$$I_D = \frac{k_n}{2}\cdot(V_{GS} - V_T)^2 \quad \Leftrightarrow \quad \sqrt{I_D} = \sqrt{\frac{k_n}{2}}\cdot(V_{GS} - V_T)$$

Let (V_{GS1}, I_{D1}) and (V_{GS2}, I_{D2}) be any two current-voltage pairs obtained from the table. Then, the square-root of the transconductance parameter k_n can be calculated.

$$\sqrt{\frac{k_n}{2}} = \frac{\sqrt{I_{D1}} - \sqrt{I_{D2}}}{V_{GS1} - V_{GS2}} = \frac{\sqrt{433\ \mu A} - \sqrt{97\ \mu A}}{5\ V - 3\ V} = 5.48\times10^{-3}\ A^{1/2}/V$$

Thus, the transconductance parameter of this n-channel MOSFET is:

$$k_n = 2\cdot\left(5.48\times10^{-3}\right)^2 = 60\times10^{-6}\ A/V^2 = 60\ \mu A/V^2$$

The *extrapolated threshold voltage* V_{T0} at zero substrate bias can be found by calculating the x-axis intercept of the *square-root of* (I_D) versus V_{GS} curve.

$$V_{T0} = V_{GS} - \sqrt{\frac{2\cdot I_D}{k_n}} = 1.2\ V$$

To find the substrate bias coefficient γ, we must first determine the threshold voltage V_T at the source-to-substrate voltage of 3 V. Using one of the current-voltage data pairs corresponding to $V_{SB} = 3$ V, V_T can be calculated as follows:

$$V_T(V_{SB} = 3\ V) = V_{GS} - \sqrt{\frac{2\cdot I_D}{k_n}} = 4\ V - \sqrt{\frac{2\cdot 173\ \mu A}{60\ \mu A/V^2}} = 1.6\ V$$

Finally, the substrate bias coefficient is found as:

$$\gamma = \frac{V_T(V_{SB} = 3\ V) - V_{T0}}{\sqrt{|2\phi_F| + V_{SB}} - \sqrt{|2\phi_F|}} = \frac{1.6\ V - 1.2\ V}{\sqrt{0.6\ V + 3\ V} - \sqrt{0.6\ V}} = 0.36\ V^{1/2}$$

3.5. MOSFET Scaling and Small-Geometry Effects

The design of high-density chips in MOS VLSI (Very Large Scale Integration) technology requires that the packing density of MOSFETs used in the circuits is as high as possible and, consequently, that the sizes of the transistors are as small as possible. The reduction of the size, i.e., the dimensions of MOSFETs, is commonly referred to as *scaling*. It is expected that the operational characteristics of the MOS transistor will

change with the reduction of its dimensions. Also, some physical limitations eventually restrict the extent of scaling that is practically achievable. There are two basic types of size-reduction strategies: *full scaling* (also called constant-field scaling) and *constant-voltage scaling*. Both types of scaling approaches will be shown to have unique effects upon the operating characteristics of the MOS transistor. In the following, we will examine in detail the scaling strategies and their effects, and we will also consider some of the physical limitations and small-geometry effects that must be taken into account for scaled MOSFETs.

Scaling of MOS transistors is concerned with systematic reduction of overall dimensions of the devices as allowed by the available technology, while preserving the geometric ratios found in the larger devices. The proportional scaling of all devices in a circuit would certainly result in a reduction of the total silicon area occupied by the circuit, thereby increasing the overall functional density of the chip. To describe device scaling, we introduce a constant *scaling factor* $S > 1$. All horizontal and vertical dimensions of the *large-size* transistor are then divided by this scaling factor to obtain the scaled device. The extent of scaling that is achievable is obviously determined by the fabrication technology and more specifically, by the minimum feature size. Table 3.1 below shows the recent history of reducing feature sizes for the typical CMOS gate-array process. It is seen that a new generation of manufacturing technology replaces the previous one about every two or three years, and the down-scaling factor S of the minimum feature size from one generation to the next is about 1.2 to 1.5.

Year	1985	1987	1989	1991	1993	1995	1997	1999
Feature size (μm)	2.5	1.7	1.2	1.0	0.8	0.5	0.35	0.25

Table 3.1. Reduction of the minimum feature size (minimum dimensions that can be defined and manufactured on chip) over the years, for a typical CMOS gate-array process.

We consider the proportional scaling of all three dimensions by the same scaling factor S. Figure 3.24 shows the reduction of key dimensions on a typical MOSFET, together with the corresponding increase of the doping densities.

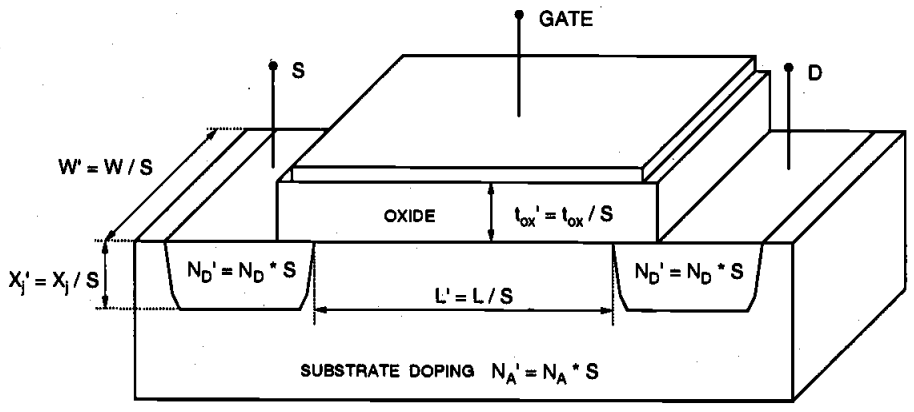


Figure 3.24. Scaling of a typical MOSFET by a scaling factor of S.

The primed quantities in Fig. 3.24 indicate the scaled dimensions and doping densities. It is easy to recognize that the scaling of all dimensions by a factor of $S > 1$ leads to the reduction of the area occupied by the transistor by a factor of S^2. To better understand the effects of scaling upon the current-voltage characteristics of the MOSFET, we will examine two different scaling options in the following sections.

Full Scaling (Constant-Field Scaling)

This scaling option attempts to preserve the magnitude of internal electric fields in the MOSFET, while the dimensions are scaled down by a factor of S. To achieve this goal, all potentials must be scaled down proportionally, by the same scaling factor. Note that this potential scaling also affects the threshold voltage V_{T0}. Finally, the Poisson equation describing the relationship between charge densities and electric fields dictates that the charge densities must be *increased* by a factor of S in order to maintain the field conditions. Table 3.2 lists the scaling factors for all significant dimensions, potentials, and doping densities of the MOS transistor.

Quantity	Before Scaling	After Scaling
Channel length	L	$L' = L/S$
Channel width	W	$W' = W/S$
Gate oxide thickness	t_{ox}	$t_{ox}' = t_{ox}/S$
Junction depth	x_j	$x_j' = x_j/S$
Power supply voltage	V_{DD}	$V_{DD}' = V_{DD}/S$
Threshold voltage	V_{T0}	$V_{T0}' = V_{T0}/S$
Doping densities	N_A N_D	$N_A' = S \cdot N_A$ $N_D' = S \cdot N_D$

Table 3.2. Full scaling of MOSFET dimensions, potentials, and doping densities.

Now consider the influence of full scaling described here upon the current-voltage characteristics of the MOS transistor. It will be assumed that the surface mobility μ_n is not significantly affected by the scaled doping density. The gate oxide capacitance per unit area, on the other hand, is changed as follows.

$$C_{ox}' = \frac{\varepsilon_{ox}}{t_{ox}'} = S \cdot \frac{\varepsilon_{ox}}{t_{ox}} = S \cdot C_{ox} \tag{3.67}$$

The aspect ratio W/L of the MOSFET will remain unchanged under scaling. Consequently, the transconductance parameter k_n will also be scaled by a factor of S. Since all terminal voltages are scaled down by the factor S as well, the linear-mode drain current of the scaled MOSFET can now be found as:

$$I_D'(lin) = \frac{k_n'}{2} \cdot \left[2 \cdot (V_{GS}' - V_T') \cdot V_{DS}' - V_{DS}'^2\right]$$
$$= \frac{S \cdot k_n}{2} \cdot \frac{1}{S^2} \cdot \left[2 \cdot (V_{GS} - V_T) \cdot V_{DS} - V_{DS}^2\right] = \frac{I_D(lin)}{S} \tag{3.68}$$

Similarly, the saturation-mode drain current is also reduced by the same scaling factor.

$$I_D'(sat) = \frac{k_n'}{2} \cdot (V_{GS}' - V_T')^2 = \frac{S \cdot k_n}{2} \cdot \frac{1}{S^2} \cdot (V_{GS} - V_T)^2 = \frac{I_D(sat)}{S} \tag{3.69}$$

Now consider the power dissipation of the MOSFET. Since the drain current flows between the source and the drain terminals, the instantaneous power dissipated by the device (before scaling) can be found as:

$$P = I_D \cdot V_{DS} \tag{3.70}$$

Notice that full scaling reduces both the drain current and the drain-to-source voltage by a factor of S; hence, the power dissipation of the transistor will be reduced by the factor S^2.

$$P' = I_D' \cdot V_{DS}' = \frac{1}{S^2} \cdot I_D \cdot V_{DS} = \frac{P}{S^2} \tag{3.71}$$

This significant reduction of the power dissipation is one of the most attractive features of full scaling. Note that with the device area reduction by S^2 discussed earlier, we find the *power density* per unit area remaining virtually unchanged for the scaled device.

Finally, consider the gate oxide capacitance defined as $C_g = WLC_{ox}$. It will be shown later in Section 3.6 that charging and discharging of this capacitance plays an important role in the transient operation of the MOSFET. Since the gate oxide capacitance C_g is scaled down by a factor of S, we can predict that the transient characteristics, i.e., the charge-up and charge-down times, of the scaled device will improve accordingly. In addition, the proportional reduction of all dimensions on-chip will lead to a reduction of various parasitic capacitances and resistances as well, contributing to the overall performance improvement. Table 3.3 summarizes the changes in key device characteristics as a result of full (constant-field) scaling.

Constant-Voltage Scaling

While the full scaling strategy dictates that the power supply voltage and all terminal voltages be scaled down proportionally with the device dimensions, the scaling of

voltages may not be very practical in many cases. In particular, the peripheral and interface circuitry may require certain voltage levels for all input and output voltages, which in turn would necessitate multiple power supply voltages and complicated level-shifter arrangements. For these reasons, constant-voltage scaling is usually preferred over full scaling.

