

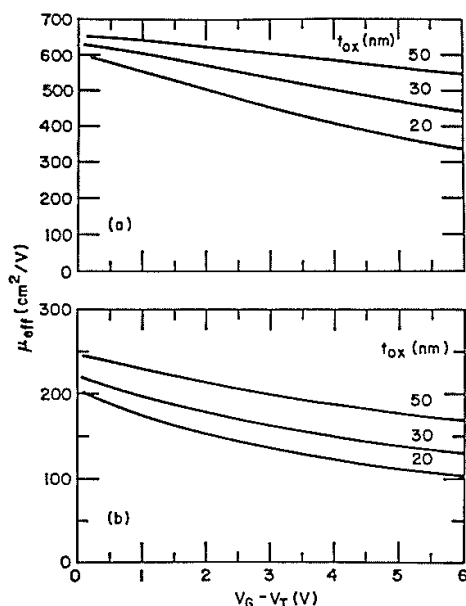


Fig. 5-34 Calculated μ_{eff} of current carrier versus $(V_{GS} - V_T)$ for three different values of t_{ox} : (a) NMOS device and (b) PMOS devices. The "buried-channel" parameters in Table 5-2 are used for PMOS. (From ³⁶ © 1986 IEEE) and ⁴⁶ © 1988 IEEE).

$$\text{Model II} \quad v = \mu_{eff} \mathcal{E} / \{1 + (\mathcal{E}/\mathcal{E}_{sat})\} \quad (5 - 30)$$

A comparison of these two models with experimental data for electrons is shown in Fig. 5-35.⁴³ It is seen that Model I fits the measured data quite well, while Model II significantly underestimates the velocity at moderate fields (the same value of saturation velocity is used in both models plotted in Fig. 5-35). On this basis Model I would be preferred for estimating v . However, if Model I was inserted into an expression to calculate I_D , a solution could only be obtained by carrying out a prohibitively complex calculation.⁴⁴ Thus, to improve the accuracy of the carrier velocity model without incurring the complexity of Model I, a third, simpler, two-region piecewise model can be used. This model is defined by

$$\text{Model III} \quad v = \mu_{eff} \mathcal{E} / \{1 + (\mathcal{E}/\mathcal{E}_{sat})\} \quad \mathcal{E} < \mathcal{E}_{sat} \quad (5 - 31a)$$

$$= v_{sat} \quad \mathcal{E} > \mathcal{E}_{sat} \quad (5 - 31b)$$

By using this model in the expression which calculates I_D a closed-form solution of the I_D - V_{DS} characteristics of the MOSFET is obtained. In addition, the carrier velocity saturates at $\mathcal{E}_{sat}$. $\mathcal{E}_{sat}$ can be expressed in terms of v_{sat} and μ_{eff} by matching the carrier velocities given by Eqs. 5-31a and 5-31b at $\mathcal{E} = \mathcal{E}_{sat}$. Therefore,

$$v_{sat} = \mu_{eff} \mathcal{E}_{sat} / [1 + (\mathcal{E}_{sat} / \mathcal{E}_{sat})] \quad (5-32)$$

or

$$\mathcal{E}_{sat} = 2 v_{sat} / \mu_{eff}. \quad (5-33)$$

This model is also plotted in Fig. 5-35 over a wide range of electric fields. Note that it slightly overestimates the carrier velocity near $\mathcal{E}_{sat}$.

5.4.2 A Simple, Analytic DC Circuit Model of the Short-Channel MOSFET in Strong Inversion

Now that we have described the influence of the transverse and lateral fields on the velocity of the channel carriers, we can use the models of these effects to modify the dc circuit models of the MOSFET operating beyond threshold. In this way we develop a simple analytical dc circuit model that is valid for short-channel MOSFETs beyond threshold, and which can be used for first-order, hand calculations of I_D - V_{DS} and g_m for devices with channel lengths down to $\sim 1 \mu\text{m}$.^{45,46}

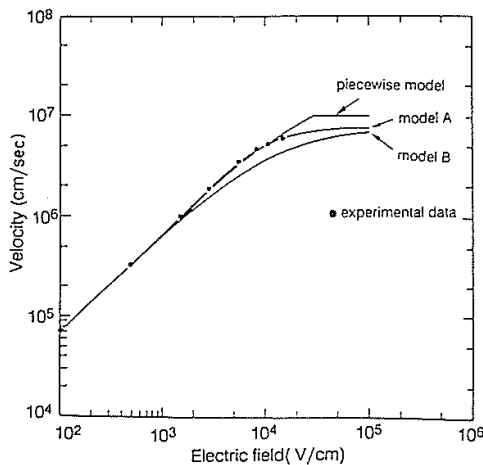


Fig. 5-35 Comparison of velocity-field models: models A and B as given in Eqs. 5-29 and 5-30, respectively. The piecewise model is given in Eq. 5-31, and $\mu_0 = 710 \text{ cm}^2/\text{V-sec}$, $v_{sat} = 1 \times 10^7 \text{ cm/sec}$.³⁸ Reprinted with permission of *Solid State Electronics*.

We first note that the factors resulting from both mobility degradation due to transverse fields and velocity saturation effects are outside of the integral in the equations used to calculate I_D (see Eqs. 4-16 and 4-19). Hence, they merely serve as multiplicative factors if incorporated into any of the long-channel dc models described in chap. 4. This means that any one of the long-channel models could be used to demonstrate the impact of these mobility effects the I_D - V_{DS} characteristics. Here, for mathematical simplicity, we elect to use the square-law model of section 4.5.5 for this purpose. To complete the dc short-channel MOSFET model, however, we will later still need to add some discussion of the short-channel effects that occur when the MOSFET is operated in saturation (i.e., the effects of channel-length modulation).

5.4.2.1 DC Circuit Model of the Short-Channel MOSFET in the Linear Regime. As noted above, all of the long-channel MOSFET dc model equations can be modified in the same manner to take into account both effects described above. For simplicity, the square-law long-channel model is chosen to be modified. In the linear-region, the modified (i.e., short-channel) version of dc I_D - V_{DS} equation of the square-law model (Eq. 4-68) is

$$I_D = \frac{\mu_{\text{eff}} C_{\text{ox}} \frac{Z}{L}}{\left(1 + \frac{V_{DS}}{E_{\text{sat}} L}\right)} \left[(V_{GS} - V_T) V_{DS} - \frac{V_{DS}^2}{2} \right] \quad (5-34)$$

Taking the ratio of the two currents given by Eqs. 5-22 and 4-68, we get

$$I_D (5-34) / I_D (4-68) = (\mu_{\text{eff}} / \mu_n) [1 / \{1 + (V_{DS} / E_{\text{sat}} L)\}] \quad (5-35)$$

Note that since E_{sat} and L appear as a product term in both Eqs. 5-22 and 5-23, a MOSFET is "long-channel-like" if either E_{sat} or L is large. Therefore, a MOSFET with a thinner gate oxide is more "long-channel-like" because the thinner oxide will produce a larger transverse field, making the carrier mobility lower so that E_{sat} is high.

The saturation drain voltage V_{DSSat} can next be calculated by equating the current at the drain to the current given by Eq. 5-34 when $V_{DS} = V_{DSSat}$. The current at the drain is given by the product of the carrier density and the drift velocity. The charge density at the drain, assuming that the GCA is still valid, is $ZQ_1 = C_{\text{ox}}(V_{GS} - V_T - V_{DSSat})$ and the drift velocity is v_{sat} . Therefore the current is

$$I_D = Z v_{\text{sat}} C_{\text{ox}} [V_{GS} - V_T - V_{DSSat}] \quad (5-36a)$$

$$= \frac{Z}{L \mu_{\text{eff}} C_{\text{ox}}} [V_{GS} - V_T - (V_{DSSat}/2)] V_{DSSat} \left[\frac{1}{1 + (V_{DS}/E_{\text{sat}} L)} \right] \quad (5-36b)$$

After some algebra, an expression for V_{DSSat} is obtained

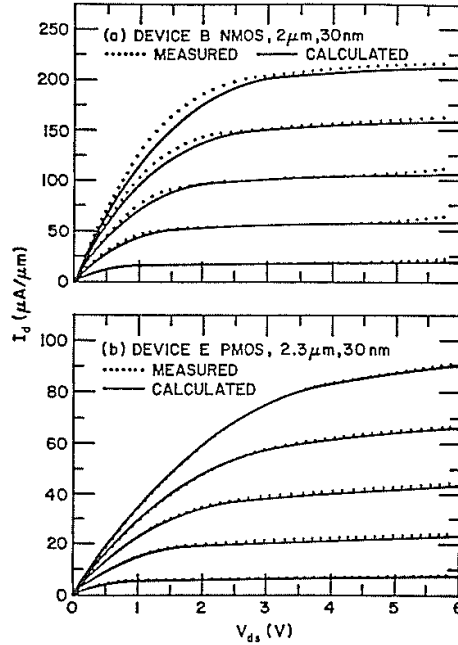


Fig. 5-36 Measured and calculated I-V characteristics for a (a) NMOS device and (b) a PMOS device.⁴⁶ (© 1988 IEEE).

$$V_{Dssat} = \frac{\mathcal{E}_{sat} L [V_{GS} - V_T]}{(V_{GS} - V_T) + \mathcal{E}_{sat} L} \quad (5 - 37)$$

where $\mathcal{E}_{sat}$ is the critical field for velocity saturation ($\sim 4 \times 10^4$ V/cm).

The saturation drain current can be determined by substituting Eq. 5-37 into Eq. 5-36a to get

$$I_{Dsat} = Z v_{sat} C_{ox} \frac{[V_{GS} - V_T]^2}{(V_{GS} - V_T) + \mathcal{E}_{sat} L} \quad (5 - 38)$$

The small-signal saturation transconductance g_{msat} can be found by differentiating Eq. 5-36a with respect to V_{GS} :

$$g_{msat} = Z v_{sat} C_{ox} [1 - (\partial V_{Dssat} / \partial V_{GS})] \quad (5 - 39)$$

Equations 5-36, 5-38, and 5-39 form the core equations of this simple model for calculating the dc drain characteristics for short-channel MOSFETs. Some modeling results using these equations are shown in Figs. 5-36a and b.⁴⁶ (Note that the I_D values for $V_{DS} > V_{Dssat}$ in these curves are calculated using a saturation region model to be

described in the next section.) Good agreement with measured data is seen in these figures, and other work has shown good agreement for devices with channel lengths as small as $0.2 \mu\text{m}$.⁴⁷ For example, agreement between calculated and measured values of g_m is better than ± 6 percent over most of the gate voltage range for channel lengths from $3 \mu\text{m}$ down to $1 \mu\text{m}$.

Equation 5-37 indicates that as the channel length is decreased, the channel field at the drain end will reach $\mathcal{E}_{\text{sat}}$ at a lower value of V_{DS} . Thus, V_{DSSat} is predicted to decrease with decreasing L as well. When the gate oxide thickness is made smaller, μ_{eff} decreases because $\mathcal{E}_{\text{eff}}$ is larger (see Eq. 5-27). Consequently, $\mathcal{E}_{\text{sat}}$ will increase, and so will V_{DSSat} . MOSFETs with thin gate oxides thus appear more "long-channel-like," as was discussed earlier.

The short-channel dc model of the MOSFET given by Eq. 5-36 also predicts four effects that impact the value of I_{DSat} in short-channel devices not addressed by the long-channel dc models. First, mobility degradation reduces I_{DSat} in short-channel devices by a factor (μ_{eff}/μ_0) . Since the effective mobility decreases as V_{GS} increases, this indicates that I_{DSat} can be significantly reduced for large gate voltages. Second, I_{DSat} is also reduced by velocity saturation effects - to a value equivalent to the one predicted by the long-channel models (which assume a constant value of mobility, e.g., 4-68), except that the channel length appears longer than L by the factor $V_{\text{DS}}/\mathcal{E}_{\text{sat}}$.

Third, Eq. 5-36 indicates that if the product $\mathcal{E}_{\text{sat}}L$ is large compared to $(V_{\text{GS}} - V_{\text{T}})$, then I_{DSat} varies quadratically with $(V_{\text{GS}} - V_{\text{T}})$. On the other hand, if $(V_{\text{GS}} - V_{\text{T}}) \gg \mathcal{E}_{\text{sat}}L$ then I_{DSat} becomes linearly proportional to $(V_{\text{GS}} - V_{\text{T}})$. Since one of the criteria in Table 5-1 for long-channel behavior is that $I_{\text{DSat}} \propto (V_{\text{GS}} - V_{\text{T}})^2$, long-channel behavior will be exhibited by devices in which either L or $\mathcal{E}_{\text{sat}}$ is large. On the other hand, if L or $\mathcal{E}_{\text{sat}}$ are both small, such that $\mathcal{E}_{\text{sat}}L \ll (V_{\text{GS}} - V_{\text{T}})$, the saturation regime behavior of the MOSFET will be "short-channel-like." Note that even a very short-channel device will exhibit some long-channel behavior if $(V_{\text{GS}} - V_{\text{T}})$ is small enough. The drain current under this situation is given by

$$I_{\text{DSat}} = [\mu_{\text{eff}} Z C_{\text{ox}} / 2 L] (V_{\text{GS}} - V_{\text{T}})^2 \quad \text{if } \mathcal{E}_{\text{sat}}L \gg (V_{\text{GS}} - V_{\text{T}}) \quad (5-40)$$

In shorter channel MOSFETs there is an observable trend toward a linear variation of I_{DSat} with $(V_{\text{GS}} - V_{\text{T}})$ instead of the quadratic dependence, as is predicted without velocity saturation. This behavior is illustrated in Fig. 5-37a.

The fourth effect predicted by Eq. 5-36 (as noted in the introduction of this chapter) is observed in the limit of very strong velocity saturation (i.e., when the channel length is very short, such that V_{DSSat} approaches zero). In such cases, Eq. 5-36 indicates that the drain current is independent of L , and I_{DSat} is given by

$$I_{\text{DSat}} = [\nu_{\text{sat}} Z C_{\text{ox}}] (V_{\text{GS}} - V_{\text{T}}) \quad (5-41a)$$

Under this condition, even the carriers at the source [whose density is given by $C_{\text{ox}}(V_{\text{GS}} - V_{\text{T}})$] are velocity saturated.

However, I_{Dsat} of a device with practical geometries is always lower than the maximum predicted by Eq. 5-41b. As a result, an ideality factor K has been defined as a useful figure of merit which indicates how close I_{Dsat} in a real device comes to that in an ideal device. This factor is inserted in Eq. 5-41b to yield

$$I_{Dsat} = K [\nu_{sat} Z C_{ox}] (V_{GS} - V_T) \quad \text{where } K < 1 \quad (5-41b)$$

Then if we compare Eqs. 5-36 and 5-41b, we obtain a definition of as

$$K = (V_{GS} - V_T - V_{Dsat}) / (V_{GS} - V_T) \quad (5-41c)$$

This factor is a very handy parameter for device performance evaluation. That is, if an NMOS device with $t_{ox} = 60$ nm and $L = 1$ μm saturates at a V_{DS} of 1.3V when $(V_{GS} - V_T) = 4.3$, the value of K is $(4.3 - 1.3)/4.3 = 0.72$. This implies that shrinking the device would improve its driving current by no more than 30%. The K factor given by Eq. 5-41c is plotted against L for two different oxide thicknesses in Fig. 5-37c. The

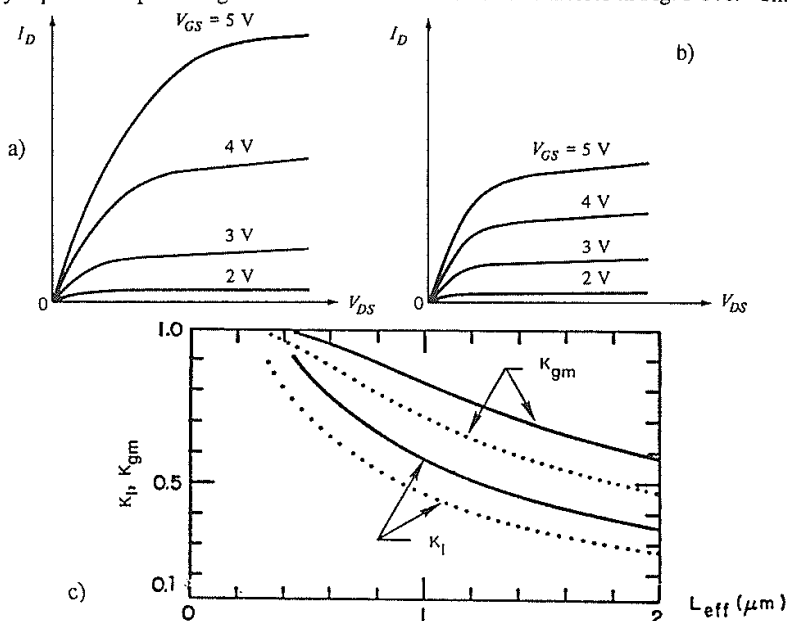


Fig. 5-37 Device characteristics (a) in the absence and (b) in the presence of velocity saturation effects. From Y.P Tsividis, *Operation and Modeling of the MOS Transistor*, McGraw-Hill, New York, 1987, p. 180. Reprinted with permission. (c) The calculated drain current ideality factor K_I versus L for different gate-oxide thicknesses. Solid lines: $t_{ox} = 250$ Å; dotted lines $t_{ox} = 100$ Å.⁴⁶ (© 1988 IEEE).

K of the thicker oxide device saturates at longer channel lengths. As a result, decreasing the channel length of a thick oxide MOSFET cannot significantly improve the current drive. On the other hand, as L becomes very short (e.g., of the order of $0.25 \mu\text{m}$), K approaches unity independent of t_{ox} . Thus, I_{Dsat} becomes proportional to C_{ox} . Since, in MOSFETs with very short channels, I_{Dsat} becomes independent of channel length, the chief means of increasing I_{Dsat} will be to make t_{ox} as thin as possible (i.e., the device driving current will increase linearly with C_{ox}). Data and numerical simulation results support this argument.

As a last comment, note that because the value of v_{sat} for electrons and holes is comparable, n -channel and p -channel MOSFETs tend to perform similarly under velocity saturation, other things being equal. This is not the case under normal operation. That is, because I_{D} is proportional to μ_0 in normal operation, and in this case the value of μ_0 for holes in p -channel devices is one-half to one-fourth the corresponding value of μ_0 for n -channel devices.

5.4.2.2 Effects of Scaling on g_{msat} . From Eq. 5-39 the maximum available value of g_{msat} for a given t_{ox} is given by

$$g_{\text{msat}} = Z v_{\text{sat}} C_{\text{ox}} \quad (5-42)$$

This is the ultimate relationship for very-short-channel MOSFETs. However, the g_{msat} of a device with practical geometries with finite channel length is always lower than this maximum. In general,

$$g_{\text{msat}} = K_{\text{gm}} Z v_{\text{sat}} C_{\text{ox}}, \quad \text{where } K_{\text{gm}} < 1 \quad (5-43)$$

Compared to Eq. 5-39, K_{gm} can be written as

$$K_{\text{gm}} = [1 - (\partial V_{\text{DSSat}} / \partial V_{\text{GS}})] \quad (5-44)$$

It is observed that g_{msat} of a short-channel device often peaks at some moderate gate bias, beyond which it appears to either remain flat or sometimes decrease slightly, as shown in Fig. 5-38. This is a consequence of the term $\partial V_{\text{DSSat}} / \partial V_{\text{GS}}$ in Eq. 5-44, which becomes almost constant beyond a gate bias roughly equal to $\mathcal{E}_{\text{sat}} L$.

