if V_t is to be kept constant. However, W_{dep} in an ideal retrograde device can be about half the X_{dep} of a uniformly doped device 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 Sec. 2.2.2) in the inversion layer can be reduced and surface mobility can be higher. However, dopant diffusion makes it difficult to fabricate a retrograde profile with a very thin lightly doped layer, i.e. a very small W_{dep} unless process temperature is further lowered.

Predicting the Ultimate Low Limit of Channel Length – A Retrospective

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

In the 1970's the consensus in the semiconductor industry was that the ultimate lower limit of channel length is 500nm. In the 80's, the consensus was 250nm. In the 90's, it was 100nm. Now it is shorter than 10nm. What made the most knowledgeable experts in the industry and universities underestimate how far channel length can be scaled?

A review of the historical literature reveals that the researchers were mistaken about the lower limit of gate oxide thickness. In the 70's it was thought ~15nm would be the limit. In the 80's, it was 8nm, and so on. Since the T_{ox} estimate was off, the estimats of the minimum acceptable W_{dep} and therefore the minimum L would be off according to Eq. 7.3.4.

Here is an intriguing note about W_{dep} scaling. A higher N_{sub} in Eq. 7.5.1 (and therefore a smaller W_{dep}) is allowable if V_t is allowed to be larger. This larger V_t can be brought back down with a body (or well) to source bias voltage, Vbs. The required V_{bs} is a forward bias across the body-source junction. The forward bias is acceptable, i.e. the forward bias current is small as long as V_{bs} is kept below 0.6V.

7.6 Shallow Junction Technology

Fig. 7-8, first introduced as Fig. 6-24(b), 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 short annealing at the lowest necessary temperature is used to activate the dopants and anneal out the implantation damages in the crystal. Because dopant diffusion can not be totally avoided, the doping concentration in the shallow junction extension must be kept low (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. 7-8, a deeper N^+ junction is used to minimize total parasitic resistance. However, even the depth of the N^+ junction should be kept shallow to help V_t roll-off.

One possible way to beat the tradeoff between the junction depth and low series resistance is to replace the shallow junction extension with a thin layer of metal or silicide. This is theoretically possible but the metal or silicide must be chosen such that there is not a large energy barrier (see Ch. 9) between it and the silicon channel and not a large leakage between it and the substrate [9].

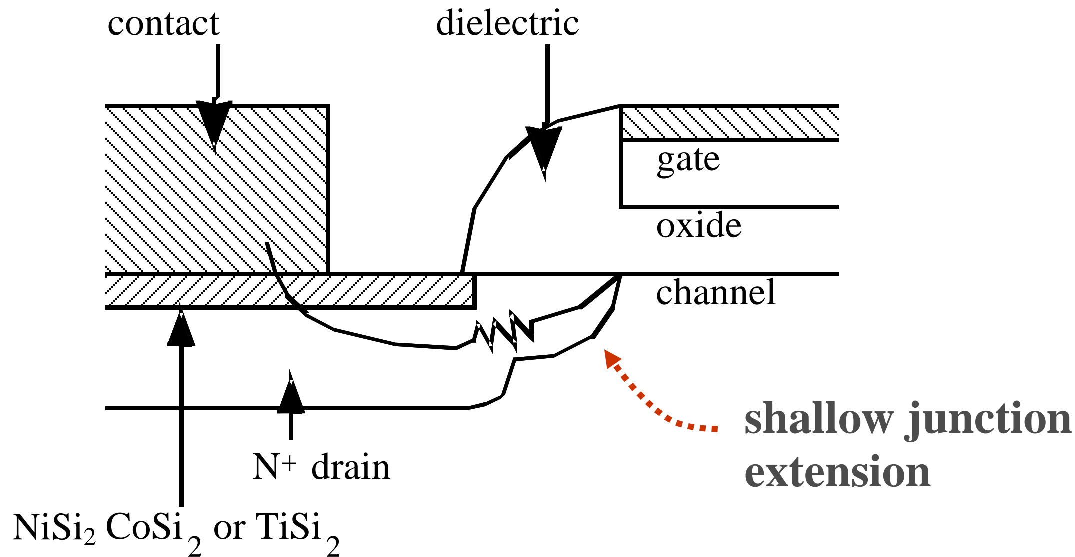


Fig. 7-8 Cross-sectional view of a MOSFET drain junction. The shallow junction extension next to the channel helps to suppress the V_t roll-off.

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

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 an acceptable solution because a larger V_{dd} would raise the power consumption, which is already too large for comfort. Most other changes that could reduce the leakage would also hurt I_{on}.

The salient exception is to use a smaller T_{ox}. That improves both I_{on} and V_t roll-off. Unfortunately, even T_{ox} reduction is no longer a cure without a serious side effect. In fact, the side effect--large dielectric tunneling leakage--has made SiO_2 thickness reduction beyond 1nm more harmful than helpful.

Question: *Does any of the following changes contribute to both leakage reduction and I_{on} enhancement? A larger V_t. A larger L. A shallower junction. A smaller V_{dd} (hint: the worst case V_{ds} in Eq. 7.3.3 is V_{dd}).*

Fig. 7-9 shows a plot of Log I_{off} versus I_{on} [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 intentional difference in drawn gate length.

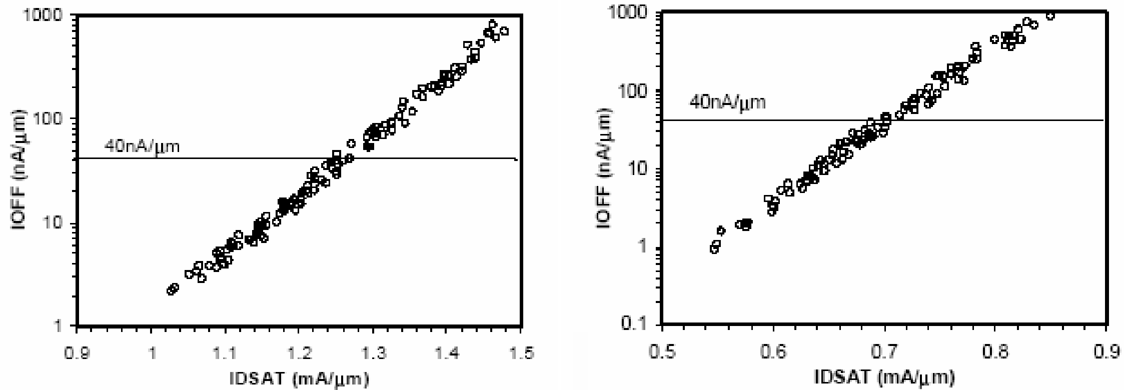


Figure 7-9 Log I_{off} versus I_{on}. The spread in I_{on} (and I_{off}) is due to a combination of intentional differences and unintentional variances in L_g. After Ref. [2]

There are several techniques at the border between device technology and circuit design that can help to relax the conflict between I_{on} and leakage. 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 substrate or 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 technique is practical and often used. It also provides a way to compensate for the chip-to-chip and block-to-block variations in V_t that results from non-uniformity among devices due to imperfect manufacturing equipment and process. An interesting alternative is to apply a forward source-body bias to reduce V_t when and where high speed is needed. If the forward bias is lower than 0.6V, the diode forward current is acceptable due to the small junction area. The advantage of this alternative is that W_{dep} is reduced by the forward bias and V_t roll-off is improved (see Eq. 7.3.4).