Quantity	Before Scaling	After Scaling
Oxide capacitance	C_{ox}	$C_{ox}' = S \cdot C_{ox}$
Drain current	I_D	$I_D' = I_D / S$
Power dissipation	P	$P' = P / S^2$
Power density	$P / Area$	$P' / Area' = P / Area$

Table 3.3. Effects of full scaling upon key device characteristics.

In constant-voltage scaling, all dimensions of the MOSFET are reduced by a factor of S, as in full scaling. The power supply voltage and the terminal voltages, on the other hand, remain unchanged. The doping densities must be increased by a factor of S^2 in order to preserve the charge-field relations. Table 3.4 shows the constant-voltage scaling of key dimensions, voltages, and densities. Under constant-voltage scaling, the changes in device characteristics are significantly different compared to those in full scaling, as we will demonstrate. The gate oxide capacitance per unit area C_{ox} is increased by a factor of S, which means that the transconductance parameter is also increased by S. Since the terminal voltages remain unchanged, the linear mode drain current of the scaled MOSFET can be written as:

$$\begin{aligned} I_D'(lin) &= \frac{k_n'}{2} \cdot \left[2 \cdot (V_{GS}' - V_T') \cdot V_{DS}' - V_{DS}'^2\right] \\ &= \frac{S \cdot k_n}{2} \cdot \left[2 \cdot (V_{GS} - V_T) \cdot V_{DS} - V_{DS}^2\right] = S \cdot I_D(lin) \end{aligned} \tag{3.72}$$

Quantity	Before Scaling	After Scaling
Dimensions	W, L, t_{ox}, x_j	reduced by S $(W' = W / S, \ldots)$
Voltages	V_{DD}, V_T	remain unchanged
Doping densities	N_A, N_D	increased by S^2 $(N_A' = S^2 \cdot N_A, \ldots)$

Table 3.4. Constant-voltage scaling of MOSFET dimensions, potentials, and doping densities.

Also, the saturation-mode drain current will be increased by a factor of S after constant-voltage scaling. This means that the drain current *density* (current per unit area) is increased by a factor of S^3, which may cause serious reliability problems for the MOS transistor.

$$I_D'(sat) = \frac{k_n'}{2} \cdot (V_{GS}' - V_T')^2 = \frac{S \cdot k_n}{2} \cdot (V_{GS} - V_T)^2 = S \cdot I_D(sat) \tag{3.73}$$

Next, consider the power dissipation. Since the drain current is increased by a factor of S while the drain-to-source voltage remains unchanged, the power dissipation of the MOSFET increases by a factor of S.

$$P' = I_D' \cdot V_{DS}' = (S \cdot I_D) \cdot V_{DS} = S \cdot P \tag{3.74}$$

Finally, the power density (power dissipation per unit area) is found to increase by a factor of S^3 after constant-voltage scaling, with possible adverse effects on device reliability.

Quantity	Before Scaling	After Scaling
Oxide capacitance	C_{ox}	$C_{ox}' = S \cdot C_{ox}$
Drain current	I_D	$I_D' = S \cdot I_D$
Power dissipation	P	$P' = S \cdot P$
Power density	$P/Area$	$P'/Area' = S^3 \cdot (P/Area)$

Table 3.5. Effects of constant-voltage scaling upon key device characteristics.

To summarize, constant-voltage scaling may be preferred over full (constant-field) scaling in many practical cases because of the external voltage-level constraints. It must be recognized, however, that constant-voltage scaling increases the drain current density and the power density by a factor of S^3. This large increase in current and power densities may eventually cause serious reliability problems for the scaled transistor, such as electromigration, hot-carrier degradation, oxide breakdown, and electrical over-stress.

As the device dimensions are systematically reduced through full scaling or constant-voltage scaling, various physical limitations become increasingly more prominent, and ultimately restrict the amount of feasible scaling for some device dimensions. Consequently, scaling may be carried out on a certain subset of MOSFET dimensions in many practical cases. Also, the simple gradual channel approximation (GCA) used for the derivation of current-voltage relationships does not accurately reflect the effects of scaling in smaller-size transistors. The current equations have to be modified accordingly. In the following, we will briefly investigate some of these small-geometry effects.

MOS Transistor

As a working definition, a MOS transistor is called a short-channel device if its channel length is on the same order of magnitude as the depletion region thicknesses of the source and drain junctions. Alternatively, a MOSFET can be defined as a short-channel device if the effective channel length L_{eff} is approximately equal to the source and drain junction depth x_j. The short-channel effects that arise in this case are attributed to two physical phenomena : (i) the limitations imposed on electron drift characteristics in the channel, and (ii) the modification of the threshold voltage due to the shortening channel length.

Note that the lateral electric field E_y along the channel increases, as the effective channel length is decreased. While the electron drift velocity v_d in the channel is proportional to the electric field for lower field values, this drift velocity tends to saturate at high channel electric fields. For channel electric fields of $E_y = 10^5$ V/cm and higher, the electron drift velocity in the channel reaches a saturation value of about $v_d(sat) = 10^7$ cm/s. This velocity saturation has very significant implications upon the current-voltage characteristics of the short-channel MOSFET. Consider the saturation-mode drain current, under the assumption that carrier velocity in the channel has already reached its limit value. The effective channel length L_{eff} will be reduced due to channel-length shortening.

$$I_D(sat) = W \cdot v_d(sat) \cdot \int_0^{L_{eff}} q \cdot n(x)\, dx = W \cdot v_d(sat) \cdot |Q_I| \tag{3.75}$$

Since the channel-end voltage is equal to V_{DSAT}, the saturation current can be found as follows:

$$I_D(sat) = W \cdot v_d(sat) \cdot C_{ox} \cdot V_{DSAT} \tag{3.76}$$

Carrier velocity saturation actually reduces the saturation-mode current below the current value predicted by the conventional long-channel current equations. The current is no longer a quadratic function of the gate-to-source voltage V_{GS}, and it is virtually independent of the channel length. Also note that under these conditions, the device is defined to be in saturation when the carrier velocity in the channel approaches about 90% of its limit value.

In short-channel MOS transistors, the carrier velocity in the channel is also a function of the normal (vertical) electric-field component E_x. Since the vertical field influences the scattering of carriers (collisions suffered by the carriers) in the surface region, the surface mobility is reduced with respect to the bulk mobility. The dependence of the surface electron mobility on the vertical electric field can be expressed by the following empirical formula :

$$\mu_n(eff) = \frac{\mu_{no}}{1 + \Theta \cdot E_x} = \frac{\mu_{no}}{1 + \dfrac{\Theta\ \varepsilon_{ox}}{t_{ox}\ \varepsilon_{Si}} \cdot \left(V_{GS} - V_c(y)\right)} \tag{3.77}$$

where μ_{no} is the low-field surface electron mobility and Θ is an empirical factor. For a simple estimation of field-related mobility reduction, (3.77) can be approximated by

$$\mu_n(eff) = \frac{\mu_{no}}{1 + \eta \cdot (V_{GS} - V_T)} \tag{3.78}$$

where η is also an empirical coefficient.

Next, we consider the modification of the threshold voltage due to short-channel effects. The threshold voltage expression (3.23) was derived for a long-channel MOSFET. Specifically, the channel depletion region was assumed to be created only by the applied gate voltage, and the depletion regions associated with the drain and source pn-junctions were neglected. The shape of this gate-induced bulk (channel) depletion region was assumed to be rectangular, extending from the source to the drain. In short-channel MOS transistors, however, the n^+ drain and source diffusion regions in the p-type substrate induce a significant amount of depletion charge; consequently, the long-channel threshold voltage expression derived earlier overestimates the depletion charge supported by the gate voltage. The threshold voltage value found by using (3.23) is therefore larger than the actual threshold voltage of the short-channel MOSFET.

Figure 3.25(a) shows the simplified geometry of the gate-induced bulk depletion region and the pn-junction depletion regions in a short-channel MOS transistor. Note that the bulk depletion region is assumed to have an asymmetric trapezoidal shape, instead of a rectangular shape, to represent accurately the gate-induced charge. The drain depletion region is expected to be larger than the source depletion region because the positive drain-to-source voltage reverse-biases the drain-substrate junction. We recognize that a significant portion of the total depletion region charge under the gate is actually due to the source and drain junction depletion, rather than the bulk depletion induced by the gate voltage. Since the bulk depletion charge in the short-channel device is smaller than expected, the threshold voltage expression must be modified to account for this reduction. Following the modification of the bulk charge term, the threshold voltage of the short-channel MOSFET can be written as

$$V_{T0}(short\ channel) = V_{T0} - \Delta V_{T0} \tag{3.79}$$

where V_{T0} is the zero-bias threshold voltage calculated using the conventional long-channel formula (3.23) and ΔV_{T0} is the threshold voltage shift (reduction) due to the short-channel effect. The reduction term actually represents the amount of charge differential between a rectangular depletion region and a trapezoidal depletion region.

Let ΔL_S and ΔL_D represent the lateral extent of the depletion regions associated with the source junction and the drain junction, respectively. Then, the bulk depletion region charge contained within the trapezoidal region is

$$Q_{B0} = -\left(1 - \frac{\Delta L_S + \Delta L_D}{2L}\right) \cdot \sqrt{2 \cdot q \cdot \varepsilon_{Si} \cdot N_A \cdot |2\phi_F|} \tag{3.80}$$