5.4.2.3 Comparing the Values of I_{D} Calculated with the Long- and Short-Channel MOSFET Models as a Function of Channel Length. The values of I_{Dsat} as the channel length is reduced from $5 \mu\text{m}$ to $0.3 \mu\text{m}$ as calculated both by the charge-control dc circuit model of the long-channel MOSFET (Eq. 4-68), and the short channel dc circuit model (Eq. 5-38, whose basis is also the charge-control model) are given in Table 5-3. We observe that Eq. 4-68 predicts larger values of I_{Dsat} than those calculated with Eq. 5-38 (and these in turn are much closer to the measured values). This reiterates the importance of choosing the correct device model when performing technology scaling studies.

Table 5-3 Comparison of Drive Current Values versus Gate Length with the Long-Channel (Eq.4-68) and Short-Channel MOSFET Models.

Technology/Gate Length (μm)	5	3	2	1	0.5	0.3
Gate Oxide (nm)	100	60	40	20	12.5	7.5
Calculated $I_{D\text{sat}}$ @ $V_{GS} = 5\text{V}$	1	2.61	6.25	20	99.3	206
Long-channel Model ($\mu\text{A}/\mu\text{m}$)						
Calculated $I_{D\text{sat}}$ @ $V_{GS} = 5\text{V}$	1	2.18	4.12	10.5	19.9	30
Short-Channel Model ($\mu\text{A}/\mu\text{m}$)						

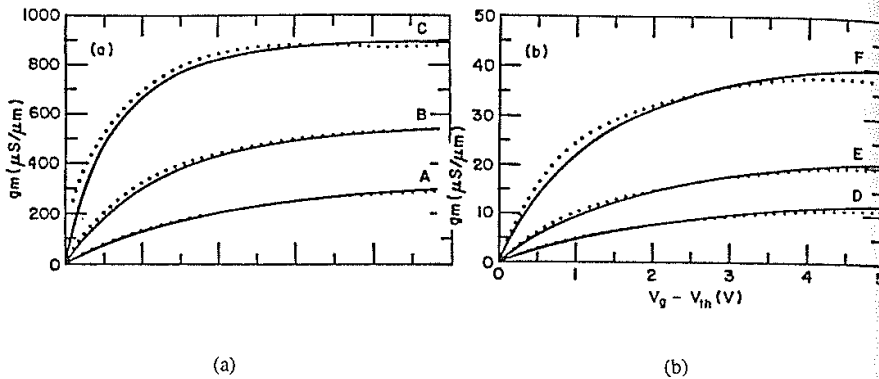


Fig. 5-38 Measured and calculated g_m vs $(V_{GS} - V_T)$ @ $V_{DS} = 5\text{V}$ for (a) NMOS; & (b) PMOS devices with $L = 3\ \mu\text{m}$ (A & D); $L = 2\ \mu\text{m}$ (B & E); and $L = 1\ \mu\text{m}$ (C & F).⁴⁶ (© 1988 IEEE).

5.4.3 Modeling the Saturation Region of Operation in Short-Channel MOSFETs

For drain voltages in excess of the pinch-off voltage ($V_{D\text{ssat}}$) the gate-to-drain voltage drops below the voltage value needed to establish an inversion layer (i.e., it becomes smaller than the threshold voltage). Hence, long-channel MOSFET theory implies that the inversion layer no longer exists in this case, and that there are no longer any mobile carriers present in the channel at such locations. The portion of the channel extending back from the drain over which the inversion layer is extinguished is called the *pinched-off region*. The length of the pinched off region is ΔL , and its length increases as V_{DS} grows beyond $V_{D\text{ssat}}$ (see Fig. 5-39). In addition, in saturation most of the drain voltage in excess of $V_{D\text{ssat}}$ is dropped across this region. From a modeling perspective this makes it appear as if the normal channel length (in which the GCA is still applicable, and in which mobile charges still exist) is growing shorter. The apparent shrinkage in channel length from this effect is known as *channel-length modulation*.

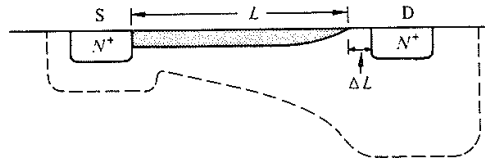


Fig. 5-39 Visualization of the velocity saturation region in a MOSFET, having a length ΔL .

In long-channel MOSFETs the length of the pinched-off region is much smaller than the original channel length (i.e. $\Delta L \ll L$), even at large values of V_{DS} . Thus, I_D remains approximately constant for drain voltages in excess of V_{DSsat} . In short-channel MOSFETs, however, ΔL can be comparable to L . Thus, the same voltage drop (V_{DSsat}) will appear across a substantially shorter channel ($L - \Delta L$). Hence, when $V_{DS} > V_{DSsat}$ in short-channel MOSFETs, I_D can increase significantly with increasing V_{DS} (as is seen in Fig. 5-31b). Since I_{Dsat} is proportional to $1/L$, the equation of drain current I_D in saturation has the general form of

$$I_D (V_D > V_{DSsat}) = I_{Dsat} [L / (L - \Delta L)] \quad (5 - 45)$$

To be able to predict the increase in I_D in short-channel MOSFETs, the value of ΔL as a function of the terminal voltages is needed. The most general approach to calculate ΔL , however, is to simultaneously solve the current transport equation (Eq. 1-4), and the two-dimensional form of the Poisson equation (Eq. 1-3) in the pinch-off region. Unfortunately, this general approach requires numerical solution of these equations, using such 2-D device simulators as MINIMOS or PISCES (see chaps. 1 and 2). To obtain an analytical expression for ΔL some approximations must be made to reduce the 2-D Poisson equation to a 1-D form. Analytical solutions of the Poisson equation can then be found, and these solutions can be used to derive analytical expressions for ΔL . The analysis used to derive models for ΔL will also yield expressions that calculate the value of the channel electric field. These can be employed to find the magnitude of the maximum channel $\mathcal{E}$ -field strength $\mathcal{E}_{y\max}$, a quantity that is needed when the subject of hot-carrier effects in MOSFETs is considered (see chap. 9).

For the purposes of modeling the MOSFET in saturation, the most common approach for reducing the 2-D Poisson equation to a 1-D form *assumes that* $\partial \mathcal{E} / \partial x$ (i.e., the gradient of the electric field in the vertical (x) direction) is much smaller than $\partial \mathcal{E} / \partial y$. In such cases, the $\partial \mathcal{E} / \partial x$ term in the 2-D Poisson equation can be dropped, turning a 2-D partial differential equation into the form of a one-dimensional (ordinary) differential equation. We note that this approach is opposite to the one taken in the GCA approximation. Namely, in the GCA approach the $\partial \mathcal{E} / \partial y$ term is considered to be much smaller than the $\partial \mathcal{E} / \partial x$ term, and thus it is the term dropped from the Poisson equation. Here we will derive three analytical models for ΔL . The simplification of dropping $\partial \mathcal{E} / \partial x$ entirely from the 2-D Poisson equation is used in the first two.

5.4.3.1 Modeling MOSFET Saturation Behavior Using the Abrupt Junction and Depletion Approximations.

The first and simplest approach to modeling saturation behavior in MOSFETs divides the MOSFET channel into two regions: one in which the long-channel I_D models are still assumed to be valid, and the pinched-off region. This simple model also invokes three approximations in the pinched-off region, which are:

- Only the drain field is assumed to be responsible for determining the current flow in the pinched-off region. (That is, the vertical field due to the gate is ignored.) In addition, the drain field is assumed to have only a horizontal component, $E_y(y)$. Furthermore, its magnitude is assumed to be constant as a function of x . The map of the drain field in the pinched-off region is thus qualitatively depicted in Fig. 5-40. The strength of the field is indicated by the length of the arrows, and the field direction at each point by the direction of the arrow.
- No mobile charge is present in the pinched-off region. That is, only the ionic charge [uniformly distributed with a concentration $\rho(x,y) = -qN_A$] is assumed to be present (i.e., the depletion approximation described in section 3.2.3 is once again invoked here).
- The abrupt junction approximation is also invoked. This implies that $E_y = 0$ at the leading edge of the pinched-off region $y = (L - \Delta L)_+$ (because the electric field at the edge of the depletion region in any abrupt junction is assumed to be zero).

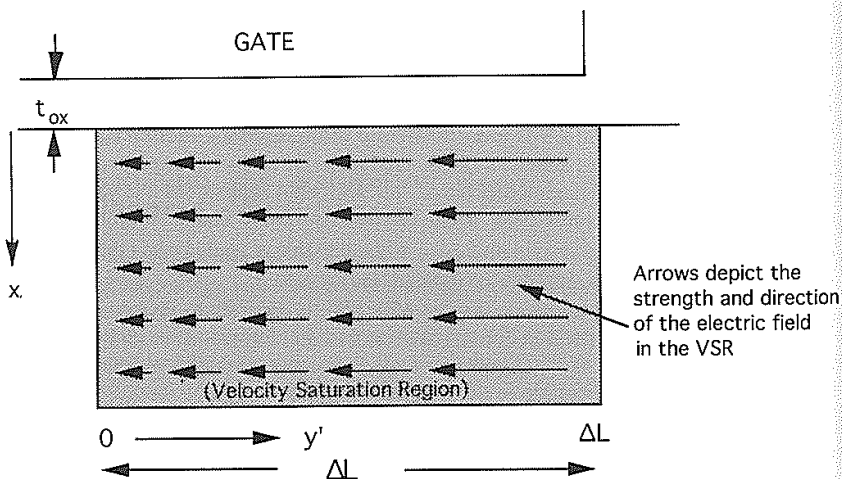


Fig. 5-40 Map of the drain field in the VSR region.

Using these approximations the Poisson equation is reduced to the form:

$$\frac{d\mathcal{E}_y(y)}{dy} = \frac{qN_A}{\kappa_{si}\epsilon_0} \quad (5-46)$$

The expression for calculating the channel length modulation is found from Eq. 5-46 to be

$$\Delta L = \sqrt{\frac{2\epsilon_{si}}{qN_A} (V_{DS} - V_{DSsat})} \quad (5-47)$$

This equation for calculating ΔL was introduced by Reddi and Sah in 1965.⁴⁸

Using this method to estimate ΔL , however, leads to some physical inconsistencies in the value of $\mathcal{E}_y(y)$ as well as yielding values of ΔL that substantially overestimate its length (i.e., it predicts that ΔL can be larger than L , even in 2 μm MOSFETs under normal operating conditions). A more complex model must therefore be invoked to reconcile these apparently unphysical results, as well as to improve the predictive capability of the ΔL model. Before we describe such more-complex models let us first discuss the shortcomings of the Reddi-Sah ΔL model in more detail.

The variation of the channel electric field when the device is saturated provides some insight into the shortcoming of this ΔL model. We calculate the channel field using the approximations that led to the charge-control long-channel dc MOSFET model (adopted here again for the sake of mathematical simplicity). The channel field $\mathcal{E}_y(y)$ for $y < L - \Delta L$ can be found by substituting Eq. 4-63 into Eq. 4-22. Combining with the pinchoff conditions that led to Eq. 5-47, we get

$$\mathcal{E}_y(y) = \frac{V_{GS} - V_T}{2(L - \Delta L)} \frac{1}{1 - [y/(L - \Delta L)]} \quad \text{for } y < L - \Delta L \quad (5-48)$$

$$= \frac{qN_A}{\epsilon_{si}} [y - (L - \Delta L)] \quad \text{for } y > L - \Delta L \quad (5-49)$$

These two equations for $\mathcal{E}_y$ are plotted as a function of y in Fig. 5-41. Note from Eq. 5-49 that $\mathcal{E}_y$ increases linearly with y in the pinched-off region. In addition, from Eq. 5-48 we see that the model predicts $\mathcal{E}_y$ goes to infinity at $y = L - \Delta L$. This occurs because at the point where the pinched-off begins (i.e., $y = L - \Delta L$), the inversion-layer charge concentration Q_i becomes zero, as required by the pinchoff (and depletion-approximation) condition. To maintain current continuity when the mobile-charge concentration approaches zero, the electric field must approach infinity. In addition, Fig. 5-41 shows that $\mathcal{E}_y$ is discontinuous at $y = L - \Delta L$ (i.e., $\mathcal{E}_y$ is infinite at $[L - \Delta L]_-$, and is zero at $[L - \Delta L]_+$). Both of these predictions are, of course, unphysical. The problem lies with the pinchoff condition. That is, the concentration of mobile charges in the pinched-off region never actually becomes zero because of carrier velocity saturation. As a result, to derive a more accurate and physically correct model of the

MOSFET in saturation, a somewhat more complex formulation of the Poisson equation than Eq. 5-46 must be employed.

It should be noted that a completely empirical approach based on the Reddi-Sah model of ΔL was developed by Frohman-Bentchkowsky and Grove.⁴⁹ In this approach, Eq. 5-47 is still used, but two terms - each containing an empirically determined constant - are added as corrections to this equation to improve the accuracy with which it predicts the value of ΔL . One of the correction terms is attributed to the fringing field between the gate and sidewall of the drain and the other to the fringing field between the gate and the end of the channel. These effects are not considered in the basic model represented by Eq. 5-47. Agreement with experiment is good for channel lengths from 2 to 4 μm and $N_{\text{SUB}} \sim 10^{15} - 10^{16} \text{ cm}^{-3}$.

5.4.3.2 Constant-Field-Gradient Model. The prediction of an infinite electric field strength at $y = L - \Delta L$ arises from the assumption that no mobile charge exists in the depletion region. If we remove this assumption (approximations **b** of the Reddi-Sah model), and also approximation **c** (which leads to the nonphysical discontinuity in the $\mathcal{E}$ -field value at the leading edge of the pinched-off region), but retain approximation **a**, the formulation of the second of the three saturation models is obtained. That is, we now assume that a uniform mobile charge concentration (with a density N_m) exists to a depth r_j in the pinched-off region, and also that these carriers are velocity saturated. Hence, it is appropriate to rename the pinched-off region as the *velocity saturation region* (VSR). Now, the mobile charge density in the VSR can be expressed in the following form⁵⁰ *

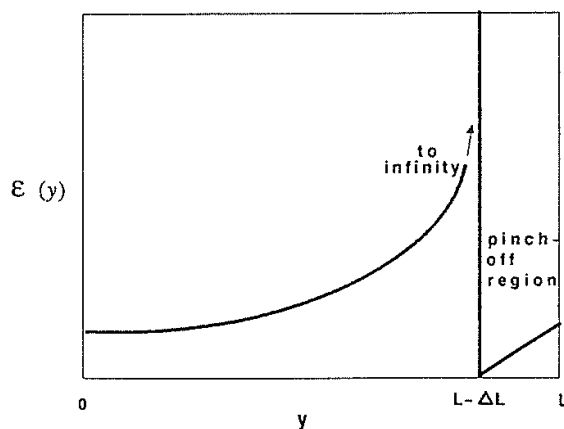


Fig. 5-41 General behavior of $\mathcal{E}(y)$, the channel field, for a device in saturation according to the long-channel model.⁴⁵ Reprinted with permission of Academic Press.

* Note that Eq. 5-50 is similar to the more familiar drift current expression $I_{\text{drift}} = J_{\text{drift}} A = q n \langle v \rangle A$, where N_m corresponds to n , v_{sat} to $\langle v \rangle$, and $r_j Z$ to A .

$$q N_m = (A_{ph} I_D) / v_{sat} r_j Z \quad (5 - 50)$$

where I_D in this region is assumed to consist of velocity saturated carriers distributed uniformly into the bulk to a depth equal to the source/drain junction depths r_j , and A_{ph} is a fitting parameter. With the total charge in the VSR now expressed as the sum of the ionic and mobile charges, the one-dimensional form of the Poisson equation (Eq. 5-46) becomes

$$\frac{d\mathcal{E}_y(y)}{dy} = \frac{q N_A}{\kappa_{si} \epsilon_0} + A_{ph} \frac{I_D}{\kappa_{si} \epsilon_0 r_j v_{sat} Z} \quad (5 - 51)$$

Using this equation, we can solve for the value of $\mathcal{E}_y(y)$ in the VSR. When removing approximation c we instead apply the boundary condition that $\mathcal{E}_y(0) = \mathcal{E}_{sat}$ at the leading edge of the VSR (instead of the boundary condition that $\mathcal{E}_y = 0$ at $[L - \Delta L]_+$, as is done in the Reddi-Sah model).

Since the terms on the right hand side of Eq. 5-51 are constant as a function of y , this model indicates that the gradient of the channel electric field in the y -direction in the VSR, $d\mathcal{E}_y(y)/dy$, is also a constant (hence the name, the *constant-field-gradient model*). As a result, $\mathcal{E}_y(y)$ is predicted by this model to increase linearly with y in the VSR (as is also predicted by Reddi-Sah model). Since more charge is now assumed to exist in the VSR, however, the magnitude of the linear gradient in this model is larger. We plot this result over the channel position ΔL as the dotted line in Fig. 5-42.⁵⁴

The channel field $\mathcal{E}_y(y)$ as calculated from a numerical solution of the two-dimensional form of the Poisson equation is also plotted in Fig. 5-42.⁵⁵ The numerical simulation predicts that $\mathcal{E}_y(y)$ exhibits a near-exponential-dependent rise towards the drain, implying that the constant-field-gradient model significantly underestimates the peak channel electric field. Because hot-carrier effects are extremely sensitive to the magnitude of the peak channel field, it has been concluded that this model is not sufficiently accurate for use in analyzing hot-carrier problems. Consequently, a more accurate analytical model of the saturation region is still needed.

5.4.3.3 Pseudo-Two-Dimensional Model of the MOSFET Operating in Saturation.

The chief reason the constant-field-gradient model underestimates the maximum value of $\mathcal{E}_y$ in the VSR is that it ignores the effect of the gate field in the analysis. A pseudo-two-dimensional approach, which includes the effect of this field, has therefore been developed to obtain a more accurate analytical model of ΔL and $\mathcal{E}_y$ in the VSR. From these relationships a model for I_{Dsat} as a function of V_{DS} can also be obtained. The original pseudo-2-D model was introduced by Elmansy and Boothroyd in 1977.⁵¹ Later, Ko et al. developed an enhanced version^{52,53} that has also been applied to modeling hot-carrier effects and subthreshold MOSFET behavior (as outlined in section 5.2.1.4). Here we present a simplified version of this model, adapted from the description by Ko in references 52 and 53.