Another technique gives circuit designers two or three (or even more) V_t to choose. 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} rather than multiple V_t. 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. This would allow a relatively large V_t to be used (to suppress leakage).

Finally, there are alternative MOSFET structures that provide superior tolerance for gate length scaling. They are introduced in the next section.

7.8 More Scalable Device Structures [3]

Fig. 7-5 gives a simple description of the competition between the gate and the drain over the control of the channel barrier height shown in Fig. 7-4. 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. Assume 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 silicon surface. The drain could still have more control than the gate along other leakage current path that is some distance below the silicon surface as shown in Fig. 7-10. At this submerged location, the gate is far away and the gate control is weak. The drain voltage can pull the potential barrier down and allow leakage current to flow along this submerged path (Fig. 7-11).

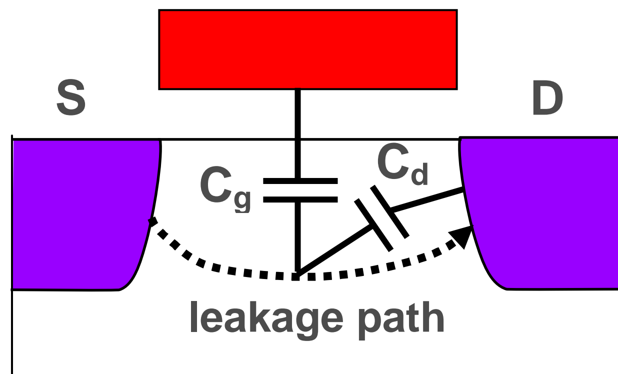


Figure 7-10 The drain could still have more control than the gate along another leakage current path that is some distance below the silicon surface.

3. This section may be omitted in an accelerated course.

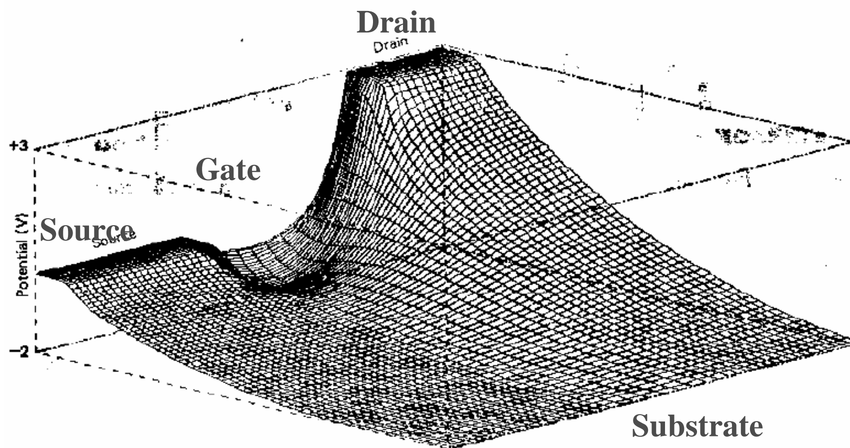


Figure 7-11 The drain voltage can pull the potential barrier down and allow leakage current to flow along a submerged path. After Ref. [10].

7.8.1 Ultra-Thin-Body MOSFET

There are two ways to eliminate these submerged leakage paths. One is to use an ultra-thin-body structure as shown in Fig. 7-12 [11]. This MOSFET is built in a thin silicon film on an insulator (SiO_2). Since the silicon film is very thin, perhaps less than 10nm, no leakage path is very far from the gate. (The worst case path is along the bottom of the silicon film.) Therefore the gate can effectively suppress the leakage. Fig. 7-13 shows that the subthreshold leakage is reduced as the silicon film is made thinner. Another benefit of this structure is that the thin silicon thickness automatically provides a shallow junction. Experiments and simulations have shown that the silicon film should be not much thicker than ¼ the gate length.

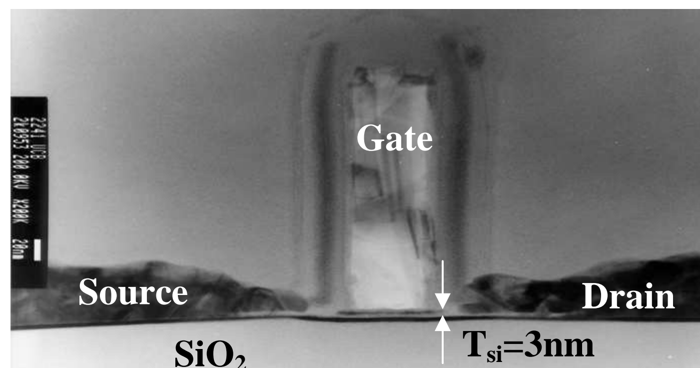


Figure 7-12. The SEM cross-section of UTB device. After Ref . [11]

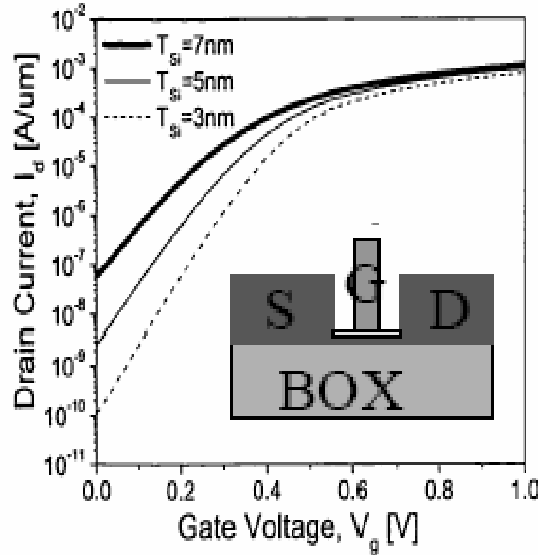


Figure 7-13 The subthreshold leakage is reduced as the silicon film is made thinner. L_g = 15nm. After Ref. [11].