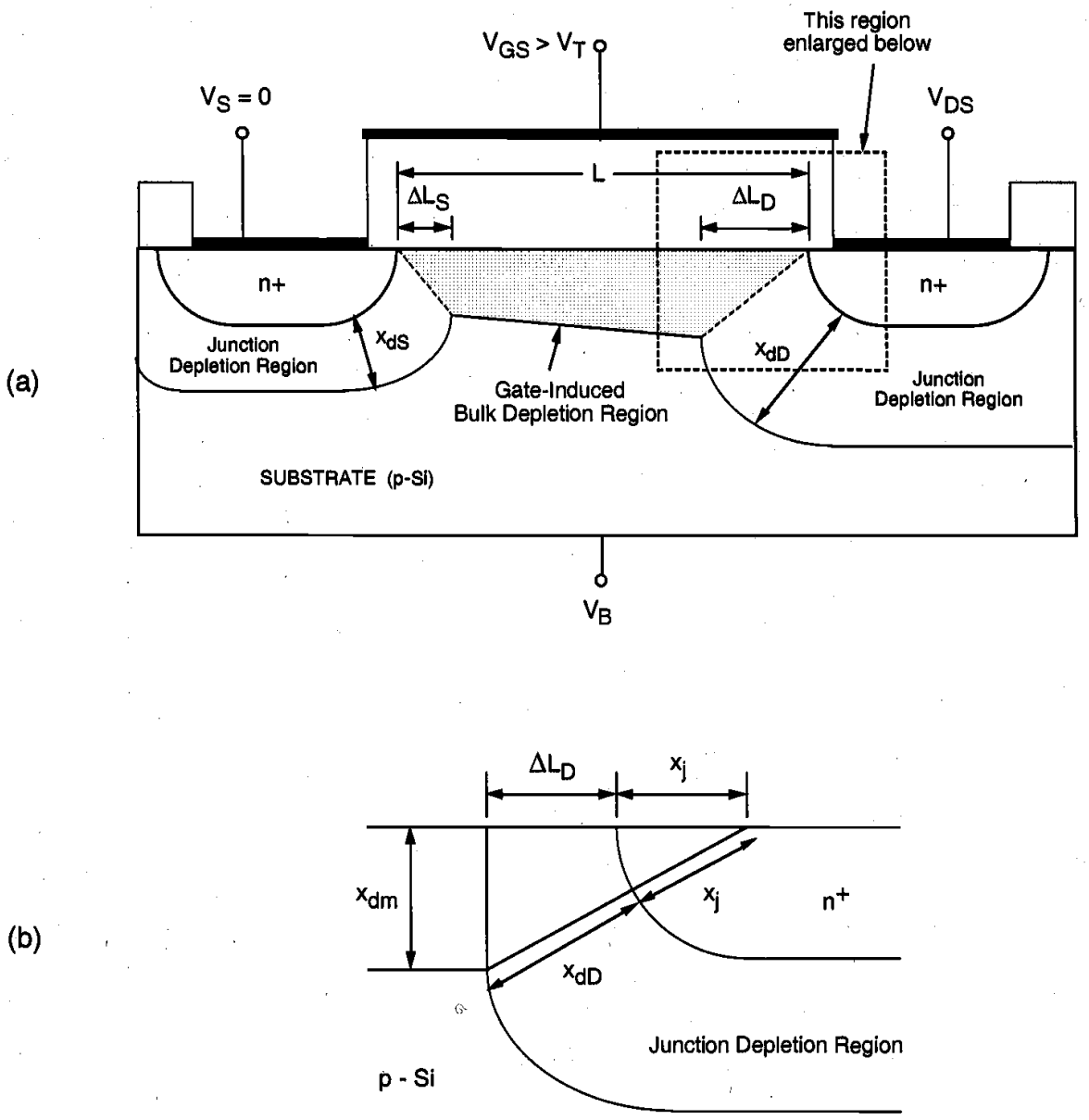


Figure 3.25. (a) Simplified geometry of the MOSFET channel region, with gate-induced bulk depletion region and the pn-junction depletion regions. (b) Close-up view of the drain diffusion edge.

To calculate ΔL_S and ΔL_D, we will use the simplified geometry shown in Fig. 3.25(b). Here, x_{dS} and x_{dD} represent the depth of the pn-junction depletion regions associated with the source and the drain, respectively. The edges of the source and drain diffusion regions are represented by quarter-circular arcs, each with a radius equal to the junction depth, x_j. The vertical extent of the bulk depletion region into the substrate is represented by x_{dm}. The junction depletion region depths can be approximated by

$$x_{dS} = \sqrt{\frac{2 \cdot \varepsilon_{Si}}{q \cdot N_A} \cdot \phi_0} \tag{3.81}$$

$$x_{dD} = \sqrt{\frac{2 \cdot \varepsilon_{Si}}{q \cdot N_A} \cdot \left(\phi_0 + V_{DS}\right)} \tag{3.82}$$

with the junction built-in voltage

$$\phi_0 = \frac{kT}{q} \cdot ln\left(\frac{N_D \cdot N_A}{n_i^2}\right) \tag{3.83}$$

From Fig. 3.25(b), we find the following relationship between ΔL_D and the depletion region depths.

$$\left(x_j + x_{dD}\right)^2 = x_{dm}^2 + \left(x_j + \Delta L_D\right)^2 \tag{3.84}$$

$$\Delta L_D^2 + 2 \cdot x_j \cdot \Delta L_D + x_{dm}^2 - x_{dD}^2 - 2 \cdot x_j \cdot x_{dD} = 0 \tag{3.85}$$

Solving for ΔL_D, we obtain:

$$\Delta L_D = -x_j + \sqrt{x_j^2 - \left(x_{dm}^2 - x_{dD}^2\right) + 2 x_j x_{dD}} \cong x_j \cdot \left(\sqrt{1 + \frac{2 x_{dD}}{x_j}} - 1\right) \tag{3.86}$$

Similarly, the length ΔL_S can also be found as follows:

$$\Delta L_S \cong x_j \cdot \left(\sqrt{1 + \frac{2 x_{dS}}{x_j}} - 1\right) \tag{3.87}$$

Now, the amount of threshold voltage reduction ΔV_{T0} due to short-channel effects can be found as:

$$\Delta V_{T0} = \frac{1}{C_{ox}} \cdot \sqrt{2 q \varepsilon_{Si} N_A \left|2\phi_F\right|} \cdot \frac{x_j}{2L} \cdot \left[\left(\sqrt{1 + \frac{2 x_{dS}}{x_j}} - 1\right) + \left(\sqrt{1 + \frac{2 x_{dD}}{x_j}} - 1\right)\right] \tag{3.88}$$

The threshold voltage shift term is proportional to (x_j/L). As a result, this term becomes more prominent for MOS transistors with shorter channel lengths, and it approaches zero

for long-channel MOSFETs where $L >> x_j$. The following example illustrates the variation of the threshold voltage as a function of channel length in short-channel devices.

Example 3.6.

Consider an n-channel MOS process with the following parameters: substrate doping density $N_A = 10^{16}$ cm^{-3}, polysilicon gate doping density N_D (gate) = 2 x 10^{20} cm^{-3}, gate oxide thickness t_{ox} = 50 nm, oxide-interface fixed charge density N_{ox} = 4 x 10^{10} cm^{-2}, and source and drain diffusion doping density $N_D = 10^{17}$ cm^{-3}. In addition, the channel region is implanted with p-type impurities (impurity concentration N_I = 2 x 10^{11} cm^{-2}) to adjust the threshold voltage. The junction depth of the source and drain diffusion regions is x_j = 1.0 μm.

Plot the variation of the zero-bias threshold voltage V_{T0} as a function of the channel length (assume that $V_{DS} = V_{SB} = 0$). Also find V_{T0} for L = 0.7 μm, V_{DS} = 5 V, and V_{SB} = 0.

First, we have to find the zero-bias threshold voltage using the conventional formula (3.23). The threshold voltage *without* the channel implant was already calculated for the same process parameters in Example 3.2, and was found to be V_{T0} = 0.40 V. The additional p-type channel implant will increase the threshold voltage by an amount of qN_I / C_{ox}. Thus, we find the long-channel zero-bias threshold voltage for the process described above as

$$V_{T0} = 0.40\,\text{V} + \frac{q \cdot N_I}{C_{ox}} = 0.40\,\text{V} + \frac{1.6\times10^{-19}\cdot 2\times10^{11}}{7.03\times10^{-8}} = 0.855\,\text{V}$$

Next, the amount of threshold voltage reduction due to short-channel effects must be calculated using (3.88). The source and drain junction built-in voltage is

$$\phi_0 = \frac{kT}{q}\cdot\ln\left(\frac{N_D\cdot N_A}{n_i^2}\right) = 0.026\ \text{V}\cdot\ln\left(\frac{10^{17}\cdot 10^{16}}{2.1\times10^{20}}\right) = 0.76\ \text{V}$$

For zero drain bias, the depth of source and drain junction depletion regions is found as

$$x_{dS} = x_{dD} = \sqrt{\frac{2\cdot\varepsilon_{Si}}{q\cdot N_A}\cdot\phi_0} = \sqrt{\frac{2\cdot 11.7\cdot 8.85\times10^{-14}}{1.6\times10^{-19}\cdot10^{16}}\cdot 0.76}$$

$$= 31.4\times10^{-6}\,\text{cm} = 0.314\ \mu\text{m}$$

Now, the threshold voltage shift ΔV_{T0} due to short-channel effects can be calculated as a function of the gate (channel) length L.

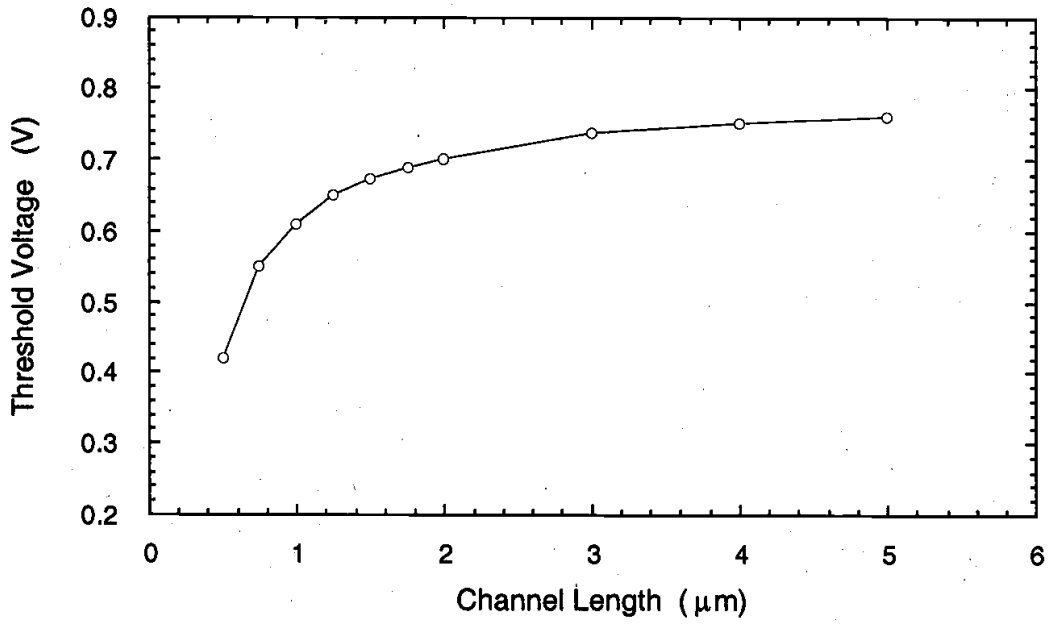
$$\Delta V_{T0} = \frac{1}{C_{ox}} \cdot \sqrt{2\,q\,\varepsilon_{Si}\,N_A\,|2\phi_F|} \cdot \frac{x_j}{2L} \cdot \left[\left(\sqrt{1+\frac{2x_{dS}}{x_j}} - 1 \right) + \left(\sqrt{1+\frac{2x_{dD}}{x_j}} - 1 \right) \right]$$

$$= \frac{4.82\times10^{-8}\ \text{C/cm}^2}{7.03\times10^{-8}\ \text{F/cm}^2} \cdot \frac{1.0\ \mu\text{m}}{L} \cdot \left(\sqrt{1+\frac{2\cdot 0.314\ \mu\text{m}}{1.0\ \mu\text{m}}} - 1 \right)$$

Finally, the zero-bias threshold voltage is found as

$$V_{T0}(\textit{short channel}) = 0.855\,\text{V} - 0.19\,\text{V} \cdot \frac{1}{L[\mu\text{m}]}$$

The following plot shows the variation of the threshold voltage with the channel length. The threshold voltage decreases by as much as 50% for channel lengths in the submicron range, while it approaches the value of 0.8 V for larger channel lengths.