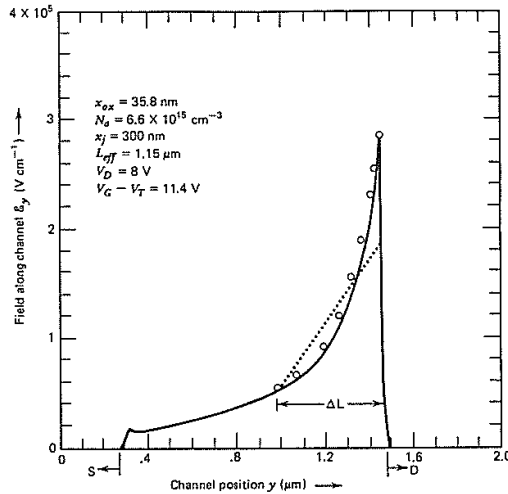


Fig. 5-42 Calculated channel field region versus position in the channel near the drain end for three models ($t_{OX} = 358 \text{ \AA}$, $N_A = 6.6 \times 10^{15} \text{ cm}^{-3}$, $r_j = 0.3 \text{ } \mu\text{m}$, $L_{eff} = 1.15 \text{ } \mu\text{m}$, $V_{DS} = 8 \text{ V}$, $(V_{GS} - V_T) = 11.4 \text{ V}$. Electrons in the channel have saturated velocities over the region ΔL which extends from the channel pinch-off point to the drain. From R.S. Muller and T.I. Kamins, *Device Electronics for Integrated Circuits*, 2nd Ed. Copyright John Wiley & Sons, New York, 1986. p. 491, Reprinted with permission.

In this approach we again remove approximations **b** and **c**, but also modify approximation **a** of the Reddi-Sah model. In Fig. 5-43a a schematic diagram of the VSR is shown. As in the previous approach the mobile carriers in this region are assumed to be velocity saturated and are uniformly distributed in the VSR to the depth of the source and drain junctions r_j (with a concentration N_m). The VSR is thus approximated as a box bounded on the sides by ABCD. This box encloses all of the VSR mobile charges and most of the ionic charges of interest. The electric field and voltage at the leading edge of the VSR (point A) are assumed to have values E_{sat} and V_{Dssat} , respectively. Since in steady-state the total number of mobile charges in the VSR is constant with time the phenomenon being analyzed is essentially an electrostatic problem, solvable using the Poisson equation.

Approximation **a** is modified as follows: First, let the coordinate system be selected such that $y = 0$ at A and $y = \Delta L$ at D. Next, we see that the charge enclosed in the VSR is the sum of the mobile charge and ionic charge concentrations, qN_A and qN_m , respectively, or:

$$\rho(x, y) = qN_A + qN_m \quad (5 - 53)$$

The drain E -field is still assumed to have only a horizontal component E_y that varies with position y , but not with position x . The map of the drain field component in the

VSR bounded by ABCD is shown in Fig. 5-43b. The effect of the field due to the gate charge, however, is now also included in the Poisson equation. This field is assumed to have only a vertical component E_x which varies with position y (i.e., E_x gets weaker as the drain is approached). E_x has a value at the Si-SiO₂ interface given by $E_x(0,y)$ and is assumed to be zero at the bottom edge of the depletion region [$E_x(d_{\max},y) = 0$]. The value of $E_x(r_j,y)$ is also assumed to be close to zero, even though r_j is not as deep as $d_{\max}$. (This latter assumption is plausible since the MOSFET is biased into strong inversion, making qN_m in the VSR larger than qN_A . Thus, most of the total charge in the VSR region of the channel [i.e., from $y = 0$ to $y = \Delta L$], is confined within the depth from $x = 0$ to $x = r_j$). It is now assumed that the value of $\partial E_x/\partial x$ at each point y can be replaced by the average value of $\partial E_x/\partial x$ from $x = 0$ to $x = r_j$. This average value is given by $[E_x(0,y) - E_x(r_j,y)/r_j]$. Since $E_x(r_j,y)/r_j \sim 0$, the average value of $\partial E/\partial x$ can be approximated as just $E_x(0,y)/r_j$.

The above modifications to approximation a result in the following important conclusion: *The term $\partial E/\partial x$ in the Poisson equation can be approximated as $E_x(0,y)/r_j$.* With this approximation, the 2-D Poisson equation which governs the relation between the electric-field and the charge density in the VSR, can be expressed as follows (with $E_x(0,y)$ being written as $E_{si}(0,y)$):

$$\frac{dE_y(y)}{dy} + \frac{E_{si}(0,y)}{r_j} = \frac{q N_A}{\kappa_{si} \epsilon_0} + \frac{q N_m}{\kappa_{si} \epsilon_0} \quad (5-54)$$

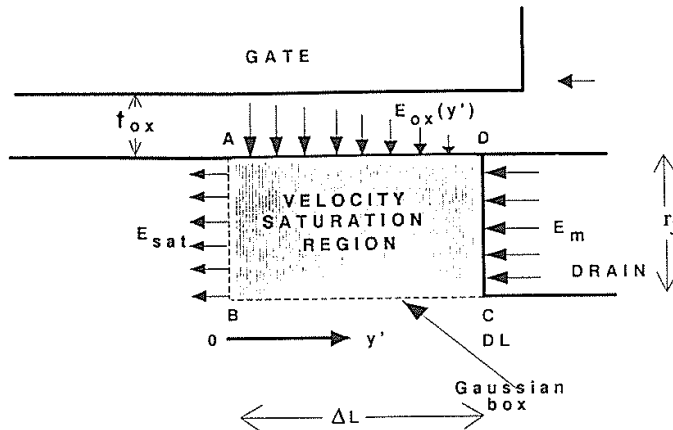


Fig. 5-43 Schematic diagram to illustrate the analysis of the velocity saturation region (VSR). From P.K. Ko, "Approaches to Scaling," Chap. 1, in *Advanced MOS Device Physics*, N.G. Einspruch and G. Gildenblatt, Eds., Vol. 18 VLSI Electronics. Academic Press 1989. Reprinted with permission.

Thus, by using this modified approximation, not only is the Poisson equation reduced to a 1-D form (i.e., dependent only on y), but in addition, the effects of the both the vertical (gate) and the horizontal (drain) fields in the VSR are included. Consequently, the name chosen for this model (i.e., the *pseudo-2-D model*) is appropriate.

Now, we can further simplify Eq. 5-54 by noting that $\mathcal{E}_{si}(0,y)$ can be described in terms of $\mathcal{E}_{ox}$ (according to Eq. 3-15) as

$$\mathcal{E}_{si}(0,y) = (\epsilon_{ox}/\kappa_{si}\epsilon_o) \mathcal{E}_{ox}(0,y) \quad (5-55)$$

Substituting the right hand side of Eq. 5-55 into Eq. 5-54 for $\mathcal{E}_{si}$, and then recalling that we since we can express $\mathcal{E}_{ox}(0,y)$ as

$$\mathcal{E}_{ox}(0,y) = [V_{GS} - V_{FB} - 2\phi_B - V(y)]/t_{ox} \quad (5-56)$$

[see Eq. 3-14, with $\phi_{ox} = V_{GS} - V_{FB} - 2\phi_B - V(y)$], we can make these both of changes in Eq. 5-54 to get

$$\frac{d\mathcal{E}_y(y)}{dy} + \frac{\epsilon_{ox}}{\kappa_{si}\epsilon_o} \frac{1}{t_{ox}} \frac{[V_{GS} - V_{FB} - 2\phi_B - V(y)]}{r_j} = \frac{q N_A}{\kappa_{si}\epsilon_o} + \frac{q N_m}{\kappa_{si}\epsilon_o} \quad (5-57)$$

Since $V(y=0) = V_{DSSat}$ we can use Eq. 5-56 to write

$$\mathcal{E}_{ox}(y=0) = \frac{1}{t_{ox}} [V_{GS} - V_{FB} - 2\phi_B - V_{DSSat}] \quad (5-58)$$

In addition, since the GCA approximation is still valid at the leading edge of the VSR ($y=0$), the total charge in the silicon (Q_C) at that point can still be calculated by Eq. 4-1c ($Q_C = \phi_{ox} C_{ox}$), where $C_{ox} = \epsilon_{ox}/t_{ox}$ and $\mathcal{E}_{ox} = \phi_{ox}/t_{ox}$. Thus, since Q_C (cm^{-2}) is $(qN_A r_j + qN_m r_j)$, and $Q_C = \mathcal{E}_{ox} \epsilon_{ox}$, we can also write

$$\mathcal{E}_{ox}(y=0) = (q N_A r_j + q N_m r_j)/\epsilon_{ox} \quad (5-59)$$

Hence,

$$\frac{\epsilon_{ox}}{\kappa_{si}\epsilon_o} \frac{1}{t_{ox}} [V_{GS} - V_{FB} - 2\phi_B - V_{DSSat}] = \frac{q N_A}{\kappa_{si}\epsilon_o} r_j + \frac{q N_m}{\kappa_{si}\epsilon_o} r_j \quad (5-60)$$

Now we can substitute $(\epsilon_{ox}/\kappa_{si}\epsilon_o t_{ox} r_j) [V_{GS} - V_{FB} - 2\phi_B - V_{DSSat}]$ for the right hand side of Eq. 5-57, allowing Eq. 5-57 to be simplified to

$$\frac{d\mathcal{E}_y(y)}{dy} = \frac{\epsilon_{ox}}{\kappa_{si}\epsilon_o t_{ox} r_j} [V(y) - V_{DSSat}] \quad (5-61)$$

or

$$\frac{d\mathcal{E}_y(y)}{dy} = \frac{[V(y) - V_{DSSat}]}{l^2} \quad \text{where } l^2 = \frac{\kappa_{si}\epsilon_o}{\epsilon_{ox}} t_{ox} r_j \quad (5-62)$$

or

$$\frac{d^2[V(y) - V_{DSSat}]}{dy^2} = \frac{[V(y) - V_{DSSat}]}{l^2} \quad (5 - 63)$$

where the parameter l is the effective length of the VSR.

A qualitative physical interpretation of equations 5-62 and 5-63 can be given as follows: At $y = 0$, where $V(0) = V_{DSSat}$, the charge in the VSR is still controlled entirely by the vertical field ($\mathcal{E}_x$). Since, for conditions of dc current flow the total charge density (i.e., the sum of the mobile and ionic charge densities) must remain constant along the VSR toward the drain, the total electric field at each point y must also remain constant. However, the voltage $V(y)$ increases toward the drain. Thus, from Eqs. 5-55 and 5-56, we see that $\mathcal{E}_x$ must decrease toward the drain, and so must the amount of charge in the VSR controlled by it. According to the Poisson equation, this charge must be taken up by the gradient of $\mathcal{E}_y$. Thus, the gradient of $\mathcal{E}_y$ must increase toward the drain. In Eq. 5-62, the RHS is the amount of charge released by $\mathcal{E}_x$ as a result of the rise in the channel voltage [equal to $V(y) - V_{DSSat}$], while the LHS is the corresponding increase of the gradient in $\mathcal{E}_y$ in order to support these charges.

The general solution of Eq. 5-63 is:

$$[V(y) - V_{DSSat}] = A e^{y/l} + B e^{-y/l} \quad (5 - 64)$$

By applying the boundary conditions: $V(0) = V_{DSSat}$, and $\mathcal{E}_y(0) = [\partial V(y)/\partial y]_{y=0} = \mathcal{E}_{ysat}$, the coefficients in Eq. 5-64 are found to be $A = l\mathcal{E}_{sat}/2$ and $B = -l\mathcal{E}_{sat}/2$. Thus, we can write the expressions for $V(y)$ and $\mathcal{E}_y(y)$ in the VSR as

$$V(y) = V_{DSSat} + l\mathcal{E}_{sat} \sinh(y/l) \quad (5 - 65a)$$

and

$$\mathcal{E}_y(y) = dV/dy = \mathcal{E}_{sat} \cosh(y/l) \quad (5 - 66a)$$

Note that near the drain end of the VSR

$$V(y) \approx V_{DSSat} + l\mathcal{E}_{sat} \exp(y/l) \quad (5 - 65b)$$

and

$$\mathcal{E}_y(y) \approx [(l\mathcal{E}_{sat})/l] \exp(y/l) \quad (5 - 66b)$$

At the drain end of the channel $V(y)$ and $\mathcal{E}_y$ have their maximum values V_{DSmax} and $\mathcal{E}_{ymax}$, as given by

$$V_{DSmax} = V_{DSSat} + l\mathcal{E}_{sat} \sinh(\Delta L/l) \quad (5 - 67)$$

and

$$\mathcal{E}_{ymax} = \mathcal{E}(y = \Delta L) = \mathcal{E}_{sat} \cosh(\Delta L/l) \quad (5 - 68)$$

Equations 5-67 and 5-68 can be combined to yield an expression for ΔL

$$\Delta L = l \ln \frac{\{[V(y) - V_{DSSat}]/l\} + \mathcal{E}_{y\max}}{\mathcal{E}_{sat}} \quad (5 - 69)$$

and

$$\mathcal{E}_{y\max} = \left[\frac{[V_{DS} - V_{DSSat}]^2}{l^2} + \mathcal{E}_{sat}^2 \right]^{1/2} \quad (5 - 70a)$$

Equations 5-66 to 5-70 describe important aspects of the behavior of the MOSFET in saturation. First, Eq. 5-66b indicates that $\mathcal{E}_y$ grows almost exponentially towards the drain, as was predicted by 2-D numerical simulation. In Fig. 5-42, Eq. 5-66a is plotted along with the results of the 2-D numerical simulation. Close agreement is observed. However, the parameter l must be adjusted to give approximately the same peak channel field. We indicate below how this has been done using both experimental data and numerical simulation.

From Eq. 5-70, $\mathcal{E}_{y\max}$ is seen to depend explicitly on $[V(y) - V_{DSSat}]$, but the channel length L only influences $\mathcal{E}_{y\max}$ indirectly. Furthermore, if V_{DS} is more than about 2V greater than V_{DSSat} , $(V_{DS} - V_{DSSat})/l \gg \mathcal{E}_{sat}$. In this case Eq. 5-70a for $\mathcal{E}_{y\max}$ is simplified to

$$\mathcal{E}_{y\max} \approx [V_{DS} - V_{DSSat}]/l \quad (5 - 70b)$$

From experimental data and 2-D device simulations using MINIMOS the following expressions for l that appear valid over a wide range of t_{ox} , N_A , and r_j have been deduced:¹

$$l \approx 0.22 t_{ox}^{1/3} r_j^{1/3} \quad \text{if } t_{ox} \geq 15 \text{ nm} \quad (5 - 70c1)$$

and

$$l \approx 1.7 \times 10^{-2} t_{ox}^{1/8} r_j^{1/3} \quad \text{if } t_{ox} < 15 \text{ nm} \quad (5 - 70c2)$$

where l , r_j and t_{ox} are in centimeters.

By substituting Eq. 5-70c1 into Eq. 5-70b a simple, explicit expression for the maximum channel electric field is obtained

$$\mathcal{E}_{y\max} \approx [V_{DS} - V_{DSSat}]/0.22 r_j^{1/3} t_{ox}^{1/3} \quad (5 - 71)$$

where V_{DSSat} is found using Eq. 5-37 and $\mathcal{E}_{sat}$ in that equation is taken to be the critical field for velocity saturation ($\sim 5 \times 10^4$ V/cm). Equations. 5-71 and 5-37 can thus be used to estimate the peak channel electric field $\mathcal{E}_{y\max}$ as a function of device dimensions and bias voltages in conventional (i.e., non-graded drain) MOSFETs.

For example, by substituting some realistic values of t_{ox} and r_j into Eq. 5-71 (i.e., using values typical of 1- μm CMOS technology, namely, $t_{ox} = 250\text{\AA}$ and $r_j = 0.2 \mu\text{m}$), we get $l = 135 \text{ nm}$, and $1/l = 7.5 \times 10^4 \text{ cm}^{-1}$. At low $(V_{GS} - V_T)$ values, $V_{DSSat} \sim 0$. Using $\mathcal{E}_{sat} = 5 \times 10^4 \text{ V/cm}$ and $V_{DS} = 5\text{V}$, from Eq. 5-71 we get $\mathcal{E}_{y\max} = 3.85 \times 10^5 \text{ V/cm}$. Since $\mathcal{E}_{y\max}$ is proportional to $(1/r_j^{1/3} t_{ox}^{1/3})$, it is apparent that device scaling will magnify the high-field problems unless the power-supply voltage is also correspondingly reduced, or the MOSFET device structure is modified to reduce $\mathcal{E}_{y\max}$.

With an expression for ΔL also in hand (Eq. 5-69), a relationship for the drain current in saturation based on the pseudo-2D saturation model can now be derived, namely for $V_{DS} > V_{DSSat}$:

$$I_D(\Delta L) = I_{Dsat} \left[\frac{L}{L - \Delta L} \right] \left[\frac{V_{DSSat} + E_{sat} L}{V_{DSSat} + E_{sat} (L - \Delta L)} \right] \quad (5 - 72)$$

5.5. SHORT-CHANNEL MOSFET MODELS USED IN THE SPICE CIRCUIT SIMULATOR

The simple, analytical short-channel MOSFET model described in section 5.4 is adequate for hand calculations. However, for computer-aided circuit analysis and design a more accurate model is needed, which also retains the benefit of simplicity to allow rapid circuit-simulation execution times. The Level-1 and Level-2 MOSFET models in SPICE are not suitable since they do not work well for small geometry transistors. Likewise, the Level-3 model relies too heavily on a semi-empirical approach. Thus, a simple and accurate, physically based, short-channel MOSFET model suitable for circuit simulation programs was needed.

Such a model has been developed and is incorporated in the circuit simulator SPICE. This model, named BSIM (Berkeley Short-channel IgFET Model), was introduced in 1987.⁵⁶ It is a modified version of the Bell Labs model CSIM with substantial enhancements (Compact Short-Channel IgFET Model).⁵⁷ BSIM is the Level-4 MOSFET model in SPICE, and BSIM2 (introduced in 1990) is the Level-5 MOSFET model.⁵⁸ By 1993, BSIM3 was developed and incorporated into SPICE.

BSIM is capable of simulating the drain current in short channel MOSFETs in both the strong-inversion and weak inversion regimes. It can model such short-channel effects as channel-length shortening, channel narrowing, mobility degradation, drain-induced barrier lowering, nonuniform channel doping for ion-implanted devices, subthreshold slope, and geometric dependencies. Terms with strong physical meaning are employed to model the fundamental physical effects while empirically derived parameters are judiciously used to embrace less-well-understood (and generally subtle) device characteristics. This approach is well-suited for circuit-analysis purposes, especially as two- and three dimensional small geometry effects become more important.

BSIM2 is an extension of the original BSIM model, but BSIM3 is essentially a new model. That is, while BSIM has about 60 parameters (17 for modeling the dc characteristics) and BSIM2 about 120, BSIM3 has 38 (the Level-1 and Level-2 SPICE models described in chap. 4 each have a total of about 15). It is generally conceded that while the Level-1 and Level-2 models had too few parameters, BSIM and BSIM2 had too many. BSIM3 is a compromise that seeks to use just the right number. In addition, many BSIM and BSIM2 parameters were put into the model strictly for curve fitting and did not have much physical basis. This makes it difficult to generate fast and slow models using BSIM or BSIM2. For the most part, the BSIM3 parameters are based on reasonable physical assumptions, and almost all have some connection with the real world, making the results of BSIM3 easier to interpret.