SOI—Silicon on Insulator

Fig. 7-14 shows the steps of making a SOI wafer [12]. 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 4 is to polish the surface. Now the SOI wafer is ready for use.

The silicon film is of high quality and suitable for IC manufacturing. SOI provides a speed advantage because the source/drain to body junction capacitance is practically eliminated because the junctions extend vertically to the buried oxide. The cost of a SOI wafer is many times higher than an ordinary silicon wafer and can increase the total fabrication cost of IC chips by ~30%. For this reason, only some microprocessors, which command high prices and compete on speed, have employed this technology so far. Fig. 7-15 shows the cross-section SEMs of a SOI product [13]. In the future, SOI may find more compelling applications because it offers extra flexibility for making novel structures such the ultra-thin-body MOSFET and multi-gate MOSFET.

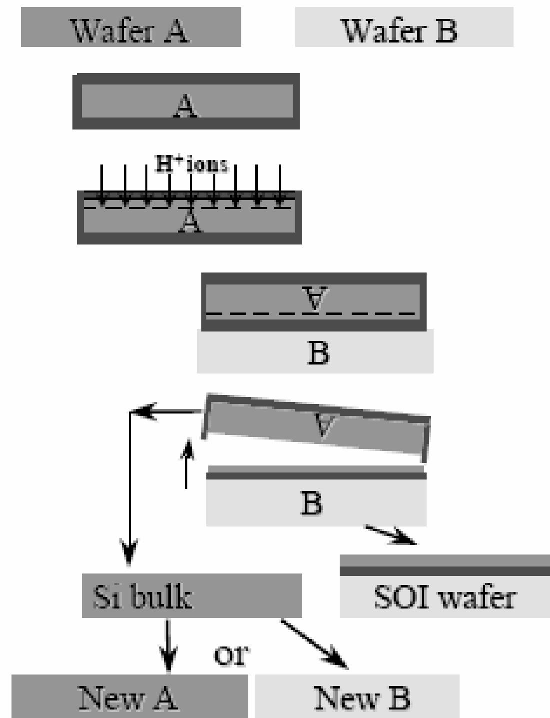


Figure 7-14 Steps of making a SOI wafer.
After Ref. [12]

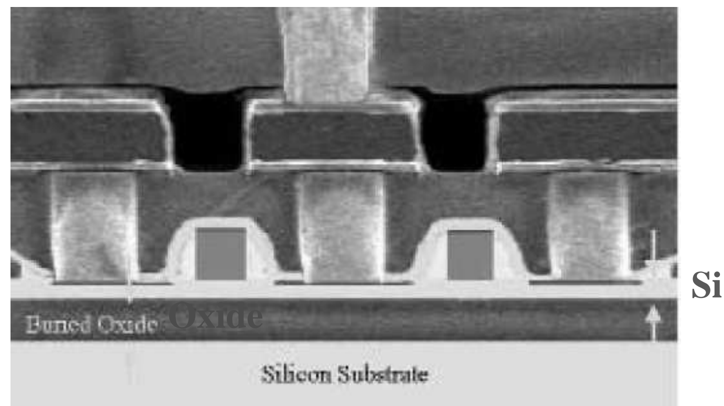


Figure 7-15 The cross-sectional electron micrograph of a SOI product. After Ref. [13]

7.8.2 Multi-gate MOSFET and FinFET

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. 7-16. The silicon film is very thin so that no leakage path is far from one of the gates. (The worst-case path is along the center of the silicon film.) Therefore, the gate(s) can suppress leakage current more effectively than the conventional MOSFET. Because there are more than one gates, the structure may be called **multi-gate MOSFET**. The structure shown in Fig. 7-16 may be called a **double-gate MOSFET**.

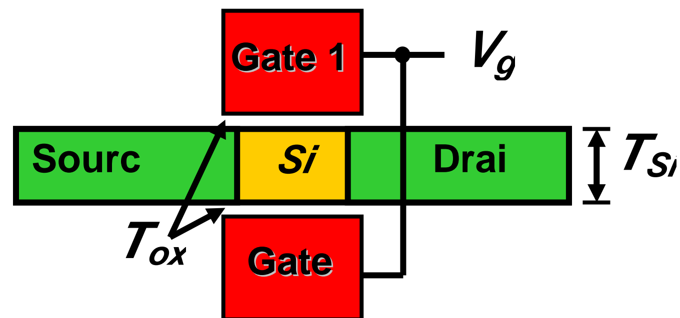


Figure 7-16 The schematic sketch of a horizontal double-gate MOSFET with gates connected.

There is one multi-gate structure that is particularly attractive for its simplicity of fabrication and it is illustrated in Fig. 7-17. The process starts with an SOI wafer. A thin fin of silicon is created by lithography and etching. Gate oxide is grown over the expose surfaces of the fin. Poly-Si gate material is deposited over the fin and gate is patterned by lithography and etching. Finally, source/drain implantation is performed. The final structure in Fig. 7-17 is basically the multi-gate structure in Fig.7-16 turned on its side. This structure is called **FinFET** because of its silicon body resembles the back fin of a fish [14]. 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.