Since the conventional threshold voltage expression (3.23) is not capable of accounting for this drastic reduction of V_{T0} at smaller channel lengths, its application for short-channel MOSFETs must be carefully restricted.

Now, consider the variation of the threshold voltage with the applied drain-to-source voltage. Equation (3.82) shows that the depth of the drain junction depletion region increases with the voltage V_{DS}. For a drain-to-source voltage of V_{DS} = 5 V, the drain depletion depth is found as :

$$x_{dD} = \sqrt{\frac{2 \cdot \varepsilon_{Si}}{q \cdot N_A} \cdot (\phi_0 + V_{DS})} = \sqrt{\frac{2 \cdot 11.7 \cdot 8.85 \times 10^{-14}}{1.6 \times 10^{-19} \cdot 10^{16}} \cdot (0.76 + 5.0)} = 0.863\ \mu\text{m}$$

The resulting threshold voltage shift can be calculated by substituting x_{dD} found above in (3.88).

$$\begin{aligned}\Delta V_{T0} &= \frac{1}{C_{ox}} \cdot \sqrt{2\, q\, \varepsilon_{Si}\, N_A\, |2\phi_F|} \cdot \frac{x_j}{2L} \cdot \left[\left(\sqrt{1 + \frac{2x_{dS}}{x_j}} - 1 \right) + \left(\sqrt{1 + \frac{2x_{dD}}{x_j}} - 1 \right) \right] \\ &= \frac{4.82 \times 10^{-8}}{7.03 \times 10^{-8}} \cdot \frac{1.0}{2 \cdot 0.7} \cdot \left[\left(\sqrt{1 + \frac{2 \cdot 0.314}{1.0}} - 1 \right) + \left(\sqrt{1 + \frac{2 \cdot 0.863}{1.0}} - 1 \right) \right] \\ &= 0.45\ \text{V}\end{aligned}$$

The threshold voltage of this short-channel MOS transistor is calculated as

$$V_{T0} = 0.855\ \text{V} - 0.45\ \text{V} = 0.405\ \text{V}$$

which is significantly lower than the threshold voltage predicted by the conventional long-channel formula (3.23).

Narrow-Channel Effects

MOS transistors that have channel widths W on the same order of magnitude as the maximum depletion region thickness x_{dm} are defined as narrow-channel devices. Similar to the short-channel effects examined earlier, the narrow-channel MOSFETs also exhibit typical characteristics which are not accounted for by the conventional GCA analysis. The most significant narrow-channel effect is that the actual threshold voltage of such a device is *larger* than that predicted by the conventional threshold voltage formula (3.23). In the following, we will briefly review the physical reasons that cause this discrepancy. A typical cross-sectional view of a narrow-channel device is shown in Fig. 3.26. The oxide thickness in the channel region is t_{ox}, while the regions around the channel are covered by a thick *field oxide* (FOX). Since the gate electrode also overlaps with the field oxide as shown in Fig. 3.26, a relatively shallow depletion region forms underneath this FOX-overlap area as well. Consequently, the gate voltage must also support this additional depletion charge in order to establish the conducting channel. The charge contribution of this fringe depletion region to the overall channel depletion charge is negligible in wider devices. For MOSFETs with small channel widths, however, the actual threshold voltage increases as a result of this extra depletion charge.

$$V_{T0}(narrow\ channel) = V_{T0} + \Delta V_{T0} \tag{3.89}$$

The additional contribution to the threshold voltage due to narrow-channel effects can be modeled as follows:

$$\Delta V_{T0} = \frac{1}{C_{ox}} \cdot \sqrt{2\,q\,\varepsilon_{Si}\,N_A\,|2\phi_F|} \cdot \frac{\kappa \cdot x_{dm}}{W} \tag{3.90}$$

where κ is an empirical parameter depending on the shape of the fringe depletion region. Assuming that the depletion region edges are modeled by quarter-circular arcs, for example, the parameter κ can be found as

$$\kappa = \frac{\pi}{2} \tag{3.91}$$

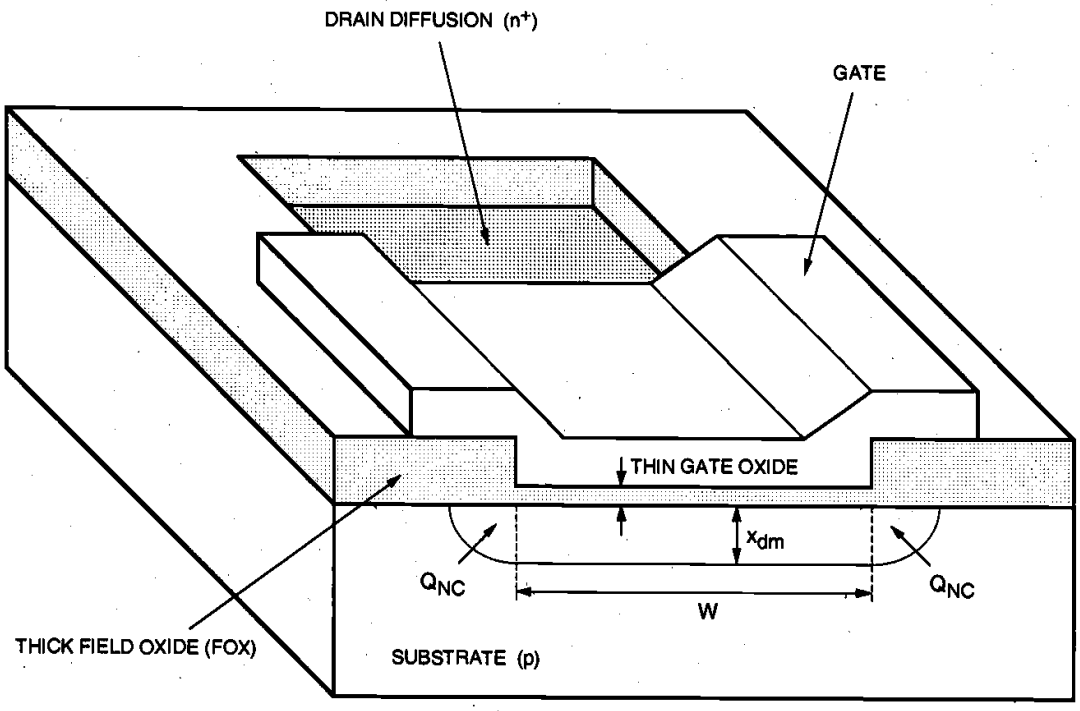


Figure 3.26. Cross-sectional view (across the channel) of a narrow-channel MOSFET. Note that Q_{NC} indicates the extra depletion charge due to narrow-channel effects.

The simple formula given in (3.90) can be modified for various device geometries and manufacturing processes, such as LOCOS, fully-recessed LOCOS, and thick-field-oxide MOSFET process. In all cases, we recognize that the additional contribution to V_{T0} is proportional to (x_{dm} / W). The amount of threshold voltage increase becomes significant only for devices which have a channel width W on the same order of magnitude

as x_{dm}. Finally, note that for minimum-geometry MOSFETs which have a small channel length *and* a small channel width, the threshold voltage variations due to short- and narrow-channel effects may tend to cancel each other out.

Other Limitations Imposed by Small-Device Geometries

In small-geometry MOSFETs, the characteristics of current flow in the channel between the source and the drain can be explained as being controlled by the two-dimensional electric field vector $\vec{E}(x, y)$. The simple one-dimensional gradual channel approximation (GCA) assumes that the electric field components parallel to the surface and perpendicular to the surface are effectively decoupled and, therefore, cannot fully account for some of the observed device characteristics. These small-geometry device characteristics, however, may severely restrict the operating conditions of the transistor and impose limitations upon the practical utility of the device. Accurate identification and characterization of these small-geometry effects are crucial, especially for submicron MOSFETs.

One typical condition, which is due to the two-dimensional nature of channel current flow, is the *subthreshold conduction* in small-geometry MOS transistors. As already discussed in the previous sections, the current flow in the channel depends on creating and sustaining an inversion layer on the surface. If the gate bias voltage is not sufficient to invert the surface, i.e., $V_{GS} < V_{T0}$, the carriers (electrons) in the channel face a *potential barrier* that blocks the flow. Increasing the gate voltage reduces this potential barrier and, eventually, allows the flow of carriers under the influence of the channel electric field. This simple picture becomes more complicated in small-geometry MOSFETs, because the potential barrier is controlled by both the gate-to-source voltage V_{GS} *and* the drain-to-source voltage V_{DS}. If the drain voltage is increased, the potential barrier in the channel decreases, leading to *drain-induced barrier lowering* (DIBL). The reduction of the potential barrier eventually allows electron flow between the source and the drain, even if the gate-to-source voltage is lower than the threshold voltage. The channel current that flows under these conditions ($V_{GS} < V_{T0}$) is called the *subthreshold current*. Note that the GCA cannot account for any nonzero drain current I_D for $V_{GS} < V_{T0}$. Two-dimensional analysis of the small-geometry MOSFET yields the following approximate expression for the subthreshold current.

$$I_D(subthreshold) \cong \frac{qD_n W x_c n_0}{L_B} \cdot e^{\frac{q\phi_r}{kT}} \cdot e^{\frac{q}{kT}(A \cdot V_{GS} + B \cdot V_{DS})} \tag{3.92}$$

Here, x_c is the subthreshold channel depth, D_n is the electron diffusion coefficient, L_B is the length of the barrier region in the channel, and ϕ_r is a reference potential. Note the exponential dependence of the subthreshold current on both the gate and the drain voltages. Identifying subthreshold conduction is very important for circuit applications where small amounts of current flow may significantly disturb the circuit operation.

We remember from the previous analysis that in small-geometry MOSFETs, the channel length is on the same order of magnitude as the source and drain depletion region thicknesses. For large drain-bias voltages, the depletion region surrounding the drain can

extend farther toward the source, and the two depletion regions can eventually merge. This condition is termed *punch-through*; the gate voltage loses its control upon the drain current, and the current rises sharply once punch-through occurs. Being able to cause permanent damage to the transistor by localized melting of material, punch-through is obviously an undesirable condition, and should be prevented in normal circuit operation.