BSIM also includes a dedicated parameter extraction program, which together with the parameter-extraction system hardware (consisting of three major elements: a desktop computer, an H-P 4145 semiconductor parameter analyzer, and a probe station) allows the BSIM model parameters for a family of devices to be automatically extracted from measurements of actual test devices.

5.5.1 BSIM Model Equations

The formulation of BSIM can be broken down into the following aspects: (1) an equation for the threshold voltage V_T of the short-channel MOSFET; (2) equations for the operation of the short-channel MOSFET in strong inversion; and (3) equations for the operation in weak inversion (subthreshold regime).

5.5.1.1 BSIM Threshold Voltage Equation. The equation used to model V_T in long-channel MOSFETs (i.e., Eq. 4-12, with V_{BC} replaced with V_{BS}), is modified in BSIM to allow the threshold lowering effect observed in short-channel MOSFETs to be modeled. The BSIM V_T equation is

$$V_T = V_{FB} + 2\phi_B + \gamma(2\phi_B + V_{BS})^{1/2} - K_T(2\phi_B - V_{BS}) - \eta V_{DS} \quad (5-73)$$

We observe that in addition to the three parameters V_{FB} , ϕ_B , and γ , Eq. 5-73 contains two more for modeling V_T , namely K_T and η . K_T is the source and drain depletion charge sharing coefficient, and η is the drain-induced barrier lowering coefficient (which is roughly inversely proportional to the channel length L). Together, K_T and η model the threshold lowering effect.

5.5.1.2 BSIM Equations for I_D when the MOSFET is Operated in Strong Inversion. The BSIM I_D equations contain three additional parameters to allow the mobility degradation and velocity saturation effects to be modeled. These are the vertical field mobility degradation coefficient U_0 , the velocity saturation coefficient U_1 , and the carrier mobility μ_0 . (Note that μ_0 in BSIM is a function of the substrate and drain biases.) In order to speed up the circuit simulation time, instead of using I_D in the form of the ionic charge model (i.e., of the form of Eq. 4-57 - which has a 3/2 power dependence of the drain current on the drain and substrate bias), a numerical approximation to the terms of Eq. 4-57 that contain V_{BS} is used. The resulting BSIM equations for I_D in the linear and saturation regions are:

Linear Region: [$V_{GS} > V_T$], and $0 < V_{DS} < V_{DSsat}$]

$$I_D = \frac{\mu_0}{[1 + U_0(V_{GS} - V_T)]} \frac{C_{ox} \frac{Z}{L}}{(1 + \frac{U_1 V_{DS}}{L})} (V_{GS} - V_T) V_{DS} - \frac{a V_{DS}^2}{2} \quad (5-74)$$

where

$$a = 1 + \frac{\gamma}{2\sqrt{2\phi_B - V_{BS}}} \left[1 - \frac{1}{1.744 + 0.8364(2\phi_B - V_{BS})} \right] \quad (5-75)$$

If we compare Eq. 5-74 with the linear I_D equation of the square-law model (Eq. 4-63b), we see that Eq. 5-74 is similar but that it also has some extra terms. The first fractional term of Eq. 5-74 models the vertical field mobility degradation effect, and has the same form as Crawford's mobility expression (Eq. 5-18). The denominator of the second fractional term models the velocity saturation effects. The last term of Eq. 5-74 is like the last term of Eq. 4-63b except that it also contains the empirical parameter 'a.' This formulation allows the terms of Eq. 4-57 containing V_{BS} to be handled via a numerical approximation, rather than by having to compute terms to the 3/2 power (as would have to be done if Eq. 4-63 was used here instead to compute I_D). The values of I_D computed with this approximation are quite close to the values that would be computed if the Eq. 4-57 was used instead (over a reasonable range of V_{BS} and V_{DS}). See Appendix C to see how this approximation is derived, starting from Eq. 4-57.

Saturation Region: [$V_{GS} > V_T$], and $V_{DS} \geq V_{DSSat}$]

$$I_D = \frac{\mu_o}{[1 + U_o(V_{GS} - V_T)]} \frac{C_{ox} \frac{Z}{L} (V_{GS} - V_T)^2}{2aV} \quad (5-76)$$

where

$$V = \frac{1 + v_c + \sqrt{1 + 2v_c}}{2} \quad \text{and} \quad v_c = U_1(V_{GS} - V_T)/a \quad (5-77)$$

and

$$V_{DSSat} = (V_{GS} - V_T)/aV \quad (5-78)$$

5.5.1.3 BSIM Equations for I_D when the MOSFET is Operated in Weak Inversion (Subthreshold Regime). In BSIM the total drain current is modeled as the linear sum of the strong and weak inversion drain current components ($I_{D,S}$ and $I_{D,W}$, respectively) to allow proper matching of these two components at a transition point close to the threshold voltage. The weak inversion component is expressed as

$$I_{D,W} = (I_{exp} I_{limit}) / (I_{exp} + I_{limit}) \quad (5-79)$$

where

$$I_{exp} = \frac{\mu_o C_{ox} Z}{L} \left(\frac{kT}{q} \right)^2 e^{1.8} e^{(q/kT)(V_{GS} - V_T)/\eta} [1 - e^{-V_{DS}(q/kT)}] \quad (5-80)$$

and

$$I_{limit} = \frac{\mu_o C_{ox} Z}{L} \left(3 \frac{kT}{q} \right)^2 \quad (5-81)$$

The factor $e^{1.8}$ is empirically chosen to achieve best fits in the subthreshold characteristics with minimum effect on the strong-inversion characteristics. The

subthreshold parameters η_0 , η_B , and η_D are used to model the subthreshold slope coefficient η , as:⁵⁹

$$\eta = \eta_0 + \eta_B V_{BS} + \eta_D V_{DS} \quad (5 - 82)$$

The subthreshold swing S_t is expressed as (Eq. 4-91)

$$S_t = (2.3kT/q) \eta = (2.3kT/q) (1 + C_d/C_{ox}) \quad (5 - 83)$$

where η is the subthreshold slope coefficient calculated from Eq. 3-82. As discussed in chap. 4 and earlier in this chapter, the subthreshold swing is a function of both substrate and drain biases, and this bias dependence is modeled in BSIM by the linear relation given by Eq. 5-82.

5.6 MOSFET SCALING

A roadmap of the scaling of MOSFET devices was proposed by Hu in 1993.²⁴ He assumed that a new generation of technology will continue to be developed every three years, with perhaps a slow-down to four years beyond the 0.35 μm generation. His proposal for MOSFET scaling is outlined as follows, with the result summarized in Table 5-4.

1. The IC industry/market will agree on the next power-supply voltage standard V_{CC} several years in advance of introducing a technology/product using that voltage.
2. For a given V_{CC} , the thinnest gate oxide (i.e., from manufacturability, reliability and GIDL considerations) will be used to get maximum I_D .
3. Junction depths will be scaled aggressively to keep the short-channel effects (i.e., subthreshold leakage current) within a desired limit.
4. V_T of general purpose technology will remain basically unchanged. A significant reduction in V_T is unacceptable for channel subthreshold leakage.
5. The well doping concentration or punchthrough implant dose will be increased, and the gate length may be chosen to be larger than the minimum feature size in order to achieve acceptable leakage and standby current.
6. Drain engineered structures will be used as necessary to meet the constraints of hot-carrier reliability, breakdown voltage, and GIDL.

5.6.1 Design of Submicron MOSFETs: Scaling Guidelines

The fundamental issue of downsizing MOSFETs is to preserve long-channel characteristics after miniaturization. Several approaches have been proposed as roadmaps for designing submicron MOSFETs ($L < 1 \mu\text{m}$) so that they exhibit such behavior.

Table 5-4 MOSFET Technology Projection²⁴ (High-Speed Scenario)

Minimum Feature Size (μm)	0.5	0.35	0.25	0.18	0.13	0.09
Year of Introduction	1991	1994	1997	2001	2005	2009
SRAM Density	4M	16M	64M	256M	1G	4G
V_{CC} (V) (+10% for max V_{CC})	5	3.3	3.3	3.3	2.2	2.2
MOSFET Gate Oxide (nm)	13.5	9	8	7	4.5	4
Junction Depth (μm)	0.15	0.15	0.1	0.08	0.08	0.07
Effective Channel Length (μm)	0.40	0.30	0.25	0.2	0.15	0.13
Threshold Voltage (V)	0.7	0.7	0.7	0.7	0.7	0.7
NMOS I_{Dsat} @ $V_{GS} = V_{CC}$ (mA/ μm)	0.64	0.48	0.55	0.65	0.51	0.57
PMOS I_{Dsat} @ $V_{GS} = V_{CC}$ (mA/ μm)	0.31	0.22	0.26	0.32	0.24	0.28

Three of the most important scaling methodologies are the following: (1) constant electric-field scaling and its derivatives (constant voltage and constant electrostatic scaling); (2) scaling based on subthreshold behavior; (3) scaling to achieve a desired value of subthreshold current, I_{off} . Because subthreshold conduction is the dominant short-channel performance issue that limits MOSFET scaling,⁶⁰ the latter two have more applicability, but for completeness we also include a brief discussion of approach #1.

5.6.1.1 Constant Electric-Field, Constant Voltage & Electrostatic Scaling. The method of *constant electric-field scaling* for designing submicron MOSFETs was proposed by Dennard et al. in 1974.⁶¹ In this approach, a successful larger device structure is selected (e.g., from the previous generation of ICs) and all its dimensions and voltages are reduced by a constant scaling factor λ (>1). The new device dimensions are then $L' = L/\lambda$, $t_{ox}' = t_{ox}/\lambda$, and $W' = W/\lambda$, while the operating voltages are $V' = V/\lambda$. With this approach, operating speed increases, component density increases, and power density remains constant. Unfortunately, current density also increases by the scaling factor (a condition that aggravates electromigration failure in metal interconnect lines). In addition, the threshold voltage V_T' is reduced to V_T/λ , which makes the subthreshold current larger at $V_{GS} = 0$. The worst shortcoming, however, is that the power-supply voltage must be reduced by the same factor as the channel length. This is not compatible with circuit requirements such as noise margin, current driving capability, and power-supply voltage compatibility from one generation to the next. Thus, this method has not been adopted in its original form.

Indeed, scaling of the power-supply voltage has not been implemented in digital ICs down to device dimensions of $0.8\ \mu\text{m}$. Instead, it has been maintained at $5.0\ \text{V}$, primarily to permit the next generation of semiconductor products to "plug in" to existing systems. Only for devices with smaller dimensions has a reduced power-supply voltage been implemented ($3.3\ \text{V}$). This has led to the consideration of *constant-voltage scaling* as a miniaturization alternative. In this approach, device dimensions are scaled by the same factor λ as in constant- $\mathcal{E}$ -field scaling, but the applied voltages remain constant. While this solves the circuit incompatibility problems, it leads to higher electric fields inside the device. These high fields cause mobility degradation, hot-carrier effects, oxide tunneling, and other reliability problems.

A compromise approach named *constant-electrostatic scaling* (or *quasi-constant voltage scaling*) was thus proposed by Bacarani et al.⁶² Dimensions are still scaled by λ but the potentials are allowed to be scaled by a separate factor κ (e.g., $\kappa = \sqrt{\lambda}$). The operating voltages can then correspondingly decrease to $V' = V/\kappa$. This has allowed suitable fabrication parameters for MOSFETs to be derived for devices down to channel lengths of $0.25\ \mu\text{m}$ (some degree of power supply voltage reduction is used).

However, even with the latter method, the depletion regions of the source and drain junctions scale by a factor less than λ , and hence devices specified according to this approach are subject to such two-dimensional problems as punchthrough and DIBL. In addition, subthreshold turn-off does not scale because it is dominated by the exponential dependence on potential of the carrier density (and the subthreshold current is dominated by a diffusion rather than a drift component). Thus, the subthreshold current does not scale with field. Consequently, the scaled down device fails to turn off at $V_{\text{GS}} = 0$ (because the voltage swing needed to drive the device "off" - which is in the range of $0.5\text{--}0.8\ \text{V}$ - cannot be scaled because the subthreshold portion of this swing $\{\Delta I_{\text{DST}}/\Delta V_{\text{GS}}\}$ is invariant). As a result, the target V_{T} value of a new, reduced-design-rule process will not decrease with the other physical parameters relative to an existing, larger-design-rule process. Hence, for ICs in which subthreshold leakage currents must be minimized, even this scaling method is not adequate.

5.6.1.2 Subthreshold Scaling. Another approach to scaling which overcomes the above difficulty was proposed by Brews et al.⁶³ They introduced an empirical formula which determines the minimum gate length a MOSFET can have so that its subthreshold behavior remains insensitive to drain bias. That is, by using this formula, a constraint upon each single combination of device parameters is identified, such that if this constraint is satisfied with this combination of parameter values, the resulting device exhibits long-channel subthreshold behavior.

We recall that when long-channel devices are operated in subthreshold, I_{DST} is independent of drain-to-source voltage once V_{DS} exceeds a few kT/q (see Eqs. 4-87 and 4-89). Thus, the criterion selected to represent acceptable long-channel behavior was that no more than a 10% change in drain current could occur for a $0.5\ \text{V}$ change in V_{DS} (e.g., I_{DST} should not vary more than 10% as V_{DS} is increased from 0.5 to $1.0\ \text{V}$).

Brews et al. found empirically that such subthreshold long-channel behavior will be preserved provided the channel length L exceeds a minimum value L_{min}

$$L \geq L_{\min} = A [r_j t_{ox} (W_S + W_D)^2]^{1/3} \quad (5-84)$$

where, r_j is the junction depth, t_{ox} is the oxide thickness, W_S and W_D are the depletion widths of the source and drain (in a one-dimensional, abrupt junction for uniform doping). With r_j , W_S , and W_D in microns, and t_{ox} in angstroms, the fitting parameter is $A = 0.41 \text{ Å}^{-1/3}$. The substrate doping dependence is contained in the depletion-width equations, inasmuch as W_S and W_D (e.g., in an NMOSFET) are given by

$$W_S = \sqrt{2\kappa_{si}\epsilon_0 V_{bi}/qN_A} \quad (5-85)$$

$$W_D = \sqrt{2\kappa_{si}\epsilon_0 (V_{bi} + V_{DS})/qN_A} \quad (5-86)$$

According to Eq. 5-84, miniaturization requires shallower junctions (r_j), thinner oxides (t_{ox}), lower voltages, or heavier doping in the channel region (i.e., this will reduce $W_S + W_D$).

A set of experimental and computer-generated points were fitted to Eq. 5-84. The validity of the equation was established for gate lengths as small as $0.3 \text{ }\mu\text{m}$ and for device parameters in the ranges of $100 \text{ Å} \leq t_{ox} \leq 1000 \text{ Å}$, $10^{14} \text{ cm}^{-3} \leq N_A \leq 10^{17} \text{ cm}^{-3}$, $0.18 \text{ }\mu\text{m} \leq r_j \leq 1.5 \text{ }\mu\text{m}$, and for drain voltages up to 5V. Figure 5-44 shows a plot of Eq. 5-84 for $L_{\min}$ (solid line) compared with experimental measurements and computer simulations. For the ranges of device parameters examined, Fig. 5-44 indicates that Eq. 5-84 provides a reliable guideline for scaling MOSFET gate lengths down to $0.3 \text{ }\mu\text{m}$.

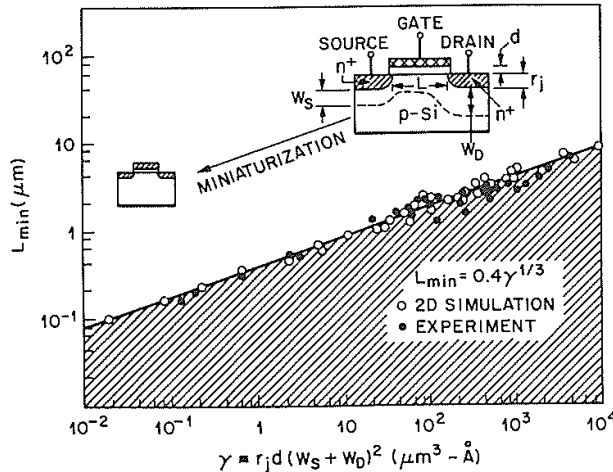


Fig. 5-44 Equation 5-84 for $L_{\min}$ compared with experimental measurements and computer simulations. The criterion for $L_{\min}$ is that for $L = L_{\min}$ a 10% variation in drain current occurs as V_{DD} varies from 0.5 to 1.0V in the subthreshold regime of gate biases. The units of the abscissa reflect the choice of units: d in Å, the other variables in μm .⁶³ (© 1980 IEEE).

This approach is also more flexible than the earlier ones because not all device parameters need to be scaled by the same factor λ . Instead, various device parameters can be adjusted independently as long as the value of $L_{\min}$ remains the same. This is shown in Fig. 5-44, which also plots $L_{\min}$ versus γ [where $\gamma = r_j t_{ox} (W_S + W_D)^2$]. All devices with channel lengths lying above the line exhibit long-channel subthreshold behavior, while those below the line (crosshatched area) exhibit short-channel subthreshold behavior.

As an example, this figure shows that in order to design a MOSFET with $L_{\min} = 0.5 \mu\text{m}$, γ should be 2. Once γ is determined, r_j , t_{ox} , W_S , and W_D can be chosen so that the resulting γ does not exceed 2. Unlike electrostatic scaling, it is not necessary to start with an optimal large device and then scale it down by a single constant factor. Rather, any device that fits the criterion will be suitable. In addition, it allows voltages to be kept fixed or only reduced modestly, provided the other factors are changed to compensate.

To some degree Eq. 5-84 appears plausible. For example, it seems reasonable that $L_{\min}$ should increase if the depletion widths increase, because such an increase enlarges the two-dimensional regions controlled by the source and drain at the expense of the region controlled by the gate. Similarly, increased junction depth increases the influence of source and drain regions, so we expect that shorter-channel devices will require smaller values of r_j (i.e., shallower junctions). Finally, as t_{ox} is increased, the screening effect of the gate is reduced, and the field lines emanating from the source and drain penetrate further into the channel region, and fewer of them terminate on the gate, causing $L_{\min}$ to increase. However, while these simple ideas appear to support the validity of Eq. 5-84, the authors of this approach in fact admit that they do not know why it works.⁶⁵ In addition, Eq. 5-84 is known to be valid only over the fitted range, and has not been tested below $0.3 \mu\text{m}$.

Nevertheless, Eq. 5-84 is useful in keeping DIBL under control. On the other hand, this approach still does not incorporate all of the necessary tradeoffs to handle such problems as hot-carrier degradation, or circuit-related constraints such as noise margin or off-current requirements. Consequently, the reduced devices designed with this criterion still need adjustments, especially as L gets deeper below $1 \mu\text{m}$.

Note as well, that in modern MOSFETs the substrate doping is not uniform. That is, ion implantation is used to increase the doping in the region between the source and drain so that the doping there is heavier than in the substrate region beneath the source and drain. This allows a shorter source-to-drain spacing. With many MOSFET designs this surface doping is nearly uniform out to the junction depth. Thus, in Eq. 5-84 it is this doping concentration (i.e., N_A in Eqs. 5-85 and 86) that must be used to calculate W_D and W_S .