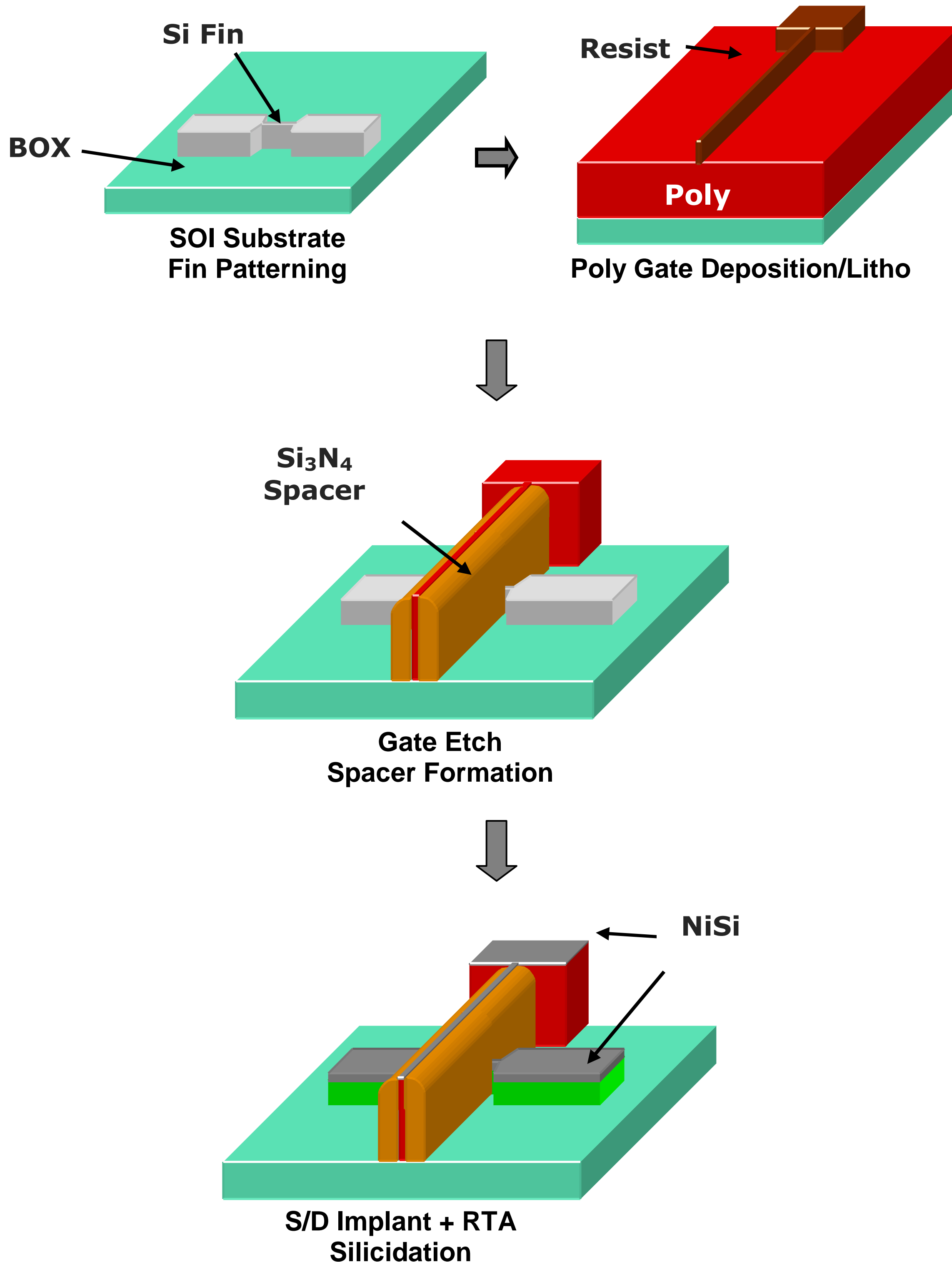


Figure 7-17 The process flow of FinFET starts with an SOI wafer. A thin fin of silicon is created by lithography and etching. Gate oxide is grown over the expose surfaces of the fin. Poly-Si gate material is deposited over the fin and gate is patterned by lithography and etching.

Several variations of FinFET are shown in Fig. 7-18 [15,16]. 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 or less challenging etching. In this case, the top surface of the fin contributes significantly to the suppression of Vt roll-off and to leakage control. This structure is also known as a **Tri-gate MOSFET**. The third variation gives the gate even more control over the silicon wire by surrounding it. It may be called a nanowire FinFET.

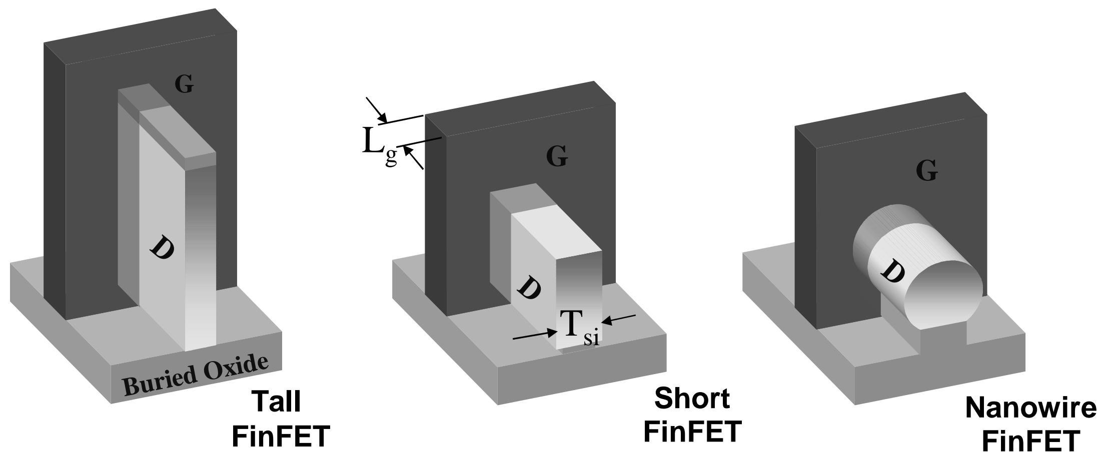


Figure 7-18 Variations of FinFET. Tall FinFET has the advantage of providing a large W and therefore large Ion while occupying a small footprint. Short FinFET has the advantage of less challenging lithography and etching. Nanowire FinFET gives the gate even more control over the silicon wire by surrounding it.

Fig. 7-19 shows the simulated density of inversion electrons in the cross-section of a FinFET body [17]. It is obvious that the inversion layer has a significant thickness (T_{inv}). Note also that there is a larger density of inversion electrons at the corners. There, a pair of gates, at right angle to each other, create a larger band bending and attract more inversion electrons. Fig. 7-20 shows the simulated I-V curves of a nanowire MOSFET .

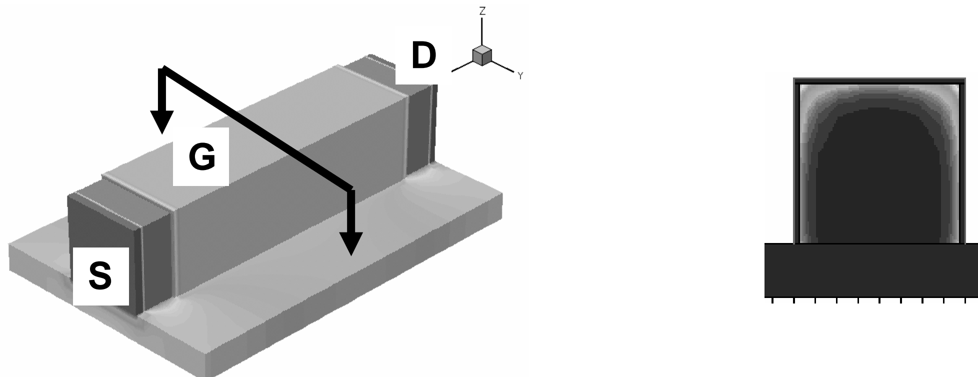


Fig. 7-19 Simulated density of inversion electrons in the cross-section of a FinFET body. After Ref. [17].