As some device dimensions, such as the channel length, are scaled down with each new generation, we find that some dimensions cannot be arbitrarily scaled because of physical limitations. One such dimension is the gate oxide thickness t_{ox}. The reduction of t_{ox} by a scaling factor of S, i.e., building a MOSFET with $t_{ox}' = t_{ox} / S$, is restricted by processing difficulties involved in growing very thin, uniform silicon-dioxide layers. Localized sites of nonuniform oxide growth, also called *pinholes*, may cause electrical shorts between the gate electrode and the substrate. Another limitation on the scaling of t_{ox} is the possibility of *oxide breakdown*. If the oxide electric field perpendicular to the surface is larger than a certain *breakdown field*, the silicon-dioxide layer may sustain permanent damage during operation, leading to device failure.

Finally, we will consider another reliability problem caused by high electric fields within the device. We have seen that advances in VLSI fabrication technologies are primarily based on the reduction of device dimensions, such as the channel length, the junction depth, and the gate oxide thickness, without proportional scaling of the power supply voltage (constant-voltage scaling). This decrease in critical device dimensions to submicron ranges, accompanied by increasing substrate doping densities, results in a significant increase of the horizontal and vertical electric fields in the channel region. Electrons and holes gaining high kinetic energies in the electric field (*hot carriers*) may, however, be injected into the gate oxide, and cause permanent changes in the oxide-interface charge distribution, degrading the current-voltage characteristics of the MOSFET (Fig. 3.27). Since the likelihood of hot-carrier induced degradation increases with shrinking device dimensions, this problem was identified as one of the important factors that may impose strict limitations on maximum achievable device densities in VLSI circuits.

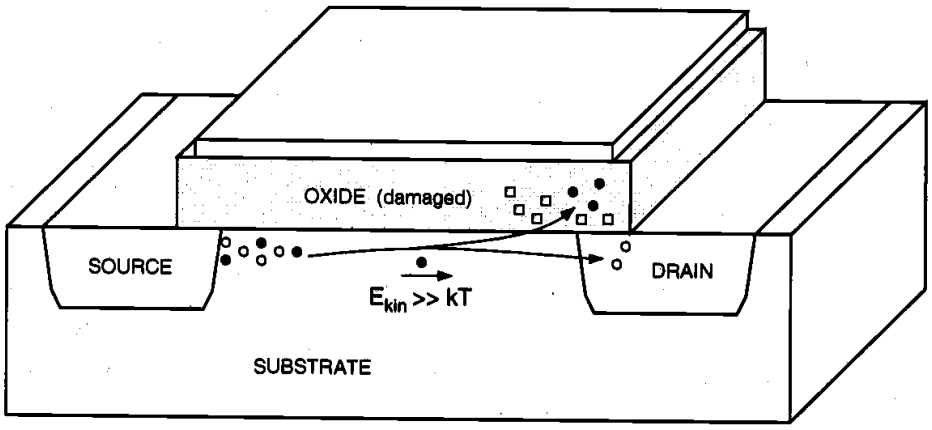
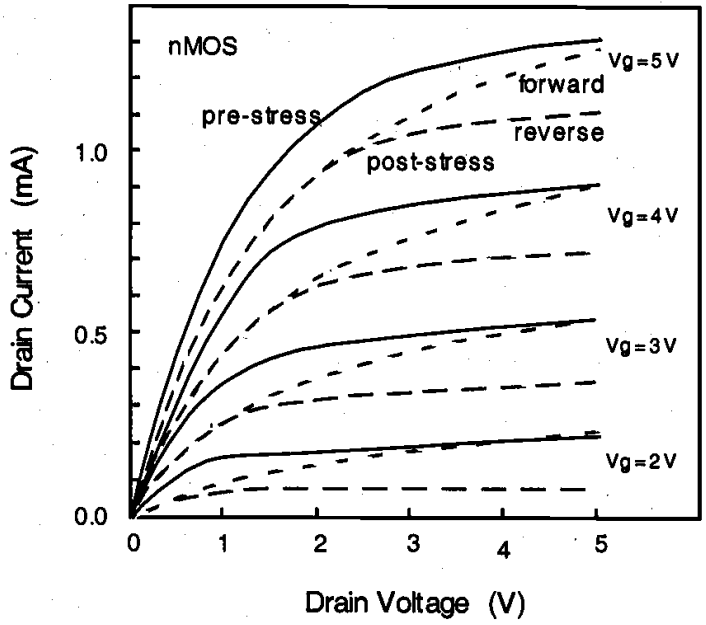


Figure 3.27. Hot-carrier injection into the gate oxide and resulting oxide damage.

The channel hot-electron (CHE) effect is caused by electrons flowing in the channel region, from the source to the drain. This effect is more pronounced at large drain-to-

source voltages, at which the lateral electric field in the drain end of the channel accelerates the electrons. The electrons arriving at the Si-SiO_2 interface with enough kinetic energy to surmount the surface potential barrier are injected into the oxide. Electrons and holes generated by impact ionization also contribute to the charge injection. Note that the channel hot-electron current and the subsequent damage in the gate oxide are localized near the drain junction (Fig. 3.27).



Stress conditions :

V_g = 3 V

V_d = 8 V

stress time = 14 h

Figure 3.28. Typical drain current vs. drain voltage characteristics of an n-channel MOS transistor before and after hot-carrier induced oxide damage.

The hot-carrier induced damage in nMOS transistors has been found to result in either trapping of carriers on defect sites in the oxide or the creation of interface states at the silicon-oxide interface, or both. The damage caused by hot-carrier injection affects the transistor characteristics by causing a degradation in transconductance, a shift in the threshold voltage, and a general decrease in the drain current capability (Fig. 3.28). This performance degradation in the devices leads to the degradation of circuit performance over time. Hence, new MOSFET technologies based on smaller device dimensions must carefully account for the hot-carrier effects and also ensure reliable long-term operation of the devices.

Other reliability concerns for small-geometry devices include interconnect damage through electromigration, electrostatic discharge (ESD) and electrical over-stress (EOS).

3.6. MOSFET Capacitances

The majority of the topics covered in this chapter has been related to the steady-state behavior of the MOS transistor. The current-voltage characteristics investigated here can be applied for investigating the DC response of MOS circuits under various operating conditions. In order to examine the transient (AC) response of MOSFETs and digital

circuits consisting of MOSFETs, on the other hand, we have to determine the nature and the amount of parasitic capacitances associated with the MOS transistor.

The on-chip capacitances found in MOS circuits are in general complicated functions of the layout geometries and the manufacturing processes. Most of these capacitances are not lumped, but *distributed*, and their exact calculations would usually require complex, three-dimensional nonlinear charge-voltage models. In the following, we will develop simple approximations for the on-chip MOSFET capacitances that can be used in most hand calculations. These capacitance models are sufficiently accurate to represent the crucial characteristics of MOSFET charge-voltage behavior, and the equations are all based on fundamental semiconductor device theory, which should be familiar to most readers. We will also stress the distinction between the device-related capacitances and the interconnect capacitances. The capacitive contribution of metal interconnections between various devices is a very important component of the total parasitic capacitance observed in digital circuits. The estimation of this interconnect capacitance will be handled in Chapter 6.

Figure 3.29 shows the cross-sectional view and the top view (mask view) of a typical n-channel MOSFET. Until now, we concentrated on the cross-sectional view of the device, since we were primarily concerned with the flow of carriers within the MOSFET. As we study the parasitic device capacitances, we will have to become more familiar with the top view of the MOSFET. In this figure, the *mask length* (drawn length) of the gate is indicated by L_M, and the actual channel length is indicated by L. The extent of both the gate-source and the gate-drain overlap are L_D; thus, the channel length is given by

$$L = L_M - 2 \cdot L_D \tag{3.93}$$

Note that the source and drain overlap region lengths are usually equal to each other because of the symmetry of the MOSFET structure. Typically, L_D is on the order of 0.1 μm. Both the source and the drain diffusion regions have a width of W. The typical diffusion region length is denoted by Y. Note that both the source diffusion region and the drain diffusion region are surrounded by a p^+ doped region, also called the channel-stop implant. As the name indicates, the purpose of this additional p^+ region is to prevent the formation of any unwanted (parasitic) channels between two neighboring n^+ diffusion regions, i.e., to ensure that the surface between two such regions cannot be inverted. Hence, the p^+ channel-stop implants act to electrically isolate neighboring devices built on the same substrate.

We will identify the parasitic capacitances associated with this typical MOSFET structure as lumped equivalent capacitances *observed* between the device terminals (Fig. 3.30), since such a lumped representation can be easily used to analyze the dynamic transient behavior of the device. The reader must always be reminded, however, that in reality most parasitic device capacitances are due to three-dimensional, distributed charge-voltage relations within the device structure. Based on their physical origins, the parasitic device capacitances can be classified into two major groups: *oxide-related capacitances* and *junction capacitances*. First, the oxide-related capacitances will be considered.