In addition, another aspect of modern NMOSFETs is the formation of shallow junction extensions by implantation using the gate as a mask (the LDD structure). It is the depth of this shallow junction extension that is used in Eq. 5-84. (The entire length of the junction is not made shallow, to avoid junction series resistance and junction leakage problems.)

Finally, Eq. 5-84 does not take into account the lateral ballooning of the drain depletion region in the lightly doped substrate beneath the source and drain regions. That is, this effect may allow bulk (subsurface) punchthrough to occur between drain and source. To prevent such subsurface punchthrough the substrate doping cannot be made as low as one would like in an attempt to reduce junction capacitance and threshold variation under source-to-substrate bias. A rough guideline for selecting the minimum value of substrate doping to avoid subsurface punchthrough has been suggested by Klaassen⁶⁴ to be

$$N_{\text{SUB}} \geq N_{\text{ch}}/10 \quad (5 - 87)$$

where N_{SUB} is the substrate doping and N_{ch} is the doping in the region between the source and drain. Equation 5-87 sets a lower bound on the junction-to-substrate capacitance.

5.6.1.3 Off-Current Scaling Approach. Another scaling approach employing a more flexible criterion for acceptable "off"-state behavior (but which is also somewhat more complex than the previous subthreshold approach), has been recently proposed.⁶⁶ We will provide a brief outline of this method in this section. The crux of the procedure involves finding a channel doping profile which yields an acceptable combination of off-current I_{off} (I_D when $V_{\text{GS}} = 0\text{V}$) and V_T in a device with as small an L as possible. The following three parameters are varied to arrive at the appropriate channel doping profile: the exposed dose D_1 and centroid x_c of the V_T implant, and the substrate doping N_{SUB} . The profile is initially determined assuming that the device exhibits long-channel behavior (i.e., no significant punchthrough should be observed in the subthreshold operating region). However, the acceptable value of "off"-current in the long-channel device ($I_{\text{off,lc}}$) is deliberately picked to be smaller than I_{off} in the short-channel device being designed - in anticipation of the increase in I_{off} due to surface-DIBL in the shrunk device. The estimate of the increase in $I_{\text{off,lc}}$ due to surface-DIBL depends on the DIBL model chosen by the designer. To ensure that the subthreshold punchthrough current remains insignificant at maximum V_{DS} , a minimum N_{SUB} is identified. Finally, in the process of selecting appropriate values of D_1 and x_c , the criterion that the channel depletion-region width d should be a minimum is applied (to keep ΔV_{TSC} as small as possible when L is shrunk, as suggested by either Eq. 5-2 or barrier-lowering analysis).

In the report that describes this scaling strategy, the example to which it is applied is a symmetrical CMOS process (i.e., in which NMOSFETs are built with n^+ -poly gates and PMOSFETs with p^+ poly gates), a technological approach that is gaining more popularity as L is decreased to $0.5\ \mu\text{m}$ or smaller. Let us now also consider this scaling strategy as applied to a symmetrical CMOS process in more detail.

First, the approach assumes that the following parameters are set by system and processing constraints: V_{DD} ; the type of gate material (e.g., heavily doped poly); and minimum values of t_{ox} and r_f . The minimum manufacturable L is also limited by the lithographic process. However, I_{off} , V_T , and the actual value of L are left as parameters that can be traded off later, based on the specific circuit or system application

requirements. The tradeoffs between V_T and long-channel I_{off} ($I_{off,lc}$) values are carried out with the aid of families of curves that relate V_T and $I_{off,lc}$ for a fixed t_{ox} (with N_{SUB} as a parameter). Figure 5-45 is an example of one such family of curves, in which $t_{ox} = 100\text{\AA}$. Notice that the x-axis is expressed in $N_{off,lc}/N_{mg}$ instead of $I_{off,lc}$. However, since $I_{off,lc}$ is directly proportional to $N_{off,lc}$ (the latter being defined as the free-carrier density in the channel per unit area - e.g., per cm^{-2} - when $V_{GS} = 0V$) according to

$$I_{off,lc} = \mu (Z/L)(kT/q)qN_{off,lc} \quad (5 - 88)$$

$I_{off,lc}$ is implicitly expressed in these curves. In Eq. 5-88, Z and L are the device width and length, respectively. In addition, N_{mg} is the free-carrier concentration per unit area in the channel at midgap,* and this value can be found from

$$N_{mg} (\text{cm}^{-2}) = n_i (\kappa_{si} / \kappa_{ox})(2kT/E_g) t_{ox} \quad (5 - 89)$$

N_{mg} is therefore the free-carrier density in the channel at the point when the surface is undergoing its field-induced reversal of polarity (e.g., from p -type to n -type). Thus, N_{mg} represents the smallest free-carrier concentration in the channel in weak inversion. As is also indicated by Eq. 5-89, N_{mg} is independent of N_{SUB} . In an Si-SiO₂ MOS device at 300K, Eq. 5-89 can be written to explicitly express the relation between N_{mg} and t_{ox} as

$$N_{mg} (@ 300K, \text{ in } \text{cm}^{-2}) = 10^{10}(\text{cm}^{-3})(11.8/3.9) 22.4 t_{ox}(\text{cm}) = 1.33 \times 10^9 t_{ox}(\text{cm}).$$

Thus, for a MOSFET with $t_{ox} = 15 \text{ nm} = 1.5 \times 10^{-6} \text{ cm}$

$$N_{mg} \sim 2 \times 10^3 \text{ cm}^{-2}.$$

Thus, if $I_{off,lc}$ is given, and it is desired to relate this value to a point on a curve in which one axis is expressed in terms of $N_{off,lc}/N_{mg}$ (such as Fig. 5-45), it can be done as follows:

Example 5-1: If a maximum off-current $I_{off,lc}$ at 300K of 10 pA is required in an NMOSFET with $L = 3 \mu\text{m}$, $Z = 6 \mu\text{m}$, $t_{ox} = 15 \text{ nm}$, and $\mu_n = 400 \text{ cm}^2/\text{Vsec}$, find the value of $N_{off,lc}/N_{mg}$ that corresponds to this $I_{off,lc}$ value.

Solution: Since $N_{off,lc} = I_{off,lc} / [\mu_n (Z/L) (kT/q)q]$

$$= 1 \times 10^{-11} \text{ A} / [400 \text{ cm}^2/\text{Vsec} (6/3) (0.025V) 1.6 \times 10^{-19} \text{ C}]$$

$$= 3.1 \times 10^7 \text{ carriers/cm}^2$$

* *Midgap* is the state of a MOSFET in which band bending is causing E_F at the Si surface to co-incide with the energy level at the middle of the bandgap (which is also where E_i resides). Also at midgap, $\Phi_{surf}/\Phi_t = \ln(N_{SUB}/n_i)$ as shown in Fig. 5-46. In the special case of midgap occurring at $V_{GS} = 0V$, N_{mg} is the free-carrier concentration (cm^{-2}) in the channel for that condition.

and since N_{mg} (300K, $t_{ox} = 15$ nm) = 2×10^3 cm⁻²,

$$N_{off,lc}/N_{mg} = 1.5 \times 10^4$$

To arrive at the V_T values exhibited in a family of curves such as those in Fig. 5-45, we follow the general procedure given in section 4.6.3. That is, the value of $V_{Tud,lc}$ (assuming a uniformly doped channel region with a doping of N_{SUB}) is calculated using Eq. 3-71, and this is added to $\Delta V_T(II)$ using Eq. 4-73, to get $V_{Tud,lc}$ (Eq. 4-74 or Eq. 4-75). However, to accurately determine $\Delta V_T(II)$ the value of both D_I and x_c must be specified. In addition, the impact of N_{SUB} must be included.

That is, in the simple long-channel analysis of a MOSFET in section it was assumed that $V_{Tud,lc}$ can be shifted by choosing D_I and x_c independent of N_{SUB} .

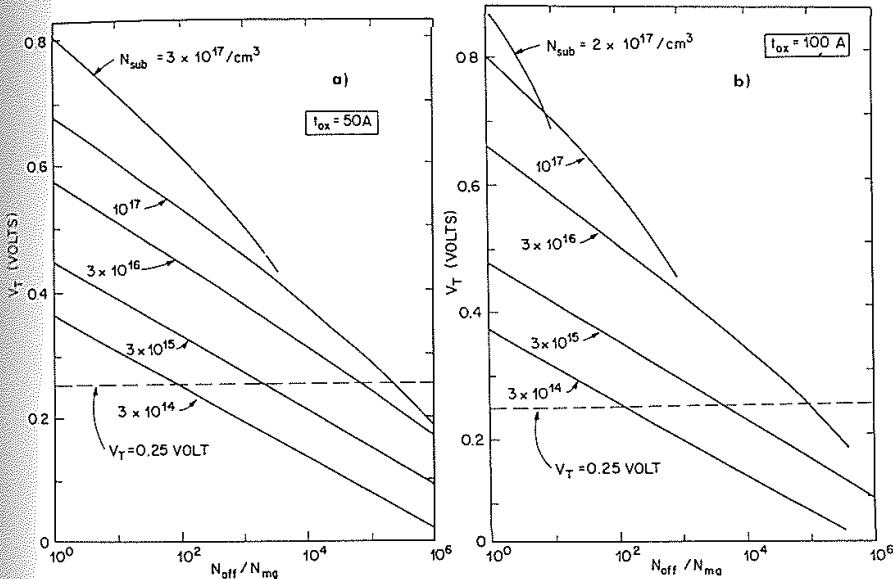


Fig. 5-45 Threshold voltage for implanted structures with the indicated substrate doping as a function of N_{off}/N_{mg} . The dose and centroid of the implants are chosen for minimum depletion width. (a) $t_{ox} = 50 \text{ \AA}$; (b) $t_{ox} = 100 \text{ \AA}$. From J.R. Brews, K.K. Ng, and R.K. Watts, Chap. 1, "The Submicrometer Silicon MOSFET," in *Submicron Integrated Circuits*, R.K. Watts, Ed., Wiley-Interscience, New York, Copyright 1989 p. 34. Reprinted with permission.

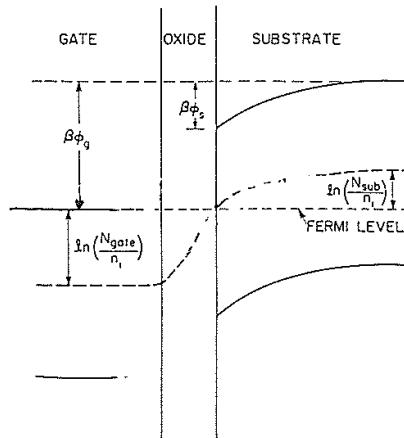


Fig. 5-46 An MOS structure at zero gate bias with midgap level and the Fermi level at the surface being equal at zero bias (*midgap* occurs).

However, for the following reasons N_{SUB} cannot be ignored, especially when designing short-channel devices: First, a minimum N_{SUB} must be identified to prevent subsurface punchthrough, e.g., $N_{SUB} \geq N_{CH}/10$. (Note that in this analysis, the effect of a punchthrough implant is not considered, and hence only the substrate doping plays a role in preventing subsurface punchthrough.) Second, the choice of N_{SUB} impacts d_i (the channel depletion region width in the presence of an V_T implant) in that larger values of N_{SUB} will cause d_i to decrease (see Eq. 4-80). To make d_i smaller (i.e., to keep ΔV_{TSC} as small as possible, as described earlier), larger N_{SUB} values are needed.* Third, N_{SUB} impacts $N_{off,lc}$, because S_t depends on d (see Eq. 4-91b). In general, increasing the value of N_{SUB} will cause the value of S_t to increase. If V_T remains the same, the consequence of an increase in S_t is a larger $N_{off,lc}$ (as is also evident in Fig. 5-45). Obviously, the first two issues call for large N_{SUB} values, while the third calls for smaller values. Hence, tradeoffs may be necessary.

Values for the exposed dose D_1 and the centroid x_c that will produce a minimum value of d are plotted versus $N_{off,lc}/N_{mg}$ with N_{SUB} as a parameter, and $t_{ox} = 50 \text{ \AA}$ in Fig. 5-47. The value of D_1 is obtained from the following equation:

* The equations used to generate the V_T versus $N_{off,lc}/N_{mg}$ curves are derived by invoking the constraint that the channel depletion width in weak inversion (d) must be a minimum. This constraint is derived from Eq. 5-2, and the fact that we want ΔV_{TSC} to be as small as possible. Since it is assumed that the device will employ the smallest t_{ox} and r_j values allowed by process technology, the last parameter in Eq. 5-2 that can reduce ΔV_{TSC} when the device is shrunk to its minimum L , is d_{max} .

$$qD_1 = (Q/Q_{mg}) (\kappa_{ox}\epsilon_0/t_{ox}) (kT/q) (BE_g/2q) \quad (5-90)$$

The value of Q/Q_{mg} is found in terms of $N_{off,lc}/N_{mg}$ with the aid of Fig. 5-48.* (It turns out that for most cases of interest the substrate doping independence of D_1 assumed in section does in fact continue to hold to larger values of N_{SUB} when the devices have thin oxides. Thus, Eq. 5-90, which is independent of N_{SUB} , continues to be valid for such cases). Once D_1 is known, its value is used to find x_c from

$$x_c/d = -D_1/(N_{SUB}d) + \{(D/N_{SUB}d)^2 + 1\}^{1/2} \quad (5-91)$$

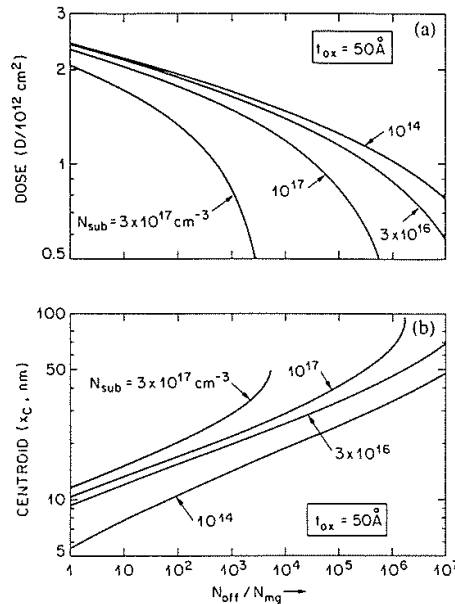


Fig. 5-47 (a) Dose in units of $10^{12}/\text{cm}^2$ as a function of N_{off}/N_{mg} with substrate doping as parameter. (b) Centroid or depletion layer width $x_c = w$ in nanometers versus N_{off}/N_{mg} . In both figures $t_{ox} = 50 \text{ \AA}$. From J.R. Brews, K.K. Ng, and R.K. Watts, Chap. 1, "The Submicrometer Silicon MOSFET," in *Submicron Integrated Circuits*, R.K. Watts, Ed., Wiley-Interscience, New York. Copyright 1989 p. 28. Reprinted with permission.

* Note that Q here is a dimensionless surface charge defined as by

$$Q = (D_1/N_{SUB}L_D) + (N_{off,lc}/N_{SUB}L_D) + (d/L_D) \quad (5-92)$$

where L_D is the Debye length, and $Q_{mg} = (\kappa_{ox}L_DE_g/2kT\kappa_s t_{ox})$. But for the purposes of our abbreviated description of this scaling strategy (described more completely in ref. 65), it suffices to note that Q/Q_{mg} can be determined in terms of $N_{off,lc}/N_{mg}$ by using Fig. 5-48.

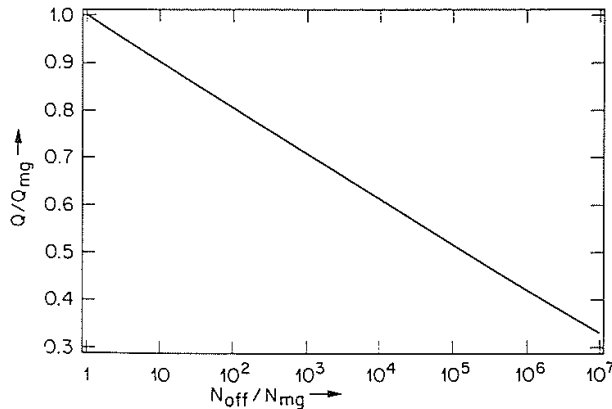


Fig. 5-48 The normalized field parameter Q/Q_{mg} of Eq. 5-90 as a function of zero-bias carrier density to midgap carrier density ratio $N_{\text{off}}/N_{\text{mg}}$. From J.R. Brews, K.K. Ng, and R.K. Watts, Chap. 1, "The Submicrometer Silicon MOSFET," in *Submicron Integrated Circuits*, R.K. Watts, Ed., Wiley-Interscience, New York, Copyright 1989 p. 28. Reprinted with permission.

Once V_T is determined for an acceptable value of $N_{\text{off},lc}$, a check is made to ensure that the chosen N_{SUB} value exceeds the minimum substrate concentration identified as being able to provide adequate subsurface punchthrough protection. If N_{SUB} is found to be large enough to prevent punchthrough, the design process can proceed to the next step - barrier-lowering analysis. (However, if the chosen value of N_{SUB} cannot prevent punchthrough, a larger N_{SUB} value may need to be accepted, which in turn may possibly require an adjustment of the maximum acceptable I_{off} .)

The barrier-lowering analysis examines the effect of shrinking L to the minimum value that can be fabricated with the best lithographic process at hand. At that length the value of ΔV_{TSC} is calculated, using the DIBL model of choice. The value of the long-channel off current $I_{\text{off},lc}$ in the device designed using the strategy outlined above now becomes the basis for determining the short-channel device I_{off} , according to

$$I_{\text{off}} = I_{\text{off},lc} \exp(q\Delta V_{\text{TSC}}/kT) \quad (5 - 93)$$

If I_{off} is less than or equal to the target value, the device with that value of L will succeed in meeting the circuit requirements and is deemed nominally manufacturable. If I_{off} is too high, either the channel length will need to be increased, or a larger value of I_{off} will need to be accommodated by changing the circuit design.

The term *nominally manufacturable* is used, because as the channel length is reduced, the sensitivity of V_T to channel length variations increases. As already noted, the impact of channel length variation is to require a more conservative setting of nominal- I_{off} so that devices with smaller than nominal V_T values will still have acceptable off-currents. As the channel length is reduced, at some point this lowered I_{off} estimate will encounter

the same punchthrough limit already discussed. If channel length is under very tight control, then V_T variations from length variation remains small, and only variation due to changes in V_{DS} will matter. However, it is more common that channel length variations will dominate the choice of L_{min} as illustrated in the next section.

5.6.2 Manufacturing Considerations in Submicron MOSFET Device Design

As noted above, the "off"-current value in submicron MOSFETs must be maintained over the range of manufacturing tolerances. For example, assume that a typical $0.8 \mu\text{m}$ NMOS transistor (i.e., the drawn channel length L is $0.8 \mu\text{m}$) has an L_{eff} of $0.7 \mu\text{m}$ and a 3-sigma manufacturing tolerance on patterned gate length of $\pm 0.12 \mu\text{m}$. Hence, the shortest L_{eff} in a device that still meets the manufacturing tolerance (i.e., patterned gate length is $0.68 \mu\text{m}$) will be $0.58 \mu\text{m}$. At this gate length, the CMOS devices will still have to possess sufficient V_T control, satisfactory hot-carrier reliability, and acceptable punchthrough characteristics. To enable a MOSFET to meet these requirements, N_{SUB} will need to be raised to a value higher than that needed for the nominal device length. As a result, the performance of the nominal gate length device ($L = 0.8 \mu\text{m}$) will be degraded through increased source/drain junction capacitance and subthreshold turnoff.