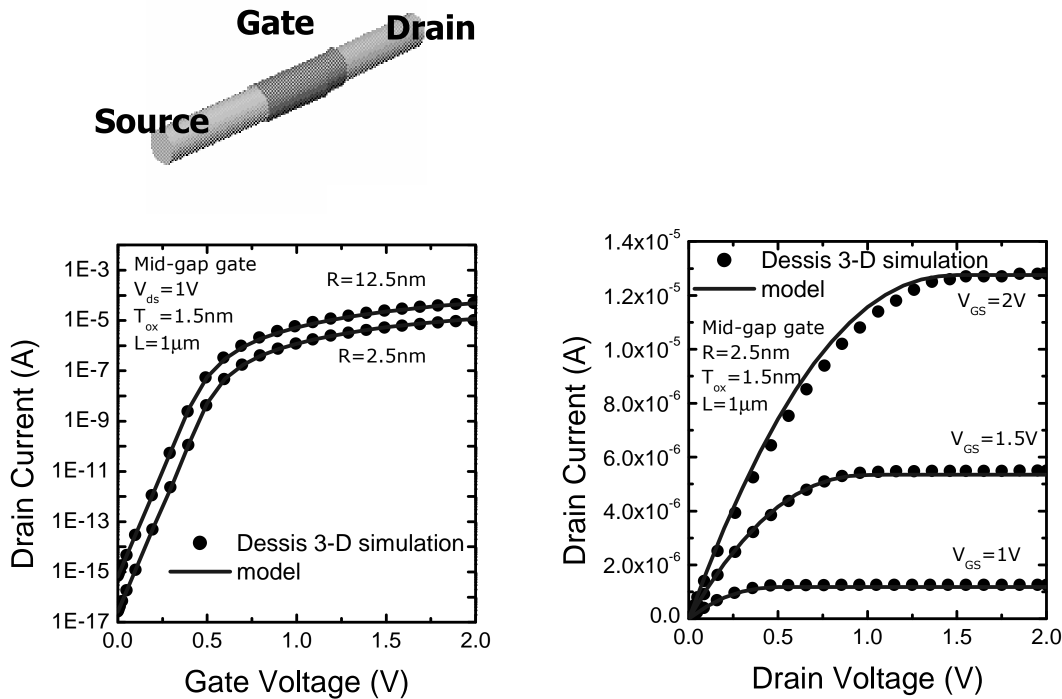


Fig. 7-20 Simulated I-V curves of a nanowire "multi-gate" MOSFET. After Ref. [17]

Device Simulation and Process Simulation

There are several commercially available computer simulation suites that solve all the equations presented in this book with few or no approximations (for example, 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 Fig. 7-11, 7-13, 7-19, and 7-20. Each of the figures takes about 30 min to several hours to generate by device simulations.

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. An example of the process simulation output is shown in Fig. 7-21 [18]. This output may be fed into a device simulator as input together with the applied voltages and the operating temperature.

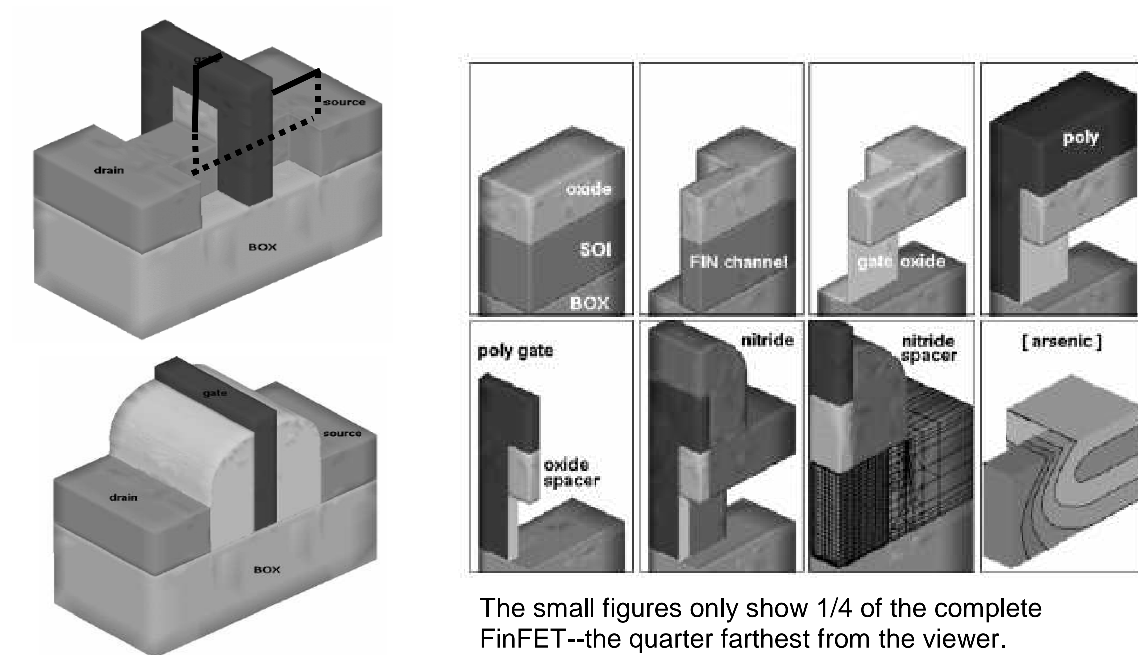


Fig. 7-21 An example of the FinFET process simulation output. After Ref. [18].

7.9 Output Conductance

Output conductance does not contribute to MOSFET leakage. In fact, it is usually discussed together with the MOSFET I_{ds}-V_{ds} theory. However, its cause and theory is actually intimately related to those of V_t roll-off. That makes the present chapter a fitting home for it, too.