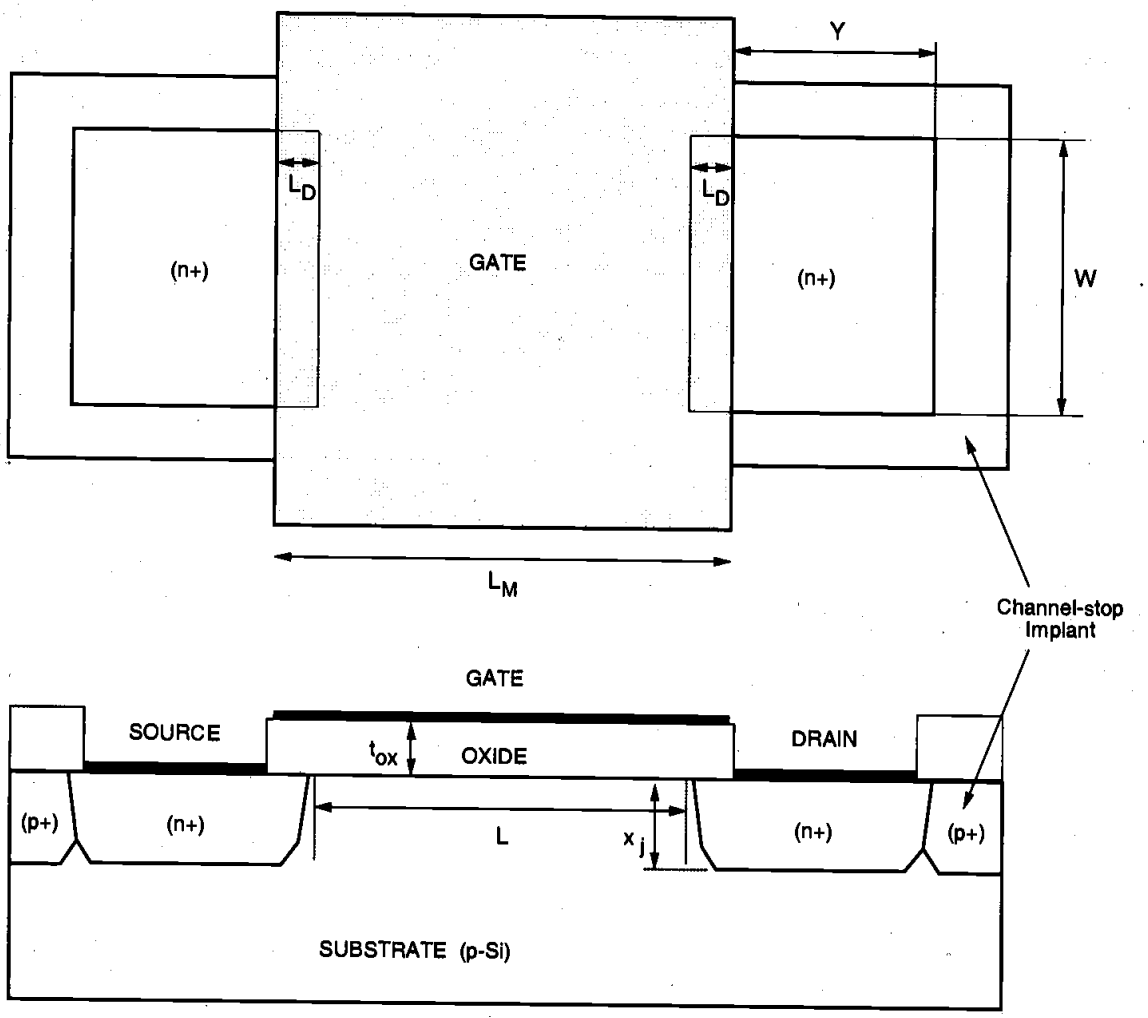


Figure 3.29. Cross-sectional view and top view (mask view) of a typical n-channel MOSFET.

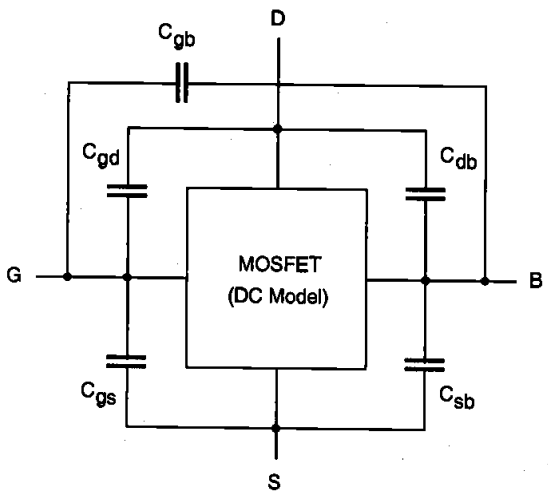


Figure 3.30. Lumped representation of the parasitic MOSFET capacitances.

It was shown earlier that the gate electrode overlaps both the source region and the drain region at the edges. The two overlap capacitances that arise as a result of this structural arrangement are called $C_{GD}\,(overlap)$ and $C_{GS}\,(overlap)$, respectively. Assuming that both the source and the drain diffusion regions have the same width W, the overlap capacitances can be found as

$$\begin{aligned} C_{GS}(overlap) &= C_{ox} \cdot W \cdot L_D \\ C_{GD}(overlap) &= C_{ox} \cdot W \cdot L_D \end{aligned} \tag{3.94}$$

with

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \tag{3.95}$$

Note that both of these overlap capacitances do not depend on the bias conditions, i.e., they are voltage-independent.

Now consider the capacitances which result from the interaction between the gate voltage and the channel charge. Since the channel region is connected to the source, the drain, and the substrate, we can identify three capacitances between the gate and these regions, i.e., C_{gs}, C_{gd}, and C_{gb}, respectively. Notice that in reality, the gate-to-channel capacitance is distributed and voltage-dependent. Then, the gate-to-source capacitance C_{gs} is actually the gate-to-channel capacitance *seen* between the gate and the source terminals; the gate-to-drain capacitance C_{gd} is actually the gate-to-channel capacitance *seen* between the gate and the drain terminals. A simplified view of their bias-dependence can be obtained by observing the conditions in the channel region during cut-off, linear, and saturation modes.

In cut-off mode (Fig. 3.31(a)), the surface is not inverted. Consequently, there is no conducting channel that links the surface to the source and to the drain. Therefore, the gate-to-source and the gate-to-drain capacitances are both equal to zero: $C_{gs} = C_{gd} = 0$. The gate-to-substrate capacitance can be approximated by

$$C_{gb} = C_{ox} \cdot W \cdot L \tag{3.96}$$

In linear-mode operation, the inverted channel extends across the MOSFET, between the source and the drain (Fig. 3.31(b)). This conducting inversion layer on the surface effectively shields the substrate from the gate electric field; thus, $C_{gb} = 0$. In this case, the distributed gate-to-channel capacitance may be viewed as being shared equally between the source and the drain, yielding

$$C_{gs} \cong C_{gd} \cong \frac{1}{2} \cdot C_{ox} \cdot W \cdot L \tag{3.97}$$

When the MOSFET is operating in saturation mode, the inversion layer on the surface does not extend to the drain, but it is pinched off (Fig. 3.31(c)). The gate-to-drain

capacitance component is therefore equal to zero ($C_{gd}=0$). Since the source is still linked to the conducting channel, its shielding effect also forces the gate-to-substrate capacitance to be zero, $C_{gb} = 0$. Finally, the distributed gate-to-channel capacitance as seen between the gate and the source can be approximated by

$$C_{gs} \cong \frac{2}{3} \cdot C_{ox} \cdot W \cdot L \tag{3.98}$$

Table 3.6 lists a summary of the approximate oxide capacitance values in three different operating modes of the MOSFET. The variation of the distributed parasitic oxide capacitances as functions of the gate-to-source voltage V_{GS} is also shown in Fig. 3.32.

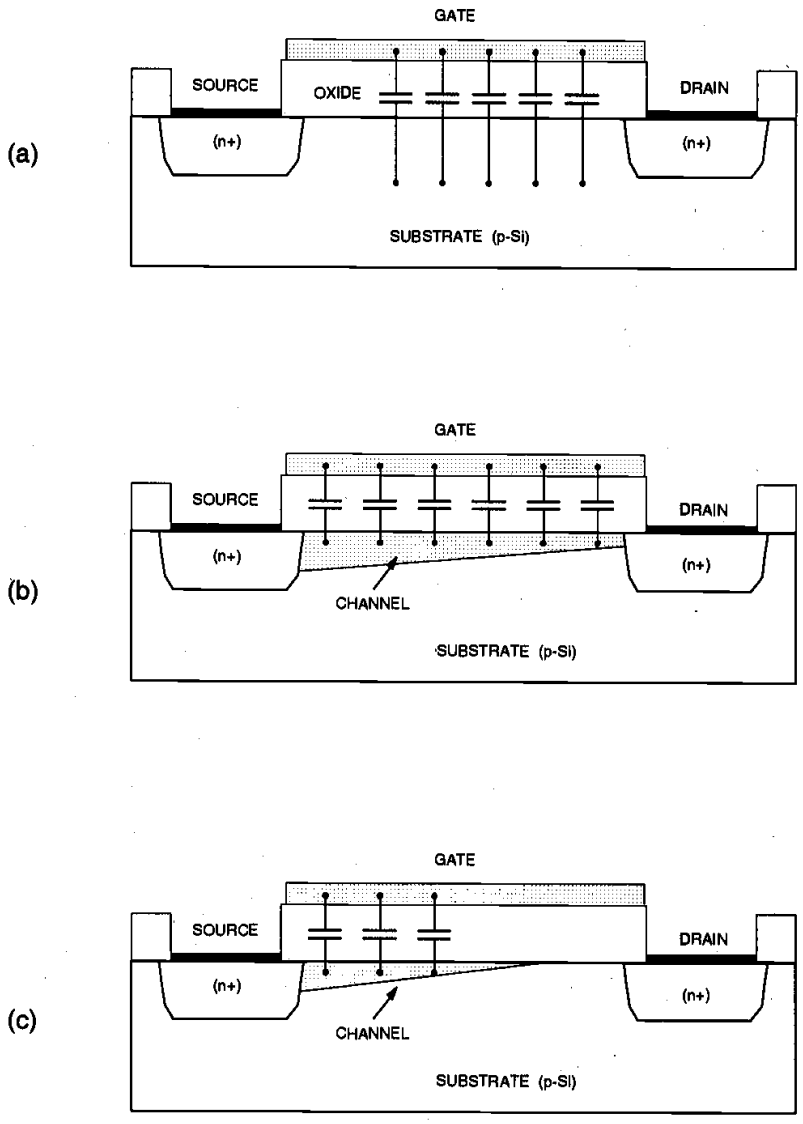


Figure 3.31. Schematic representation of MOSFET oxide capacitances during (a) cut-off, (b) linear, and (c) saturation modes.

Obviously, we have to combine the distributed C_{gs} and C_{gd} values found here with the relevant overlap capacitance values, in order to calculate the total capacitance between the external device terminals. It is also worth mentioning that the sum of all three voltage-dependent (distributed) gate oxide capacitances ($C_{gb} + C_{gs} + C_{gd}$) has a minimum value of $0.66\,C_{ox}WL$ (in saturation mode) and a maximum value of $C_{ox}WL$ (in cut-off and linear modes). For simple hand calculations where all three capacitances can be considered to be connected in parallel, a constant worst-case value of $C_{ox}W(L+2L_D)$ can be used for the sum of MOSFET gate oxide capacitances.

Capacitance	Cut-off	Linear	Saturation
C_{gb} (total)	$C_{ox}WL$	0	0
C_{gd} (total)	$C_{ox}WL_D$	$\frac{1}{2}C_{ox}WL + C_{ox}WL_D$	$C_{ox}WL_D$
C_{gs} (total)	$C_{ox}WL_D$	$\frac{1}{2}C_{ox}WL + C_{ox}WL_D$	$\frac{2}{3}C_{ox}WL + C_{ox}WL_D$

Table 3.6. Approximate oxide capacitance values for three operating modes of the MOS transistor.