The importance of manufacturing control on patterned gate length is illustrated in Fig. 5-49.⁶⁷ This figure plots the NMOS and PMOS device leakage in the "off" state

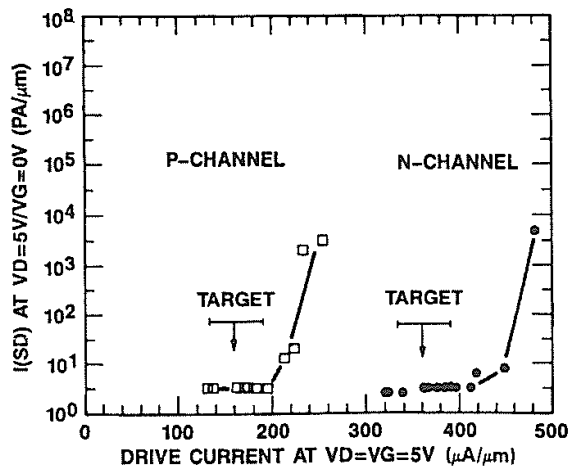


Fig. 5-49 NMOS and PMOS drive current, defined as the drain current for 5V applied to the gate and drain, versus leakage in the off state with 0V on the gate and 5V on the drain. The parameter is gate length, with a target value of $0.8 \mu\text{m}$. From R.A. Haken et al., "BiCMOS Process Technology," Chap. 3, p. 85, in *BiCMOS Technology and Applications*, A. Alvarez Ed. Kluwer, Norwell MA, Copyright 1989. Reprinted with permission.

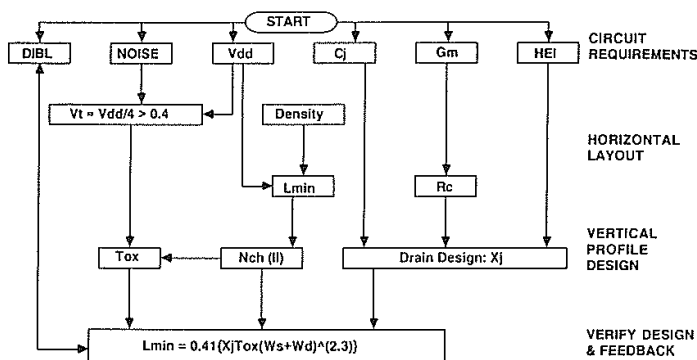


Fig. 5-50 First order MOS device design methodology. Horizontal design generally precedes the vertical profile design. The horizontal design is set to optimize intrinsic device delay (L_{eff} , R_{SD}). Vertical profile design is set to minimize short-channel effects.⁶⁸ (© 1988 IEEE).

(in this figure I_{SD} with 0V on the gate and 5V on the drain), versus drive current, defined as I_D when $V_{DS} = V_{GS} = 5V$. Although I_D is the x -axis variable, the actual parameter being varied is L , since I_D is inversely proportional to L . In this example, for the n -channel device, when $L = 0.8 \mu m$ ($L_{eff} = 0.7 \mu m$), I_D is $400 \mu A/\mu m$. The device design has been optimized for a patterned gate length of $0.8 \mu m$ with a manufacturing 3- σ control on L of $\pm 0.12 \mu m$. According to Fig. 5-49, as the 3- σ manufacturing control limit is reached, the device leakage current increases to $\sim 10^4$ pA/ μm (i.e., from Fig. 5-49 the leakage starts to rapidly rise when $I_D = 480 \mu A/\mu m$, which means it is occurring when $L = 0.68 \mu m$, or $L_{eff} = 0.58 \mu m$). Such high levels of device leakage might cause circuits to become nonfunctional, or could cause them to fail a particular circuit specification, such as maximum standby power.

5.7 DESIGN OF SUBMICRON NMOSFETS

We outline two approaches for the design of submicron NMOSFETs that have appeared in the literature. The first is fundamentally based on Eq. 5-87, while the second relies on the off-current scaling approach.

5.7.1 Design Methodology Based on the Subthreshold Scaling Equation

Figure 5-50 shows a flow-chart of the steps taken in the subthreshold-scaling approach to designing submicron MOSFETs in CMOS.⁶⁸ It assumes that the following circuit constraints on the device are defined (as dictated by the application being served): the power-supply voltage V_{DD} , the required drive current I_D in $\mu A/\mu m$ (or transconductance,

g_m), the required noise margin, the minimum intrinsic gate delay, and the maximum allowed device degradation (long-term) due to hot-carrier effects (HEI in the diagram).

The given V_{DD} and noise-margin constraint determine the choice of V_T . V_{DD} and the lithographic-processing capability limit the minimum channel length L_{min} . By thinning t_{ox} , I_D can be increased, but minimum t_{ox} is set by V_{DD} and processing constraints.

The maximum allowed parasitic resistance R_S depends on g_m (which is really the inverse of R_{ch}). For example, R_S may need to be kept smaller than 10% of R_{ch} . This may then dictate the minimum horizontal dimensions of the drain and source regions, and also the depth r_j of the source and drain junctions. The required R_S value may also determine the doping profile of the tip region of the LDD structure and may dictate whether or not salicided source/drains are necessary. The minimum gate delay τ_d will be impacted by the intrinsic device characteristics (i.e., I_D through L_{min} , and C_L through C_{gate} and C_j). By minimizing gate area and source/drain junction areas, smaller C_{gate} and C_j can be realized. Lower substrate doping also decreases C_j .

The selection of the channel doping concentration (N_{ch}) will be impacted by the substrate concentration and the V_T -adjust implant. V_T and N_A will determine the V_T adjust implant dose. The drain design (e.g., an LDD) is impacted by the desire to keep hot-carrier degradation to an acceptable value (see chap. 9). The depletion region widths of the source and drain junctions - W_S and W_D - are determined by N_{ch} and the drain structure design.

Once the values of t_{ox} , r_j , and W_S and W_D are determined, they are inserted into Eq. 5-84. If the right hand side of the equation is a number smaller than L_{min} with this group of parameters, then according to past experience it is likely that a MOSFET with this L_{min} can be successfully fabricated (i.e., the device should satisfy the subthreshold leakage current criterion given in section 5.6.1.2). Hence, if some acceptable combination of manufacturable parameters meeting this criterion can be found, fabrication success is likely. If not, a larger value of L_{min} may have to be accepted.

5.7.2 Design Methodology Based on Required Off-Current Value

The flow chart of the I_{off} design strategy of submicron MOSFETs is shown in Fig. 5-51.⁶⁹ In summary, a MOSFET is designed by selecting what appears to be an appropriate set of V_T and I_{off} values, applying the principles of long-channel MOSFET physics to derive the channel doping profile. Then the device length is shortened and punchthrough analysis and barrier-lowering analysis is performed on the shortened device to ensure that the I_{off} circuit specification is still met. If not, the design procedure must be repeated, trading off V_T and I_{off} against channel length until the best set of parameters for the application at hand is achieved. Details of the various steps in the flow chart were described in section 5.7.1.

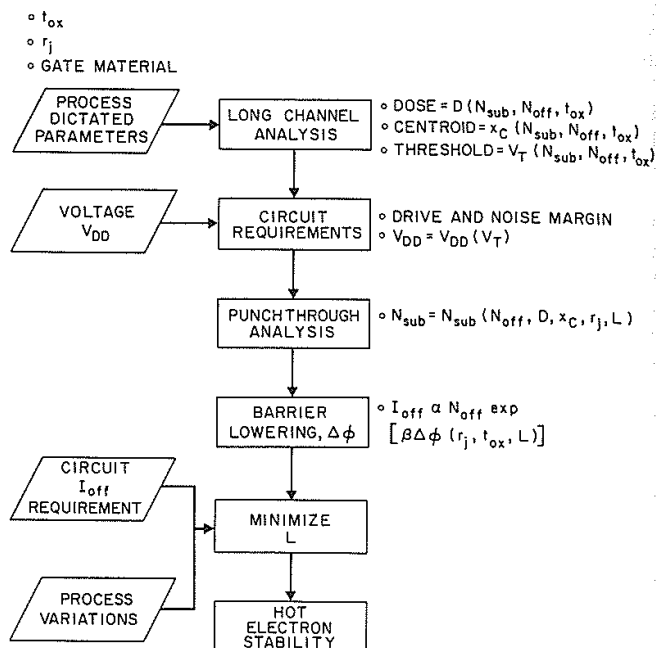


Fig. 5-51 A flow chart of the I_{OFF} strategy. From J.R. Brews, K.K. Ng, and R.K. Watts, Chap. 1, "The Submicrometer Silicon MOSFET," in *Submicron Integrated Circuits*, R.K. Watts, Ed., Wiley-Interscience, New York, Copyright 1989, p. 24. Reprinted with permission.

5.8 SHORT-CHANNEL PMOS DEVICES IN CMOS

Early MOS ICs were built using PMOS technology, primarily because enhancement-mode MOSFETs with suitable V_T values could only be fabricated as p -channel devices using aluminum or n^+ -doped polysilicon gates and uniform lightly doped n -substrates. That is, at that time it was not possible to produce enhancement-mode n -channel MOSFETs using n^+ -poly or Al gates on a uniform, lightly doped p -substrate because such structures exhibit a negative threshold voltage (see Fig. 3-20). However, with the advent of ion implantation it became possible to build NMOSFETs with positive V_T s by adjusting channel doping profiles. Thus, enhancement-mode and depletion-mode NMOSFETs could then be fabricated with little extra difficulty. Since n -channel transistors have greater drive current (and hence speed), NMOS naturally replaced PMOS as the dominant digital IC technology. During this era of NMOS dominance, interest in

p-channel MOSFETs languished. However, when CMOS arose to become the main technology for VLSI in the late 1980s, the need for enhancement-mode *p*-channel MOSFETs returned.

The fabrication of *p*-channel devices with short channels in CMOS, however, presents some unique problems which arise from having to build both NMOS and PMOS devices on the same chip. The problems revolve around the choice of a doping type for the polysilicon gate electrode and the impact that this choice has on the threshold voltage and other characteristics of PMOS devices.

5.8.1 PMOS Devices with n^+ -Polysilicon Gates

To achieve high drive current it is necessary to make the threshold voltage of a MOSFET as close to 0V as possible (with the minimum value limited by noise margin and subthreshold-leakage considerations). When MOS IC technologies were initially being developed V_T values with a magnitude of ~ 1 V were acceptable. In early PMOS (and early *p*-well CMOS) it was relatively straightforward to build *p*-channel devices with $V_T = -1$ V since the doping in the *n*-substrates was in the range of $1\text{--}3 \times 10^{15} \text{ cm}^{-3}$. As seen in Fig. 3-20, V_{Tp} of PMOS devices made on such substrates is -0.8 V to -1.0 V, which is well within the acceptable V_T range.

However, in *n*-well CMOS (which was developed after *p*-well CMOS), as well as in twin-well CMOS, the situation changes. That is, in *n*-well CMOS the doping in the well is about $10\times$ the doping in the substrate, making it at least $1 \times 10^{16} \text{ cm}^{-3}$. Twin-well CMOS becomes the well-architecture of choice as the gate lengths decrease below about $1 \mu\text{m}$. For $1 \mu\text{m}$ MOS devices (both *p*- and *n*-channel) the substrate doping density must be $2\text{--}3 \times 10^{16} \text{ cm}^{-3}$. Thus, as shown in Fig. 3-20, the V_T values of PMOS devices made with n^+ -poly gates in either *n*-well or twin-well CMOS will be at least -1.5 V, whose magnitude is too large. To make matters worse, in PMOSFETs with $L \leq 1 \mu\text{m}$, punchthrough effects are more severe than in comparably sized NMOSFETs (primarily due to the inability to make p^+ source/drain junctions as shallow as n^+ junctions, as will be described later). To suppress such punchthrough current it is necessary to increase the *n*-doping in the substrate (i.e., by an extra deep phosphorus implant into the channel of PMOS devices). The resultant increase in channel doping makes short-channel PMOS devices even more strongly enhancement mode (i.e., V_T is increased even more).

For optimal logic-gate performance the threshold voltages of the *n*- and *p*-channel devices in CMOS circuits should also have comparable magnitudes. In addition, to allow for maximum current-driving capability, they should also be as small as possible. Thus, for 5V CMOS technologies, desirable threshold voltages are 0.6 to 0.8 V for V_{Tn} and -0.6 to -0.8 V for V_{Tp} . Furthermore, as mentioned above, the most common choice for the gate material has historically been heavily doped *n*-type polysilicon. The work function of n^+ polysilicon is ideal for *n*-channel devices since these yield threshold voltages of less than 0.7 V for reasonable values of channel doping and oxide thicknesses.*

Figure 3-20 plots V_T for long-channel, uniformly doped MOSFETs with n^+ poly gates (left scale)⁷⁰ as a function of p -substrate doping (assuming the substrate is uniformly doped and that $Q_{tot} = 0$). The value of V_{Tn} is less than 1.0V for gate-oxide thicknesses of 25 nm or less and substrate dopings of less than 10^{17} cm^{-3} . Thus, V_{Tn} can easily be adjusted to the desired values by implanting boron into the channel.

However, when n^+ poly is the gate electrode of a PMOS device, adjustment of V_{Tp} to -0.7 V is not as simple, especially as the PMOS channel lengths shrink below about 2 μm . First, as we noted above, the magnitude of V_T of PMOS devices when the substrate doping exceeds $2 \times 10^{16} \text{ cm}^{-3}$ is significantly greater than -0.8V (e.g., -1.4V or larger). Even higher values of substrate doping will be needed as the channel lengths are decreased below 1 μm , implying that V_T would be even larger in submicron PMOS devices with uniformly doped substrates. In addition, such short-channel PMOSFETs are more prone to punchthrough than NMOSFETs of the same channel length. This means that the substrate doping concentration will need to be increased even further in submicron PMOS devices to suppress these effects (e.g., by performing an additional deep n -type [phosphorus] implant into the channel region).

The increased susceptibility to punchthrough of scaled-down PMOS devices with uniformly doped channels is caused in part by the dopant used to form the source and drain regions of the PMOS device, i.e., boron. Boron diffuses much more quickly than the dopant used to form the source and drain junctions in NMOSFETs (arsenic). As a result, the junction depths of the p^+ regions in 1 μm PMOS are $\sim 0.35 \mu\text{m}$, whereas the n^+ regions in 1 μm NMOS are shallower (typically only $\sim 0.2 \mu\text{m}$ deep). The deeper junctions accentuate the punchthrough susceptibility. The extra n -implant used to suppress punchthrough mentioned above must also place the additional doping at a depth below the source and drain depths to be effective (e.g., with a peak at $\sim 0.15 \mu\text{m}$ and increased doping above the background level to a depth of $\sim 0.5 \mu\text{m}$). This extra implant, however, not only increases the V_T value to about -2V, but also increases the source/drain-to-substrate capacitance.

In any case, it is obvious that the magnitude of V_T in short-channel PMOS devices of advanced CMOS technologies needs to be reduced. However, implanting the n -doped substrate with more n -type dopant would only *increase* the magnitude of V_{Tp} , rather than bringing its value closer to the desired -0.7V. To reduce the magnitude of V_{Tp} when an n^+ poly gate is used, a thin sheet of *negative* charges must therefore be placed at the Si-SiO₂ interface. This is accomplished by implanting a shallow layer of boron atoms, which as we know become negatively charged acceptor ions in the Si lattice. The boron dose must be heavy enough to overcompensate the n -surface so that a p -region (depleted of holes at $V_{GS} = 0\text{V}$) is formed. This shifts V_{Tp} toward more positive values. In practice, the boron dose depends on the substrate doping, as shown

* Note that the polysilicon layer may be combined with a layer of silicide for sheet resistance reduction; since the polysilicon remains as the underlayer of the polycide, the work function of the gate electrode will not be changed.

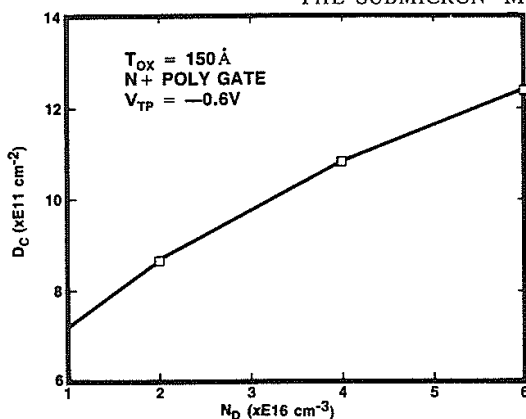


Fig. 5-52 Simulation of the counter-doping dose versus n -type substrate concentration for a fixed p -channel threshold voltage of -0.6V .¹¹⁵ (© 1984 IEEE).

in Fig. 5-52 for devices with $t_{\text{ox}} = 15 \text{ nm}$ in which a $V_{TP} = -0.6\text{V}$ is desired. Typical doses of such implanted layers are of the order of $5 \times 10^{11} \text{ cm}^{-2}$ at 25 keV .

The fact that boron is implanted to adjust both V_{TN} and V_{TP} in CMOS circuits with n^+ poly gates suggests that a single implant could be used instead of two separate implants. Figure 5-53 shows that this can be accomplished if the appropriate background dopings are chosen for the substrate and the well.⁷¹ On the other hand, it may be decided to use two separate implants in order to achieve better short-channel behavior through individual optimization of the n - and p -channel devices. In the latter case, if the boron dose needed to adjust V_{TP} is smaller than the dose needed to adjust V_{TN} , the former implant can be made into both the p - and n -channel regions without use of a mask. Then the second implant for adjusting V_{TN} is done with a mask that protects the n -well region. (A third boron implant may also be performed at that time to suppress punchthrough in the n -channel devices.) In other processes, however, it may be necessary to use a separate mask for both the V_{TP} - and V_{TN} -adjust implants.

5.8.2 Why PMOS Devices with n^+ Poly Gates and Implanted p -Surface Layers Exhibit Enhanced Susceptibility to Short-Channel Effects

The result of implanting boron atoms to adjust V_T is that such PMOS devices exhibit worse short-channel effects than NMOSFETs with the same L . This susceptibility arises mainly because the implanted boron layer has a finite thickness (i.e., instead of being an infinitely-thin sheet of charge at the surface, as would exist in the ideal case). Figure 5-54 illustrates this fact, in which the channel doping profile of a PMOS device with such a boron implant is seen to extend well below the channel surface. The penetration of the p -layer into the n -substrate gives rise to two phenomena not observed

in surface-channel NMOSFETs: (1) a pn junction is formed beneath the gate oxide in the channel, extending to a depth Y_j (as can also be seen in Fig. 5-54); and (2) the potential minimum is moved away from the surface. The latter effect gives rise to important device-behavior effects that differ from those of surface-channel MOSFETs. We explain this behavior by first comparing the location of the potential minimum in such boron-implanted-layer PMOSFETs to its location in surface-channel MOSFETs.