The saturation of I_{ds} (at $V_{ds}>V_{dsat}$) is rather clear in Fig. 6-22(b). The saturation of I_{ds} in Fig. 6-22(a) is unclear and incomplete. The reason for the difference is that the channel length is long in the former case and short in the latter. The slope of the I-V curve is called the **output conductance**,

$$g_{ds} \equiv \frac{dI_{dsat}}{dV_{ds}} \tag{7.9.1}$$

A clear saturation of I_{ds}, i.e., a small gds is desirable. The reason can be explained with the help of the amplifier in Fig. 7-22. The bias voltages are chosen such that the transistor operates in the saturation region. A small-signal input, v_{in}, is applied.

$$\begin{aligned} i_{ds} &= g_{msat} \cdot v_{gs} + g_{ds} \cdot v_{ds} \\ &= g_{msat} \cdot v_{in} + g_{ds} \cdot v_{out} \end{aligned} \quad 7.9.2$$

$$v_{out} = -R \times i_{ds} \quad 7.9.3$$

Eliminate i_{ds} from the last two equations and we obtain

$$v_{out} = \frac{-g_{msat}}{(g_{ds} + 1/R)} \times v_{in} \quad 7.9.4$$

The magnitude of the output voltage, according to Eq. 7.9.4 is amplified from the input voltage by a gain factor of $g_{msat}/(g_{ds} + 1/R)$. The gain factor can be increased by using a large R. Even with R approaching infinity, the maximum available voltage gain is

$$\text{Maximum voltage gain} = \frac{g_{msat}}{g_{ds}} \quad 7.9.5$$

If g_{ds} is large, the voltage gain will be small. As an extreme example, the maximum gain will be only 1 if g_{ds} is equal to g_{msat}. A large gain is obviously beneficial to analog circuit applications. A reasonable gain is also needed for digital circuit applications to enhance noise immunity. Therefore, g_{ds} must be kept low.

What device design parameters determine the output conductance? Let us start with Eq. 7.9.1,

$$g_{ds} \equiv \frac{dI_{dsat}}{dV_{ds}} = \frac{dI_{dsat}}{dV_t} \cdot \frac{dV_t}{dV_{ds}} \quad 7.9.6$$

Since I_{ds} is a function of V_{gs}-V_t (see Eq. 6.9.11), it is obvious that

$$\frac{dI_{dsat}}{dV_t} = \frac{-dI_{dsat}}{dV_{gs}} = -g_{msat} \quad 7.9.7$$

The last step is the definition of g_{msat}. Now, Eq. 7.9.6 can be evaluated with the help of Eq. 7.3.3.

$$g_{ds} = g_{msat} \times e^{-L/l_d} \quad 7.9.8$$

$$\text{Max voltage gain} = \frac{g_{msat}}{g_{ds}} = e^{L/l_d} \qquad 7.9.9$$

Eq. 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 V_t roll-off.** This is why g_{ds} is larger in MOSFET with shorter L. This mechanism is sometimes called **drain-induced barrier lowering**. The name refers to the concept depicted in Fig. 7-4. To reduce g_{ds} or to increase voltage gain, we can use a large L and/or reduce l_d. Circuit designers routinely use much large 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 changes that improve the suppression of V_t roll-off and subthreshold leakage also suppress gds and improve 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}. That is **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}. The distance is taken from the original channel length. 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 gds due to channel length modulation is approximately

$$g_{ds} = \frac{l_d \cdot I_{dsat}}{L(V_{ds} - V_{dsat})} \qquad 7.9.10$$

where l_d is given in Eq. 7.3.4. This component of gds can also be suppressed with larger L and smaller T_{ox}, X_j, and W_{dep}.

7-10 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 automotive products. 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 equations, of course. These models are called **compact models** to highlight their computational efficiency in contrast with the device simulators described in Sec. 7.8.
Some circuit-design methodologies use circuit simulations extensively. Other design methodologies use cell libraries, which have been carefully designed and characterized

beforehand using circuit simulations. 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.

BSIM---Berkeley Short-channel IGFET Model

At one time, nearly every company developed its own compact models. In 1996, the Compact Model Council, an industry standard setting group sponsored by most of the world's largest semiconductor manufacturers and design tool companies, set out to select one standard model. It selected **BSIM** as the world's first industry standard model in 1997. (I in BSIM is for IGFET. IG stands for insulated-gate, which is a more generic name for MOS because it does not refer to the materials used for the gate or the insulator.) Now, nearly all the semiconductor companies in the world use BSIM to some degree.

If the I_{ds} equation of BSIM is typed out on paper, it will fill two pages.

Fig. 7-22 shows selected comparisons of the BSIM 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. Fig. 7-23 illustrates this capability. Finally, a good compact model should provide fast simulation times by using simple model equations. In addition to the I-V of N-channel and P-channel transistors, the model also includes capacitance models, gate dielectric leakage current model, source and drain junction diode model.

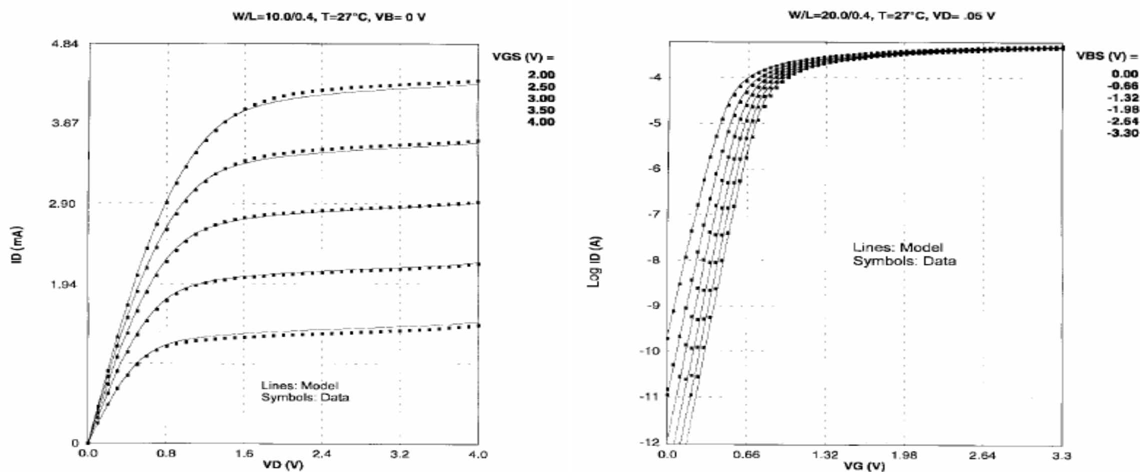


Fig. 7-22 Selected comparisons of BSIM and measured device data to illustrate the accuracy of a compact model. After Ref. [19]