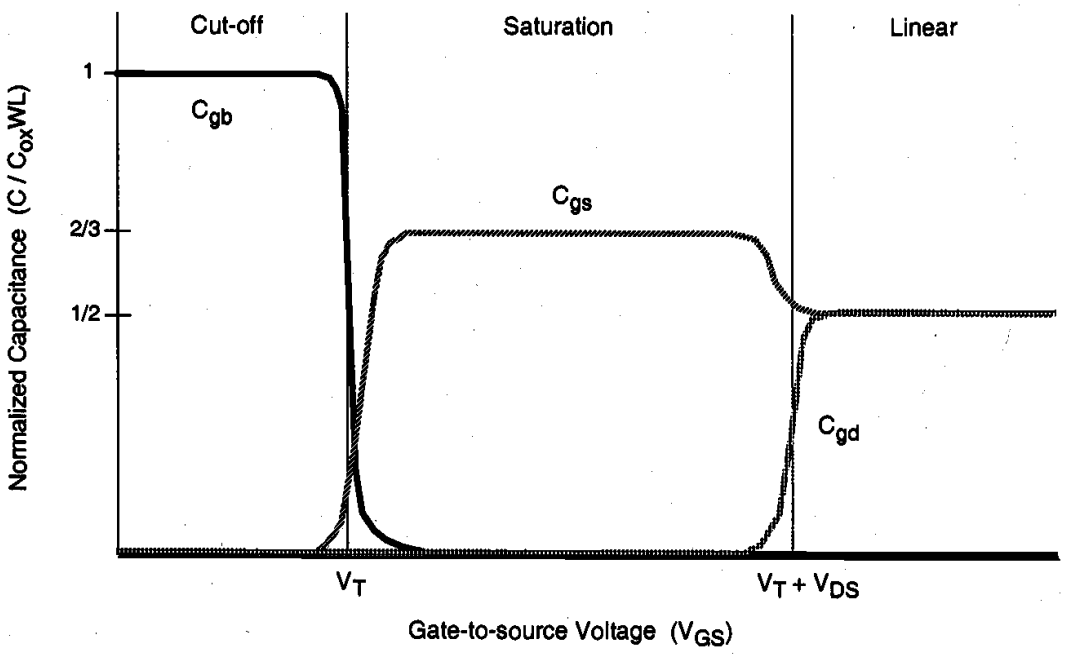


Figure 3.32. Variation of the distributed (gate-to-channel) oxide capacitances as functions of gate-to-source voltage V_{GS}.

Now we consider the voltage-dependent source-substrate and drain-substrate junction capacitances, C_{sb} and C_{db}, respectively. Both of these capacitances are due to the depletion charge surrounding the respective source or drain diffusion regions embedded in the substrate. The calculation of the associated junction capacitances is complicated by the three-dimensional shape of the diffusion regions that form the source-substrate and the drain-substrate junctions. Note that both of these junctions are reverse-biased under normal operating conditions of the MOSFET and that the amount of junction capacitance is a function of the applied terminal voltages. Figure 3.33 shows the simplified, partial geometry of a typical n-channel enhancement MOSFET, focusing on the n-type diffusion region within the p-type substrate. The analysis to be carried out in the following will apply to both n-channel and p-channel MOS transistors.

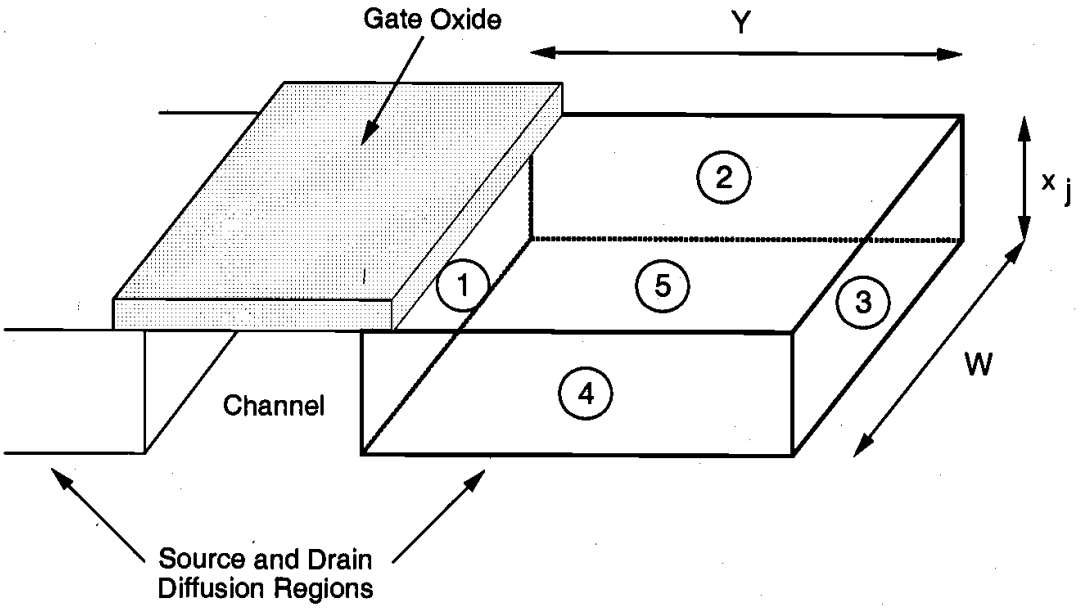


Figure 3.33. Three-dimensional view of the n^+ diffusion region within the p-type substrate.

As seen in Fig. 3.33, the n^+ diffusion region forms a number of planar pn-junctions with the surrounding p-type substrate, indicated here with 1 through 5. The dimensions of the rectangular box representing the diffusion region are given as W, Y, and x_j. Abrupt (step) pn-junction profiles will be assumed for all junctions for simplicity. Also, comparing this three-dimensional view with Fig. 3.29, we recognize that three of the five planar junctions shown here (2, 3, and 4) are actually surrounded by the p^+ channel-stop implant. The junction labeled (1) is facing the channel, and the bottom junction (5) is facing the p-type substrate, which has a doping density of N_A. Since the p^+ channel-stop implant density is usually about $10N_A$, the junction capacitances associated with these *sidewalls* will be different from the other junction capacitances (see Table 3.7). Note that in general, the actual shape of the diffusion regions as well as the doping profiles are much

more complicated. However, this simplified analysis provides sufficient insight for the first-order estimation of junction-related capacitances.

Junction	Area	Type
1	$W \cdot x_j$	n^+/p
2	$Y \cdot x_j$	n^+/p^+
3	$W \cdot x_j$	n^+/p^+
4	$Y \cdot x_j$	n^+/p^+
5	$W \cdot Y$	n^+/p

Table 3.7. Types and areas of the pn-junctions shown in Figure 3.33.

To calculate the depletion capacitance of a reverse-biased abrupt pn-junction, consider first the depletion region thickness, x_d. Assuming that the n-type and p-type doping densities are given by N_D and N_A, respectively, and that the reverse bias voltage is given by V (negative), the depletion region thickness can be found as follows:

$$x_d = \sqrt{\frac{2 \cdot \varepsilon_{Si}}{q} \cdot \frac{N_A + N_D}{N_A \cdot N_D} \cdot (\phi_0 - V)} \tag{3.99}$$

where the built-in junction potential is calculated as

$$\phi_0 = \frac{kT}{q} \cdot \ln\left(\frac{N_A \cdot N_D}{n_i^2}\right) \tag{3.100}$$

Note that the junction is forward-biased for a *positive* bias voltage V, and reverse-biased for a *negative* bias voltage. The depletion-region charge stored in this area can be written in terms of the depletion region thickness, x_d.

$$Q_j = A \cdot q \cdot \left(\frac{N_A \cdot N_D}{N_A + N_D}\right) \cdot x_d = A\sqrt{2 \cdot \varepsilon_{Si} \cdot q \cdot \left(\frac{N_A \cdot N_D}{N_A + N_D}\right) \cdot (\phi_0 - V)} \tag{3.101}$$

Here, A indicates the junction area. The junction capacitance associated with the depletion region is defined as

$$C_j = \left|\frac{dQ_j}{dV}\right| \tag{3.102}$$

By differentiating (3.101) with respect to the bias voltage V, we can now obtain the expression for the junction capacitance as follows.

$$C_j(V) = A \cdot \sqrt{\frac{\varepsilon_{Si} \cdot q}{2} \cdot \left(\frac{N_A \cdot N_D}{N_A + N_D} \right)} \cdot \frac{1}{\sqrt{\phi_0 - V}} \tag{3.103}$$

This expression can be rewritten in a more general form, to account for the junction grading.

$$C_j(V) = \frac{A \cdot C_{j0}}{\left(1 - \frac{V}{\phi_0} \right)^m} \tag{3.104}$$

The parameter m in (3.104) is called the *grading coefficient*. Its value is equal to 1/2 for an abrupt junction profile, and 1/3 for a linearly graded junction profile. Obviously, for an abrupt pn-junction profile, i.e., for $m = 1/2$, the equations (3.103) and (3.104) become identical. The zero-bias junction capacitance per unit area C_{j0} is defined as

$$C_{j0} = \sqrt{\frac{\varepsilon_{Si} \cdot q}{2} \cdot \left(\frac{N_A \cdot N_D}{N_A + N_D} \right) \cdot \frac{1}{\phi_0}} \tag{3.105}$$

Note that the value of the junction capacitance C_j given by (3.104) ultimately depends on the external bias voltage that is applied across the pn-junction. Since the terminal voltages of a MOSFET will change during dynamic operation, accurate estimation of the junction capacitances under transient conditions is quite complicated; the instantaneous values of all junction capacitances will also change accordingly. The problem of estimating capacitance values under changing bias conditions can be simplified, if we calculate a large-signal average (linear) junction capacitance instead, which, by definition, is independent of the bias potential. This *equivalent large-signal capacitance* can be defined as follows:

$$C_{eq} = \frac{\Delta Q}{\Delta V} = \frac{Q_j(V_2) - Q_j(V_1)}{V_2 - V_1} = \frac{1}{V_2 - V_1} \cdot \int_{V_1}^{V_2} C_j(V)\, dV \tag{3.106}$$

Here, the reverse bias voltage across the pn-junction is assumed to change from V_1 to V_2. Hence, the equivalent capacitance C_{eq} is always calculated for a *transition between two known voltage levels.* By substituting (3.104) into (3.106), we obtain

$$C_{eq} = -\frac{A \cdot C_{j0} \cdot \phi_0}{(V_2 - V_1) \cdot (1 - m)} \cdot \left[\left(1 - \frac{V_2}{\phi_0} \right)^{1-m} - \left(1 - \frac{V_1}{\phi_0} \right)^{1-m} \right] \tag{3.107}$$

For the special case of abrupt pn-junctions, equation (3.107) becomes

$$C_{eq} = -\frac{2 \cdot A \cdot C_{j0} \cdot \phi_0}{(V_2 - V_1)} \cdot \left[\sqrt{1 - \frac{V_2}{\phi_0}} - \sqrt{1 - \frac{V_1}{\phi_0}} \right] \tag{3.108}$$

This equation can be rewritten in a simpler form by defining a dimensionless coefficient K_{eq}, as follows:

$$C_{eq} = A \cdot C_{j0} \cdot K_{eq} \tag{3.109}$$

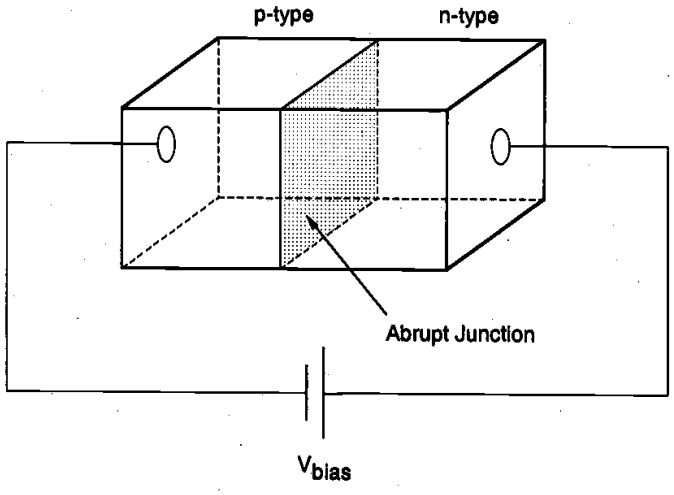
$$K_{eq} = -\frac{2\sqrt{\phi_0}}{V_2 - V_1} \cdot \left(\sqrt{\phi_0 - V_2} - \sqrt{\phi_0 - V_1} \right) \tag{3.110}$$

where K_{eq} is the *voltage equivalence factor* (note that $0 < K_{eq} < 1$). Thus, the coefficient K_{eq} allows us to take into account the voltage-dependent variations of the junction capacitance. The accuracy of the large-signal equivalent junction capacitance C_{eq} found by using (3.109) and (3.110) is usually sufficient for most first-order hand calculations. Practical applications of the capacitance calculation methods discussed here will be illustrated in the following examples.