Figure 5-55 shows the one-dimensional potential diagrams of two PMOS devices, one with a p^+ -polysilicon gate and a uniformly doped n -substrate ($3.4 \times 10^{16} \text{ cm}^{-3}$), and the other with a n^+ -polysilicon gate and a shallow boron V_T -adjust implant. In both devices $V_{GS} = 0\text{V}$, and $V_{DS} = -5\text{V}$. The potential profile is taken at the point along the channel at which the longitudinal field vanishes (to match the fact that in a MOS-C there is also no longitudinal electric field). In the p^+ -poly gate device the potential minimum is located at the surface, implying that such a device will operate much like the surface-channel NMOSFETs described in chap. 4. In the PMOS structure with the n^+ -poly gate and boron V_T -adjust implant, however, we see that the potential minimum is away from the Si surface. The mobile carriers (holes) will thus be found at highest density near the location of the potential minimum. This implies that, unlike in the

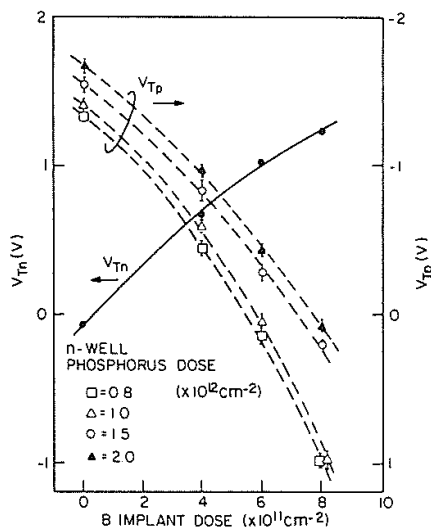


Fig. 5-53 Threshold voltages of n -channel (V_{Tn}) and p -channel (V_{Tp}) transistors as a function of boron threshold-adjustment dose. The CMOS structure uses an n -well implanted into a p -substrate (whose doping level is $6 \times 10^{14} \text{ atoms/cm}^3$). V_{Tp} results are shown for various implant doses of the n -well.⁷¹ (© 1980 IEEE).

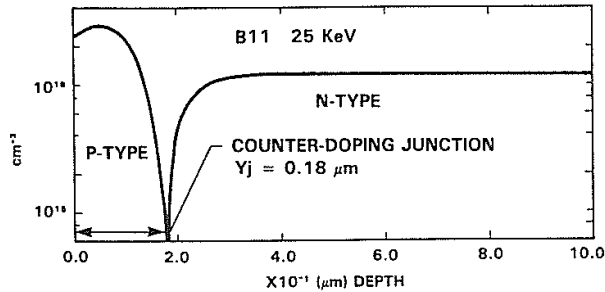


Fig. 5-54 Simulated channel profile for a p -channel MOSFET with an implanted B dose at the channel surface.¹¹⁵ (© 1984 IEEE).

case of conventional surface-channel devices, more current will flow in a channel *beneath* the surface of the device than at the surface.⁷² Such PMOS devices are therefore referred to as *buried-channel* (BC) transistors. As also seen in Fig. 5-56 (which gives the calculated variation of the potential as a function of distance below the surface as the channel pn junction depth is varied), the potential minimum moves further into the substrate as the thickness of the implanted p -layer is increased. The mobility of carriers in the buried channel is $\sim 15\%$ higher than carriers at the surface, and thus BC MOSFETs exhibit a somewhat larger drive current than comparable surface channel FETs. While this is an advantage, it is outweighed by several other disadvantages exhibited by BC FETs. Let us examine how the presence of the channel junction and

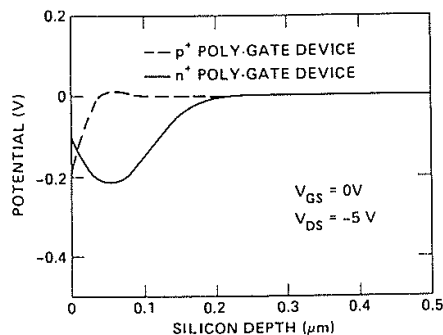


Fig. 5-55 Simulations of the electrical potential perpendicular to the surface in the channel region of two long-channel PMOS devices using GEMINI and SUPREM III simulations of devices with n^+ polysilicon and p^+ polysilicon gates with effective gate lengths of $20\ \mu\text{m}$.¹¹⁴ (© 1985 IEEE).

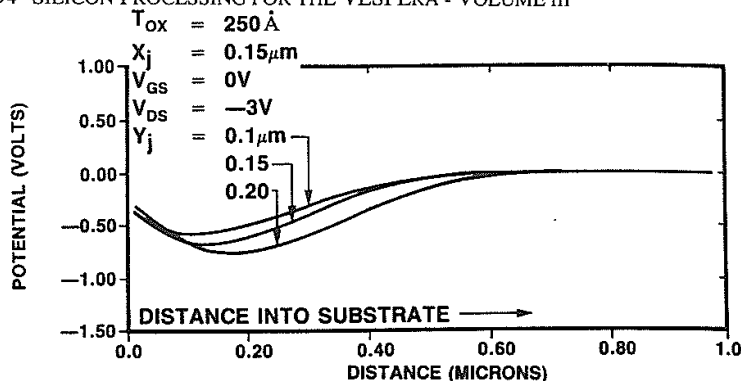


Fig. 5-56 Channel potential profile versus vertical distance for a p -channel MOSFET with $Y_j = 0.1, 0.15$, and $0.2 \mu\text{m}$, and $V_{DS} = -3\text{V}$ and $V_{GS} = 0\text{V}$.¹¹⁵ (© 1984 IEEE).

the shifting of the potential minimum beneath the Si surface worsen the susceptibility of the PMOS device to short-channel effects.

5.8.2.1 Threshold-Voltage Roll-Off in Buried-Channel PMOS.

As the channel length of MOSFETs is decreased, the threshold voltage decreases from the value it exhibits in long-channel devices. This effect, as described in section 5.2.2, is referred to as V_T -roll-off. The V_T -roll-off effect is more severe in BC PMOS devices than in surface channel FETs (Fig. 5-57a).¹¹⁴ This is due to an effect in BC FETs that does not occur in surface-channel FETs. That is, the implanted p -layer is a lightly doped p -region that extends entirely between the p^+ source and drain regions of the device. Holes diffuse from the regions of high concentration into the lightly doped layer as depicted in Fig. 5-57b.¹⁰⁶ The diffusion length d_p is given by¹⁰⁷

$$d_p = L_D \tan^{-1} [(N_p^+/N_p^-)^{1/2}] \quad (5-94)$$

where

$$L_D = [(2 \kappa_{\text{Si}} \epsilon_0 kT)/(q^2 N_p^-)]^{1/2} \quad (5-95)$$

and N_p^+ and N_p^- are the doping levels in the p^+ -source/drain and p -layer regions, respectively. If the diffusion length is smaller than half the channel length, the net effect of this hole diffusion is to increase the effective positive charge in the channel, thus making its threshold voltage smaller in magnitude. As the channel length is decreased, the proportionate contribution of the diffused mobile charge to the total channel charge increases, and as a result, the threshold is shifted to smaller values. Eventually, at short enough channel lengths, the diffused hole distribution from the source meets that from the drain.¹⁰⁸ When that occurs, the potential barrier between the source and drain is reduced, giving rise to poor subthreshold characteristics.

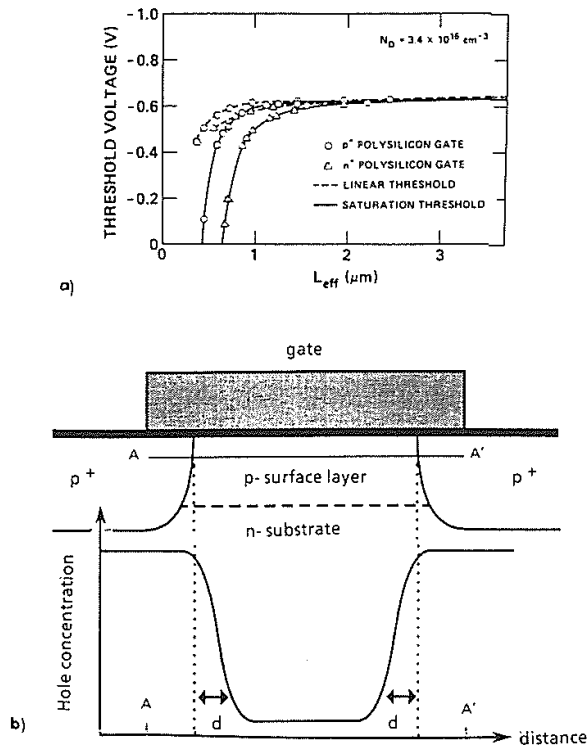


Fig. 5-57 (a) Threshold voltage variation as a function of channel length L_{eff} in surface-channel (p^+ -poly gate) and buried-channel (n^+ -poly gate) PMOSFETs.¹¹⁴ (© 1985 IEEE). (b) The origin of short-channel threshold shifts in buried-channel PMOS transistors. From A.G. Lewis and J.Y. Chen, "Current Trends in MOS Process Integration," in N.G. Einspruch and G. Gildenblatt, Eds. *Advanced MOS Device Physics*, Academic Press, San Diego CA, 1989, p. 58. Reprinted with Permission.

Experimental values of V_T in BC PMOS devices with a gate oxide thickness of 15 nm versus channel length are shown in Fig. 5-58a. As in Fig. 5-57a we see that the magnitude of V_T decreases as L is reduced. Simulation studies indicate that the severity of the V_T roll-off is increased as Y_j is made deeper (see Fig. 5-58b). This is due to the fact that more holes are able to diffuse into the thicker p -layer of the channel region from the source/drain regions as Y_j is made deeper, accentuating the roll-off effect. As in surface channel devices, V_T roll-off is decreased as t_{ox} is made smaller (Fig. 5-58c). Thus, we conclude that smaller Y_j and t_{ox} values are needed to mitigate the V_T -roll-off effect in BC PMOS.

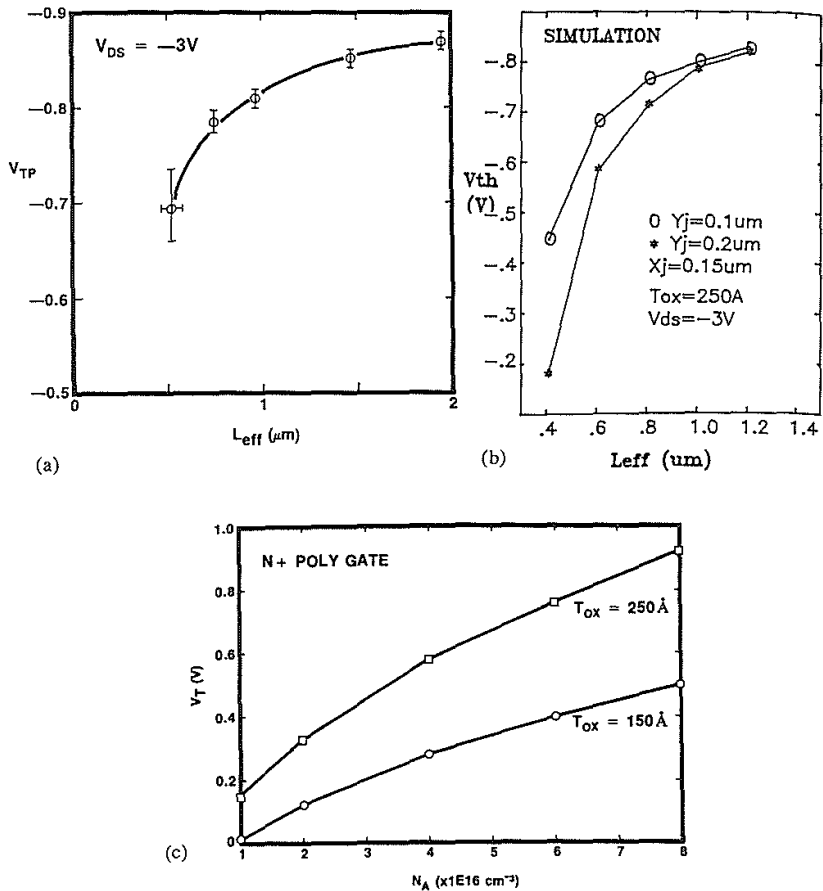


Fig. 5-58 (a) Experimental data of p -channel threshold voltage versus L_{eff} for gate oxide thickness of 15 nm. (b) Simulation of threshold voltage versus L_{eff} for p -channel MOSFETs with different Y_j . (c) Simulated p -channel threshold voltage versus L_{eff} for two different gate oxide thicknesses. From K. M. Cham, et al., *Computer Aided Design and VLSI Device Development*, p. 187-8, Copyright 1986 Kluwer Academic Publishers. Reprinted with permission.

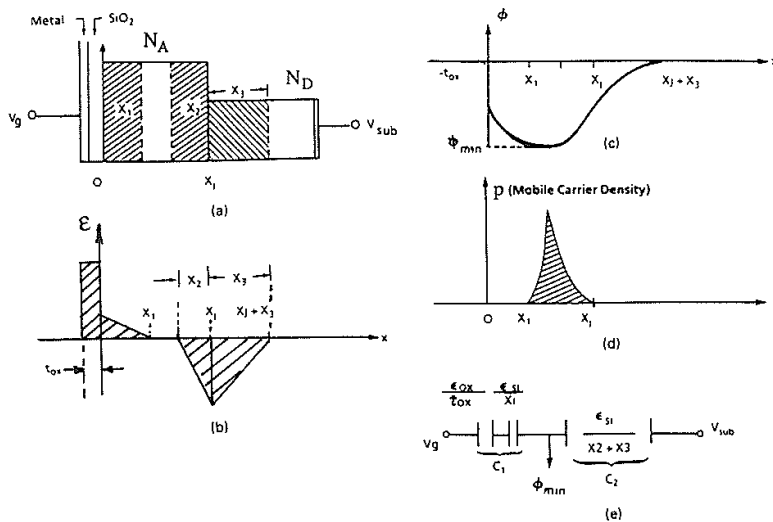


Fig. 5-59 Buried-channel PMOSFET when biased to inversion: (a) Doping profile and depletion regions; (b) electric field; (c) electron potential; (d) mobile carriers; and (e) equivalent capacitances in sub-pinchoff condition.¹²⁰ (© IEEE 1978).

5.8.2.2 Subthreshold Swing (S_t) in Buried-Channel PMOS. The PMOSFETs used in CMOS must be enhancement-mode devices; i.e., they need to be OFF when $V_{GS} = 0V$. If such PMOSFETs are buried-channel devices, the entire implanted p -layer in the BC PMOS devices must be depleted of holes when $V_{GS} = 0V$. On the other hand, upon applying a gate voltage equal to the threshold voltage, the p -layer of such devices is no longer depleted. That is, holes will have returned to the layer, forming a p -type conducting channel between the source and drain. But this channel will exist beneath instead of at the surface. Thus, for gate voltages greater than V_T , two depletion regions will be present in this p -layer; one that is formed between the surface and the buried channel (to a depth x_1) and the other due to the channel pn junction, as is shown in Fig. 5-59a. The BC MOSFETs thus have two capacitances in addition to the oxide capacitance $\kappa_{ox}\epsilon_0/t_{ox}$; namely the depletion capacitance from the surface ($\kappa_{si}\epsilon_0/x_1$), and the depletion capacitance from the pn junction ($C_2 = \kappa_{si}\epsilon_0/[x_2 + x_3] \sim \kappa_{si}\epsilon_0/x_3$). The potential minimum is located in the position between these two capacitors, as shown in Fig. 5-59e. As also noted earlier, the location of the potential minimum causes most of the free carriers to appear in the bulk Si rather than at the Si-SiO₂ interface. It is expected that since the free carriers are now farther away from the surface they will be harder to modulate by the gate voltage. (In the same way that

carriers in a surface-channel MOSFET with a thicker gate oxide are harder to modulate - because of reduced gate-to-channel capacitive coupling - than in one with a thinner gate oxide.) Hence, to achieve the same degree of turn-off, the gate voltage will need to be reduced more than in the case of a surface-channel device. Thus, we expect that the turn-off characteristic will be less sharp, which means that S_t will be larger.

Quantitatively the buried channel turn-off characteristic can be modeled by an expression that gives the subthreshold swing S_t in a structure with two such depletion regions, i.e.,

$$S_t = 2.3 (kT/q) (1 + C_2/C_1) \quad (5 - 96)$$

$$= 2.3 (kT/q) [1 + (\kappa_{ox}\epsilon_0 x_1 + \kappa_{si}\epsilon_0 t_{ox})/(\kappa_{ox}\epsilon_0 x_2 + \kappa_{ox}\epsilon_0 x_3)] \quad (5 - 97)$$

where C_1 is the oxide capacitance (C_{ox}) in series with the depletion capacitance from the surface ($\kappa_{si}\epsilon_0/x_1$), and C_2 is defined above. Let us compare Eq. 5-96 with Eq. 4-91 (which gives S_t for the case of surface-channel MOSFETs, and is repeated here for convenience).

$$S_t = \ln 10 (d \ln I_D / d V_{GS})^{-1} = 2.3 (kT/q) (1 + C_d/C_{ox}) \quad (5 - 98)$$

By noting that C_1 is smaller than C_{ox} , and C_2 is typically comparable in value to C_d , we see that S_t is expected to be larger in BC devices. Experimental results confirm that S_t is worse (i.e., larger) in buried channel devices, especially in short-channel PMOSFETs. That is, S_t values in buried-channel PMOSFETs range from 0.5 to 1.5 V/decade, compared to the 70-120 mV/decade S_t values observed in active NMOSFET devices. Figure 5-60 compares the subthreshold behavior of the two 0.8 μm BC PMOS

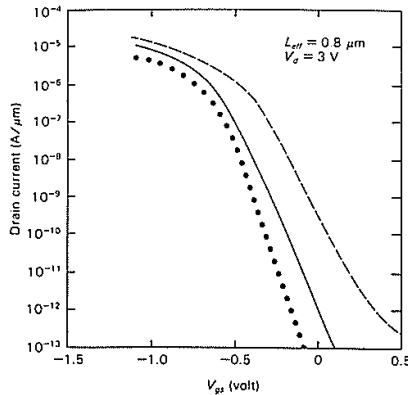


Fig. 5-60 Subthreshold I-V curves in PMOS device in CMOS. Dashed lines: buried channel PMOS in a lightly doped substrate. Solid lines: buried-channel PMOS in a retrograde n -well. Dotted lines: surface channel NMOS drawn for comparison.⁷² (© IEEE 1982).