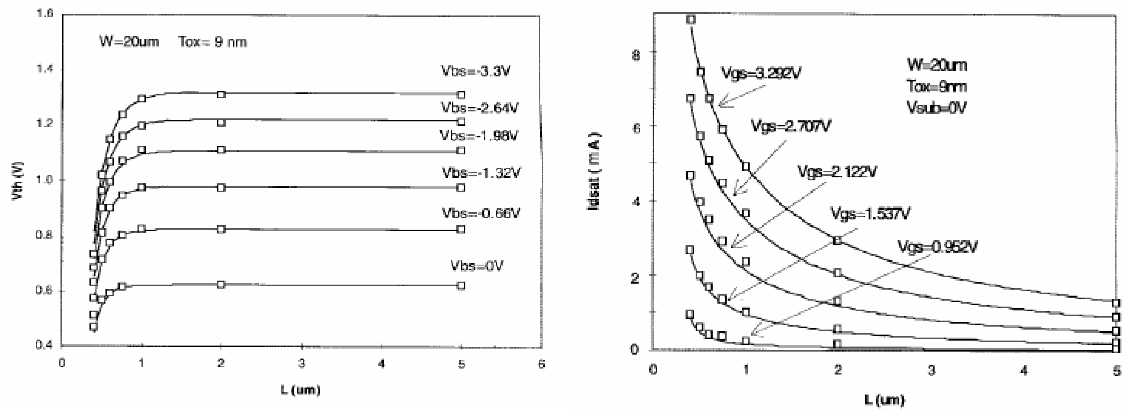


Fig. 7-23 A compact model needs to accurately model the transistor behaviors for any L and W that circuit designers may specify. After Ref. [19].

References

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

[2] T. Ghani, M. Armstrong, C. Auth, M. Bost, P. Charvat, G. Glass, T. Hoffmann, K. Johnson, C. Kenyon, J. Klaus, B. McIntyre, K. Mistry, A. Murthy, J. Sandford, M. Silberstein, S. Sivakumar, P. Smith, K. Zawadzki, S. Thompson, and M. Bohr, “A 90nm high volume manufacturing logic technology featuring novel 45nm gate length strained silicon CMOS transistors,” IEDM Technical Digest, pp. 978-980, 2003.

[3] K. Goto, Y. Tagawa, H. Ohta, H. Morioka, S. Pidin, Y. Momiyama, H. Kokura, S. Inagaki, N. Tamura, M. Hori, T. Mori, M. Kase, K. Hashimoto, M. Kojima, and T. Sugii, “High performance 25nm gate CMOSFETs for 65nm node high speed MPUs,” IEDM Technical Digest, pp. 623-626, 2003.

[4] Z.H. Liu, C. Hu, J-H. Huang, T-Y. Chan, M-C. Jeng, P.K. Ko, Y.C. Cheng, "Threshold Voltage Model for Deep-Submicrometer MOSFET's," *IEEE Trans. on Electron Devices,* Vol. 40, No. 1, January 1993, pp. 86-95.

[5] C.H. Wann, K. Noda, T. Tanaka, M. Yoshida, C. Hu, "A Comparative Study of Advanced MOSFET Concepts," *IEEE Transactions on Electron Devices*, Vol. 43, No. 10, October 1996, pp. 1742-1753.

[6] Yee-Chia Yeo; Tsu-Jae King; Chenming Hu, “MOSFET gate leakage modeling and selection guide for alternative gate dielectrics based on leakage considerations,” *IEEE Transactions on Electron Devices*, Vol. 50, No. 4, April 2003, pp. 1027-1035.

[7] W. Tsai, L.-Å Ragnarsson, L. Pantisano, P. J. Chen, B. Onsia, T. Schram, E. Cartier, A. Kerber, E. Young, M. Caymax, S. De Gendt, and M. Heyns, “Performance comparison of sub 1 nm sputtered TiN/HfO_2nMOS and pMOSFETs,” IEDM Technical Digest, pp. 311-314, 2004.

[8] I.C. Chen, S. Holland, C. Hu, "Electrical Breakdown in Thin Gate and Tunneling Oxides," IEEE Trans. Electron Devices, Vol. ED-32, February 1985, pp. 413-422 and IEEE Journal Solid-State Circuits, Vol. SC-20, February l985, pp. 333-342.

[9] J. Kedzierski, P. Xuan, E.H. Anderson, J. Bokor, T-J. King, C. Hu, "Complementary Silicide Source/Drain Thin-Body MOSFETs for the 20nm Gate Length Regime," IEDM Meeting 2000, IEDM Technical Digest, San Francisco, CA, pp. 57-60, December 10-13, 2000.

[10] Reference for Fig. 7-11

[11] Yang-Kyu Choi; Asano, K.; Lindert, N.; Subramanian, V.; Tsu-Jae King; Bokor, J.; Chenming Hu, “Ultrathin-body SOI MOSFET for deep-sub-tenth micron era,” IEEE Electron Device Letters, Vol. 21, No. 5, May 2000, pp. 254-255.

[12] George Celler, and Michael Wolf, “Smart Cut™ A guide to the technology, the process, the products,” SOITEC, July 2003.

[13] Jerry Yue and Jeff Kriz, “SOI CMOS Technology for RF System-on-Chip Applications, “Microwave Journal, January 2002.

[14] X. Huang, W-C. Lee, C. Kuo, D. Hisamoto, L. Chang, J. Kedzierski, E. Anderson, H. Takeuchi, Y-K. Choi, K. Asano, V. Subramanian, T-J. King, J. Bokor, C. Hu, "Sub 50-nm FinFET: PMOS," IEDM Technical Digest, Washington, DC, pp. 67-70, December 5-8, 1999.

[15] Fu-Liang Yang; Hao-Yu Chen; Fang-Cheng Chen; Cheng-Chuan Huang; Chang-Yun Chang; Hsien-Kuang Chiu; Chi-Chuang Lee; Chi-Chun Chen; Huan-Tsung Huang; Chih-Jian Chen; Hun-Jan Tao; Yee-Chia Yeo; Mong-Song Liang; Chenming Hu, “25 nm CMOS Omega FETs,” IEDM Technical Digest, pp. 255-258, 1999.