Example 3.7.

Consider a simple abrupt pn-junction, which is reverse-biased with a voltage V_{bias}. The doping density of the n-type region is $N_D = 10^{19}$ cm^{-3}, and the doping density of the p-type region is given as $N_A = 10^{16}$ cm^{-3}. The junction area is $A = 20$ μm x 20 μm.

First, we will calculate the zero-bias junction capacitance per unit area, C_{j0}, for this structure. The built-in junction potential is found as



$$\phi_0 = \frac{kT}{q} \cdot \ln\left(\frac{N_A \cdot N_D}{n_i^2}\right) = 0.026\ \text{V} \cdot \ln\left(\frac{10^{16} \cdot 10^{19}}{2.1 \times 10^{20}}\right) = 0.88\ \text{V}$$

Using (3.105), we can calculate the zero-bias junction capacitance :

$$C_{j0} = \sqrt{\frac{\varepsilon_{Si} \cdot q}{2} \cdot \left(\frac{N_A \cdot N_D}{N_A + N_D}\right) \cdot \frac{1}{\phi_0}}$$

$$= \sqrt{\frac{11.7 \cdot 8.85 \times 10^{-14}\ \text{F/cm} \cdot 1.6 \times 10^{-19}\ \text{C}}{2} \cdot \left(\frac{10^{16} \cdot 10^{19}}{10^{16} + 10^{19}}\right) \cdot \frac{1}{0.88\ \text{V}}}$$

$$= 3.1 \times 10^{-8}\ \text{F/cm}^2$$

Next, find the equivalent large-signal junction capacitance assuming that the reverse bias voltage changes from $V_1 = 0$ to $V_2 = -5$ V. The voltage equivalence factor for this transition can be found as follows:

$$K_{eq} = -\frac{2\sqrt{\phi_0}}{V_2 - V_1} \cdot \left(\sqrt{\phi_0 - V_2} - \sqrt{\phi_0 - V_1}\right)$$

$$= -\frac{2\sqrt{0.88}}{-5} \cdot \left(\sqrt{0.88 - (-5)} - \sqrt{0.88}\right) = 0.56$$

Then, the average junction capacitance can be found simply by using (3.109).

$$C_{eq} = A \cdot C_{j0} \cdot K_{eq} = 400 \times 10^{-8}\ \text{cm}^2 \cdot 3.1 \times 10^{-8}\ \text{F/cm}^2 \cdot 0.56 = 69\ \text{fF}$$

It was shown in Fig. 3.29 and Fig. 3.33 that the sidewalls of a typical MOSFET source or drain diffusion region are surrounded by a p^+ channel-stop implant, with a higher doping density than the substrate doping density N_A. Consequently, the sidewall zero-bias capacitance C_{j0sw}, as well as the sidewall voltage equivalence factor $K_{eq}(sw)$ will be different from those of the bottom junction. Assuming that the sidewall doping density is given by $N_A(sw)$, the zero-bias capacitance per unit area can be found as follows:

$$C_{j0sw} = \sqrt{\frac{\varepsilon_{Si} \cdot q}{2} \cdot \left(\frac{N_A(sw) \cdot N_D}{N_A(sw) + N_D}\right) \cdot \frac{1}{\phi_{0sw}}} \qquad (3.111)$$

where ϕ_{0sw} is the built-in potential of the sidewall junctions. Since all sidewalls in a typical diffusion structure have approximately the same depth of x_j, we can define a zero-bias sidewall junction capacitance per unit length.

$$C_{jsw} = C_{j0sw} \cdot x_j \tag{3.112}$$

The sidewall voltage equivalence factor $K_{eq}(sw)$ for a voltage swing between V_1 and V_2 is defined as follows:

$$K_{eq}(sw) = -\frac{2\sqrt{\phi_{0sw}}}{V_2 - V_1} \cdot \left(\sqrt{\phi_{0sw} - V_2} - \sqrt{\phi_{0sw} - V_1}\right) \tag{3.113}$$

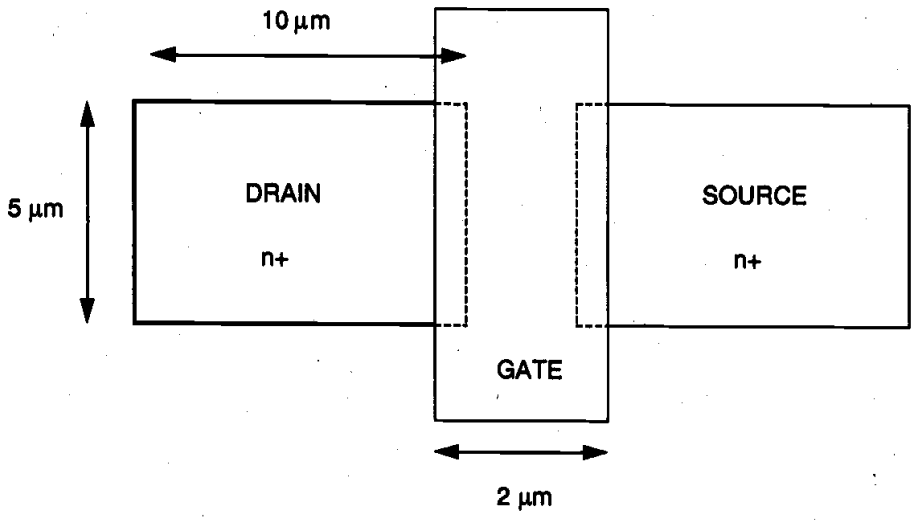
Combining the equations (3.111) through (3.113), the equivalent large-signal junction capacitance $C_{eq}(sw)$ for a sidewall of length (perimeter) P can be calculated as

$$C_{eq}(sw) = P \cdot C_{jsw} \cdot K_{eq}(sw) \tag{3.114}$$

Example 3.8.

Consider the n-channel enhancement-type MOSFET shown below. The process parameters are given as follows:

Substrate doping	$N_A = 2 \times 10^{15}$ cm^{-3}
Source / drain doping	$N_D = 10^{19}$ cm^{-3}
Sidewall (p+) doping	$N_A(sw) = 4 \times 10^{16}$ cm^{-3}
Gate oxide thickness	$t_{ox} = 45$ nm
Junction depth	$x_j = 1.0$ μm



Note that both the source and the drain diffusion regions are surrounded by p^+ channel-stop diffusion. The substrate is biased at 0 V. Assuming that the drain voltage is changing from 0.5 V to 5 V, find the average drain-substrate junction capacitance C_{db}.

First, we recognize that three sidewalls of the rectangular drain diffusion structure form n^+/p^+ junctions with the p^+ channel-stop implant, while the bottom area and the sidewall facing the channel form n^+/p junctions. Start by calculating the built-in potentials for both types of junctions.

$$\phi_0 = \frac{kT}{q} \cdot \ln\left(\frac{N_A \cdot N_D}{n_i^2}\right) = 0.026\ \text{V} \cdot \ln\left(\frac{2\times10^{15}\cdot10^{19}}{2.1\times10^{20}}\right) = 0.837\ \text{V}$$

$$\phi_{0sw} = \frac{kT}{q} \cdot \ln\left(\frac{N_A(sw) \cdot N_D}{n_i^2}\right) = 0.026\ \text{V} \cdot \ln\left(\frac{4\times10^{16}\cdot10^{19}}{2.1\times10^{20}}\right) = 0.915\ \text{V}$$

Next, we calculate the zero-bias junction capacitances per unit area:

$$C_{j0} = \sqrt{\frac{\varepsilon_{Si} \cdot q}{2} \cdot \left(\frac{N_A \cdot N_D}{N_A + N_D}\right) \cdot \frac{1}{\phi_0}}$$

$$= \sqrt{\frac{11.7 \cdot 8.85\times10^{-14}\,\text{F/cm} \cdot 1.6\times10^{-19}\,\text{C}}{2} \cdot \left(\frac{2\times10^{15}\cdot10^{19}}{2\times10^{15}+10^{19}}\right) \cdot \frac{1}{0.837\ \text{V}}}$$

$$= 1.41\times10^{-8}\ \text{F/cm}^2$$

$$C_{j0sw} = \sqrt{\frac{\varepsilon_{Si} \cdot q}{2} \cdot \left(\frac{N_A(sw) \cdot N_D}{N_A(sw) + N_D}\right) \cdot \frac{1}{\phi_{0sw}}}$$

$$= \sqrt{\frac{11.7 \cdot 8.85\times10^{-14}\,\text{F/cm} \cdot 1.6\times10^{-19}\,\text{C}}{2} \cdot \left(\frac{4\times10^{16}\cdot10^{19}}{4\times10^{16}+10^{19}}\right) \cdot \frac{1}{0.915\ \text{V}}}$$

$$= 6.01\times10^{-8}\ \text{F/cm}^2$$

The zero-bias sidewall junction capacitance per unit length can also be found as follows.

$$C_{jsw} = C_{j0sw} \cdot x_j = 6.01\times10^{-8}\ \text{F/cm}^2 \cdot 10^{-4}\ \text{cm} = 6.01\ \text{pF/cm}$$

In order to take the given drain voltage variation into account, we must now calculate the voltage equivalence factors, K_{eq} and $K_{eq}(sw)$, for both types of junctions. This will allow us to find the average large-signal capacitance values.