[16] Fu-Liang Yang; Di-Hong Lee; Hou-Yu Chen; Chang-Yun Chang; Sheng-Da Liu; Cheng-Chuan Huang; Tang-Xuan Chung; Hung-Wei Chen; Chien-Chao Huang; Yi-Hsuan Liu; Chung-Cheng Wu; Chi-Chun Chen; Shih-Chang Chen; Ying-Tsung Chen; Ying-Ho Chen; Chih-Jian Chen; Bor-Wen Chan; Peng-Fu Hsu; Jyu-Horng Shieh; Han-Jan Tao; Yee-Chia Yeo; Yiming Li; Jam-Wem Lee; Pu Chen; Mong-Song Liang; Chenming Hu, “5nm-gate nanowire FinFET,” VLSI Technology, 2004. Digest of Technical Papers, pp. 196-197.

[17] Chung-Hsun Lin, Guannan Xu, Xuemei Xi, Mansun Chan, Ali M. Niknejad, and Chenming Hu, “Corner Effect Model for Compact Modeling of Multi-Gate MOSFETs,” 2005 SRC TECHCON.

[18] Taurus Process, Synoposys TCAD Manual, Synoposys Inc, Mountain View, CA

[19] Y. Cheng, M-C. Jeng, Z. Liu, J. Huang, M. Chan, K. Chen, P. K. Ko, and C. Hu, "A Physical and Scalable I-V Model in BSIM3v3 for Analog/Digital Circuit Simulation," *IEEE Trans. on Electron Devices* , Vol. 44, No. 2, pp. 277-287, February 1997.

Problems

Subthreshold Leakage Current

Problem 7.1:

Assume the gate oxide between an n+polysilicon gate and the p-substrate is 11 Angstrom thick and that Na=1E18.

a) What is Vt for this device?
b) What is the sub threshold swing, S?
c) What is the maximum leakage current if W=1um, L=18nm? (Assume Ids = 100W/LnA at Vg=Vt).

Problem 7.2. FIELD OXIDE LEAKAGE

Assume the field oxide between an n+polysilicon wire and the p-substrate is 0.3um thick and that Na=5E17.

a)What is Vt for this field oxide device?
b)What is the subthreshold swing, S?
c)What is the maximum field leakage current if W=10um, L=0.3um, and Vdd=2.0V?
d)What is the answer to part a if there is a fixed interface trap charge of 1E10cm^-2?

Vt Roll-off

Prob. 7.3

Qualitatively sketch log(Id) vs. Vg (assume Vd=Vdd) for the following:
i) L=0.2um, Na=1E15
ii) L=0.2um, Na=1E17
iii) L=1um, Na=1E15
iv) L=1um, Na=1E17
Please pay attention to the positions of the curves relative to each other and label all curves.

Trade-off between Ioff and Ion.

Problem 7.4

Does each of the following change increase or decrease Ioff, and Ion? A larger Vt. A larger L. A shallower junction. A smaller Vdd. A smaller Tox. Which of these changes contribute to leakage reduction without reducing the precious Ion?

Problem 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 concern for high Ion and low Ioff.

(a) Answer this question using 1 paragraph of less then 50 words.
(b) Support your description in (a) with 3 hand drawn sketches of your choice.
(c) Why is it not possible to achieve high Ion and small I_{off} by picking optimal T_{ox}, Xj W_{dep} etc? Please explain in your own words.
(d) Provide three equations that help to quantify the issues discussed in part (c). (Suggestion: for fun why don't you try to do this question without copying words from the text).

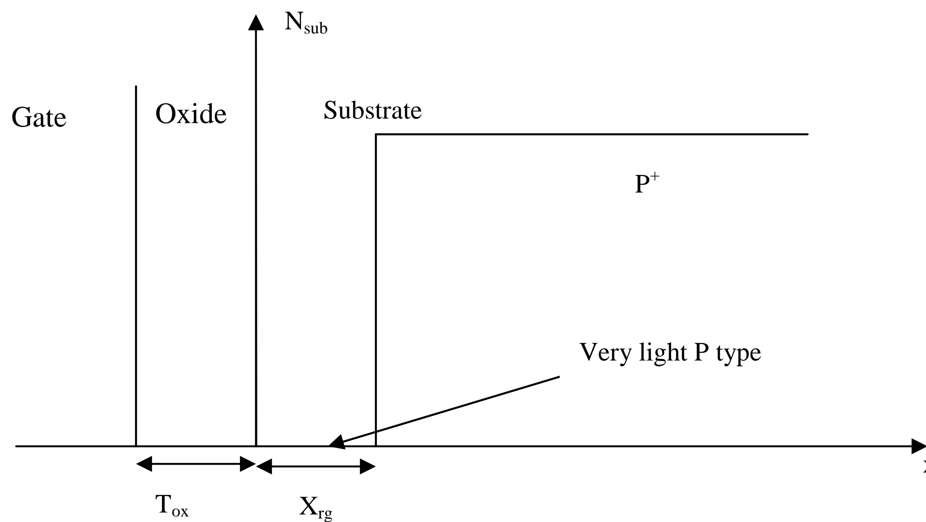
Prob. 7.6
A). Rewrite Eq. 7.4.5 in a form that does not contain Wdep but contains Vt. Do so by using Eq. 5.5.1 and Eq. 5.4.3 assuming that Vt is given. B). Based on the answer to A), state what actions can be taken to reduce the minimum acceptable channel length.

Prob. 7.7.
(a) What is the advantage of having a small Wdep? (a) For given L and Vt, what is the impact of reducing Wdep on Idsat and gate? (Hint: consider the "m" in Ch. 6) (Overall, smaller Wdep is desirable because it is important to be able to suppress Vt roll-off so that L can be scaled.)

MOSFET with Ideal Retrograde Doping Profile

Prob. 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 the figure below.



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 $d\varepsilon/dx = 0$). Assume that the Fermi level in the P^+ region coincides with Ev and the Fermi level in the N^+ gate coincides with Ec. Remember to label Ec, Ev and Ef.

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 part (a) and remember that V_t is the difference between the Fermi levels in the gate and in the substrate. At threshold, at the Si-SiO_2 interface, Ec of Si coincides with the Fermi level).

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 Nsub in the lightly doped region is almost 0, so if your answer is in terms of Nsub alone, you might want to rethink your strategy. Also remember: think carefully about how you derived Vt for a uniformly doped substrate. Maybe $\varepsilon_{ox}\varepsilon_{ox} = \varepsilon_{si}\varepsilon_{si}$ could be a starting point.

d. Show that the depletion layer width, Wdep in an ideal retrograde MOSFET can be about half the Xdep of a uniformly doped device and still yield the same Vt.