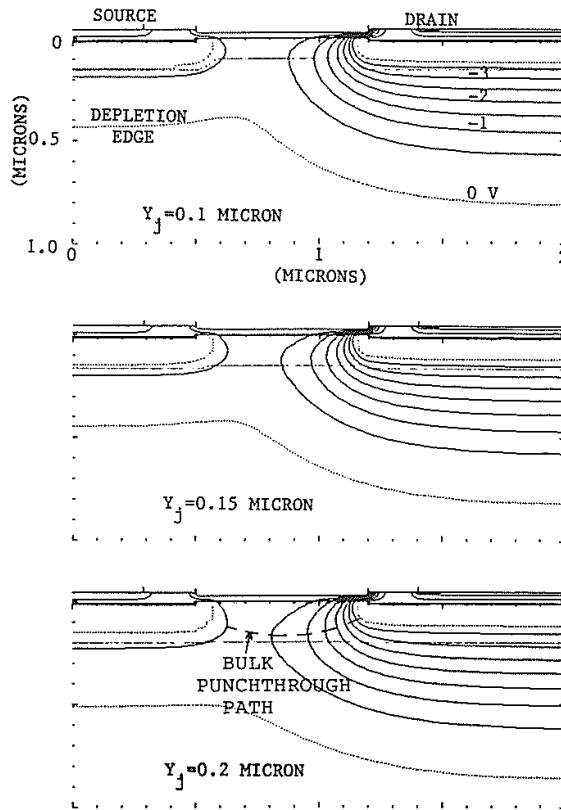


Fig. 5-61 Simulated 2-D potential profile for p -channel MOSFET with $Y_j = 0.1, 0.15$, and $0.2 \mu\text{m}$. $V_{GS} = 0\text{V}$ and $V_{DS} = -3\text{V}$.¹¹⁵ (© 1984 IEEE).

devices with that of a comparable surface-channel NMOSFET. (Note that the V_{GS} scale for the NMOSFET should be $+1.5\text{V}$ to -1.5V from left to right.) It indicates that S_t is larger for the BC PMOS devices.

The effect on S_t of the depths of the channel and source/drain junctions Y_j and r_j , respectively, and of the gate-oxide thickness t_{ox} has also been investigated with simulation studies. Figure 5-61 illustrates the 2-D simulated potential profiles for a $0.5 \mu\text{m}$ p -channel MOSFET as Y_j is increased from 0.1 to $0.2 \mu\text{m}$. It is observed that as Y_j is increased the potential contours spread more from drain to source, and the bulk punchthrough path passes through the depletion region near the pn junction. Figure 5-62 shows the calculated values of S_t for three Y_j values (i.e., $0.1, 0.15$, and $0.2 \mu\text{m}$), as r_j is varied in a BC PMOS device with an $L_{eff} = 0.5 \mu\text{m}$, a $t_{ox} = 25 \text{ nm}$,

bulk punchthrough path passes through the depletion region near the pn junction. Figure 5-62 shows the calculated values of S_t for three Y_j values (i.e., 0.1, 0.15, and $0.2\ \mu\text{m}$), as r_j is varied in a BC PMOS device with an $L_{\text{eff}} = 0.5\ \mu\text{m}$, a $t_{\text{ox}} = 25\ \text{nm}$, and $V_{\text{DS}} = -3\text{V}$. Each curve corresponds to a fixed value of Y_j . The threshold voltage [defined as the gate voltage at which the drain current is $100\ \text{nA} \cdot (Z/L)$] is specified at each point. The data shows that S_t is strongly dependent on Y_j and r_j . That is, as Y_j is increased for a fixed value of r_j , S_t increases. For a fixed value of Y_j , S_t also increases as r_j is increased. In Fig. 5-63 the simulated value of S_t for two different gate oxide thicknesses is shown as Y_j is varied. It is seen that S_t improves for thinner gate oxides, but the dependence of S_t on Y_j is only slightly reduced. The overall conclusion from these studies is that small values of Y_j , r_j , and t_{ox} are also needed to realize small values of S_t in submicron BC PMOS devices.

5.8.2.3 Punchthrough Susceptibility of Buried-Channel PMOS.

The larger S_t values exhibited by BC PMOS devices implies that they are susceptible to larger leakage currents in the OFF state. In fact, such relatively large leakage currents are commonly observed in short PMOSFETs, especially at high drain bias. For example, it has been shown that in submicron PMOS transistors ($L_{\text{eff}} = 0.8\ \mu\text{m}$), I_{DST} will increase by two orders of magnitude if V_{DS} is increased from -1V to -10V .⁷⁴ In this case, if I_{DST} was $1\ \text{nA}$ at -1V , it would increase to $0.1\ \mu\text{A}$ if V_{DS} was raised to -10V . It would be useful to be able to predict the value of I_{DST} prior to fabricating the device.

As described in section 5.3.1 an increase in S_t has been correlated to the onset of subsurface punchthrough in surface-channel MOSFETs. The increase in S_t in BC PMOS devices as a result of various device parameters (i.e., L , Y_j , r_j) discussed in the previous section can also be considered from the perspective of the phenomenon of punchthrough (subsurface-DIBL), which was introduced in section 5.3. However, we will see that the subthreshold leakage behavior in BC PMOSFETs differs from that observed in surface-channel FETs. To help explain this difference, we will once again

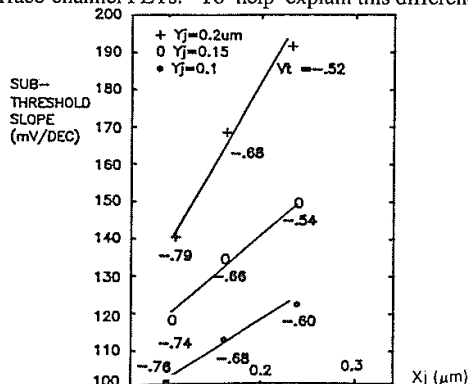


Fig. 5-62 Simulated subthreshold slope versus r_j and Y_j for p -channel MOSFETs in which $L_{\text{eff}} = 0.5\ \mu\text{m}$, $t_{\text{ox}} = 250\text{\AA}$, and $V_{\text{DS}} = -3\text{V}$. V_T is the threshold voltage.¹¹⁵ (© 1984 IEEE).

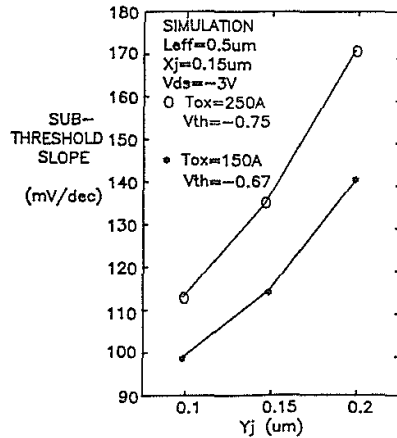


Fig. 5-63 Simulated subthreshold slope versus Y_j and t_{ox} .¹¹⁵ (© 1984 IEEE).

employ the punchthrough modeling technique described in section 5.3.2, but here it will be applied to quantitatively model I_{Dst} in BC PMOSFETs. The results of this analysis will permit us to compare the subthreshold leakage behavior of surface-channel and buried channel MOSFETs.

We recall that punchthrough current is determined by carrier injection over the potential energy barrier at the saddle point of the electrostatic potential in the bulk (see Fig. 5-64).¹¹⁷ Therefore, analysis of the channel potential, especially the change of the potential at the saddle point P of the BC PMOS device with V_{DS} is necessary. This analysis can be performed with the aid of 2-D device simulators,** and here we present the results of one report that performed such a study using the device simulator GEMINI.⁷⁴ The simulated potential along the direction from source to drain of a BC PMOSFET in an n -well with $L_{eff} = 0.55 \mu m$ as V_{DS} is increased is shown in Fig. 5-65. It is observed that the magnitude of ϕ^* (the potential at the saddle point P), begins to drop as soon as V_{DS} is increased above 0V. This is due to the fact that there is no pn junction along the channel region in a p -channel BC MOSFET. (Conversely, in the case of short, surface-channel MOSFETs, at low drain-bias values the V_{DS} drop occurs almost entirely across the drain-channel junction, and the barrier at the source does not decrease as long as V_{DS} remains small. Only when V_{DS} becomes large enough to cause subsurface-DIBL does the barrier begin to drop as V_{DS} exceeds this value.) The decrease of the potential barrier height ($\phi^* - V_S$) in a BC PMOSFET as V_{DS} is increased is plotted in Fig. 5-66. Here we see that $(\phi^* - V_S)$ is linearly proportional to the increase in drain voltage, even at low values of V_{DS} .

** Note that an entirely analytic model of punchthrough in PMOSFETs has also been published by Skotnicki et al.¹¹⁷

Once $(\phi^* - V_S)$ is found for a given value of V_{DS} , it can be used to calculate I_{Dst} with the aid of an analytic expression (Eq. 5-18b) as discussed in section 5.3. This expression for I_{Dst} is repeated here for the case of a PMOS device:

$$I_{Dst} = -q D_p Z \frac{t^*}{L^*} \frac{n_i^2}{N_D} \exp\left[\frac{q}{kT}(V_S - \phi^*)\right] \quad (5-99)$$

To exploit the results plotted in Fig. 5-66, we express Eq. 5-99 in logarithmic form as

$$\log(I_{Dst}) = \{\log[-q D_p Z (t^*/L^*) (n_i^2/N_D)]\} - q(\phi^* - V_S)/k \quad (5-100)$$

or

$$\log(I_{Dst}) = A - q\phi_b/kT \quad (5-101)$$

where $A = \log[-q D_p Z (t^*/L^*) (n_i^2/N_D)]$, and $\phi_b = (\phi^* - V_S)$. Now, Fig. 5-66 shows that the decrease in ϕ_b is linearly proportional to V_{DS} , i.e.,

$$\phi_b = (\phi_{b0} - b |V_{DS}|) \quad (5-102)$$

where ϕ_{b0} is the barrier height at $V_{DS} = 0V$, and b represents the slope of the curve of Fig. 5-66 (i.e., $d\phi_b/dV_{DS}$). We also note that the values of both ϕ_{b0} and b can be determined from Fig. 5-66. Then we can write Eq. 5-101 as

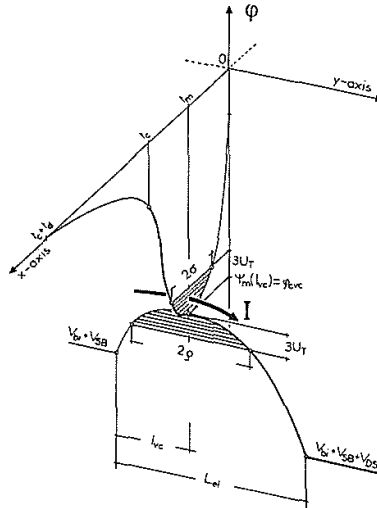


Fig. 5-64 Perspective view of the electrical potential distribution in two perpendicular cross sections of the channel of a buried-channel PMOSFET. Shaded regions are those corresponding to t^* and L^* in Eq. 5-99.¹¹⁷ (© 1989 IEEE).

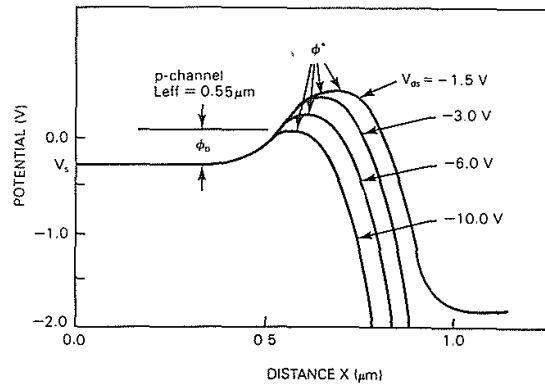


Fig. 5-65 Simulated potential distributions along the buried channel of a PMOSFET for various drain voltages.⁷⁴ (© 1988 IEEE).

$$\log(I_{Dst}) = A - q(\phi_{b0} - b|V_{DS}|)/kT \quad (5-103)$$

or

$$= A' + V_{DS}/R \quad (5-104)$$

where $A' = A - q\phi_{b0}/kT$ and $R = kT/bq$. The parameter R is the incremental voltage that the drain can sustain before I_{Dst} increases by another order of magnitude ($R = \partial V_{DS}/\partial \log I_{Dst}$). The larger the value of R , the lower the DIBL rate, hence the better the MOSFET is able to maintain low subthreshold leakage as the drain bias rises.

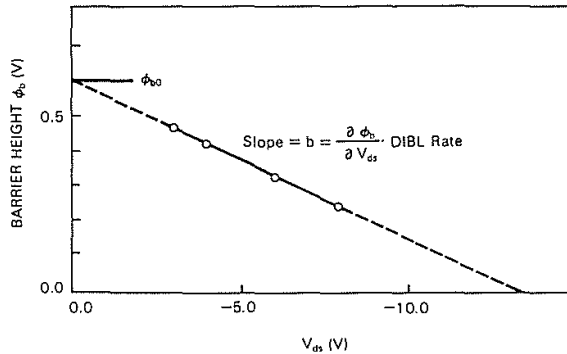


Fig. 5-66 Potential barrier height as a function of drain voltage.⁷⁴ (© 1988 IEEE).

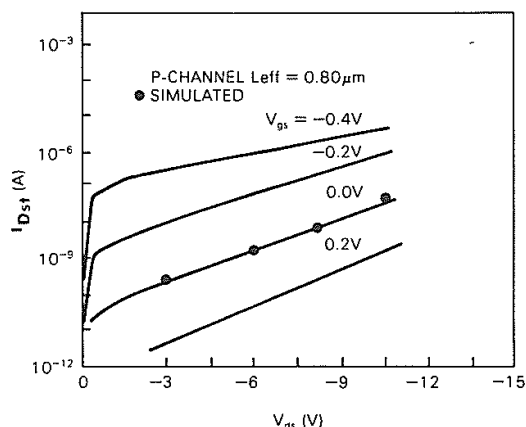


Fig. 5-67 Subthreshold I_{Dst} - V_{DS} characteristics for a submicron PMOSFET.⁷⁴ (© 1988 IEEE).

Equation 5-104 implies that the relationship between $\log I_{Dst}$ and V_{DS} should be linear in BC PMOSFETs operating in subthreshold. Figure 5-67 shows the measured $\log I_{Dst}$ - V_{DS} curves for a $0.8 \mu\text{m}$ BC PMOSFET, together with some simulated I_{Dst} values at $V_{GS} = 0\text{V}$. The agreement between the experimental and simulated data is good. We also see that R in this device is about -4.4V/decade for $V_{GS} = 0\text{V}$. This means that if I_{Dst} is 10 pA at $V_{DS} = -1\text{V}$, it will increase to 10 nA at $V_{DS} = -10\text{V}$.

Now we can compare the I_{Dst} behavior of SC and BC MOSFETs with respect to externally applied biases. We note that in the case of surface channel MOSFETs two regions of different slope are observed in the $\log I_{Dst}$ versus V_{DS} curves (see Fig. 5-11). At low values of V_{DS} the value of R is large (e.g., in Fig. 5-11b $R = 12.6\text{V/decade}$ when $V_{DS} < 6.0\text{V}$), but when V_{DS} exceeds about 8.0V , the value of R drops to about 1.0V/decade . This change implies that in surface-channel MOSFETs there is a transfer from surface-DIBL to subsurface-DIBL (punchthrough) - e.g., in the device of Fig. 5-11b at about 6.0V . Also, up to some value of V_{DS} (e.g., V_{PT}), the drain bias has little effect on I_{Dst} . In BC PMOSFETs, however, the value of R is relatively constant (e.g., it has a magnitude of $R = 4.4\text{V/decade}$ for all values of V_{DS} in the case of the device characterized in Fig. 5-67). This implies that the nature of subthreshold leakage current in BC PMOSFETs does *not* undergo a change from surface- to subsurface-DIBL, but instead the DIBL phenomenon in a BC device is a combination of both effects. Furthermore, the intermediate value of R also supports the view that *both* the gate and the drain voltages affect the potential of the p -type layer, and thus both can impact I_{Dst} at all values. Finally, this indicates that the measured slope of $\log I_{Dst}$ versus V_{DS} (defined as $1/R$), reflects the effect of V_{DS} on ϕ_b through DIBL, and permits a comparison of the subthreshold leakage behavior of various PMOS device structures.

5.8.3 Techniques Used to Suppress Short-Channel Effects in Buried-Channel PMOS

5.8.3.1 Techniques Used to Keep the Implanted p -Layer as Thin as Possible. The most obvious solution to suppressing short-channel effects in PMOS devices is to increase the PMOS device channel length. This is done in many CMOS technologies, and it is the reason that the minimum channel lengths of PMOS devices are often larger than those of the NMOS devices on the same chip. But we also noted that to achieve small values of S_T , less V_T -roll-off, and larger punchthrough voltages in BC PMOS devices, the p -buried layer depth Y_j must be as shallow as possible. That is, the thinner the p -layer, the more the device resembles a surface channel FET. Several methods have been reported to realize shallow p -layer implant profiles, including the following:

1. Using the lowest practical implant energy, which in ordinary medium-current implanters is typically 20-25 keV.
2. Implanting BF_2^+ (which produces shallower boron layers than implantation with elemental boron, see Vol. 1, chap. 9).
3. Using a high-energy n -implant (e.g., P^+ at 400 keV) to form a retrograde n -well in the regions where the PMOS devices are formed. This places more n -type dopant atoms beneath the channel pn junctions, and the more heavily doped regions absorb the drain voltage in a shorter distance, while simultaneously squeezing the channel pn junction toward the surface, as shown in Fig. 5-68.

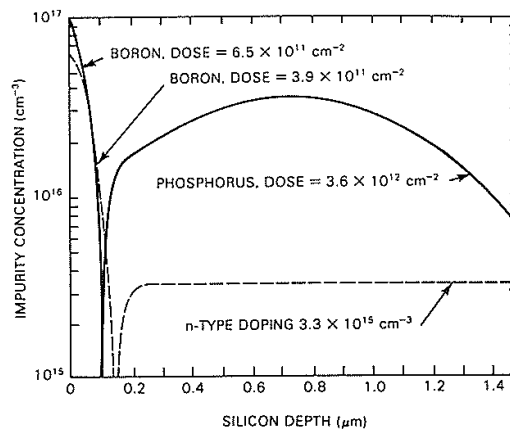


Fig. 5-68 Simulated doping profiles in PMOS channel regions: dashed lines lightly doped substrate; solid lines, retrograde n -well.⁷² (© 1982 IEEE).

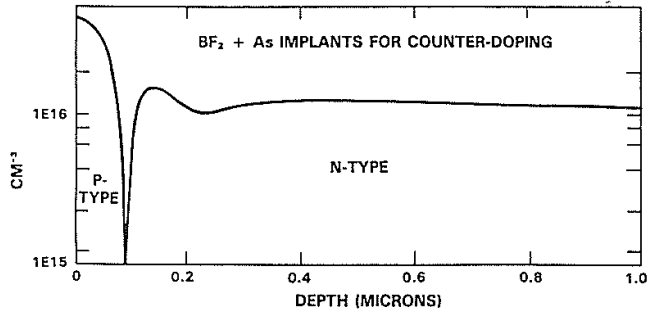


Fig. 5-69 (a) Simulated channel profile for p -channel with BF_2 and As counterdoping.¹¹⁵ (© 1984 IEEE).

4. An extra n -type implant (typically arsenic) can be performed into a conventional n -well or n -substrate, with its peak located below the p -layer immediately following the BF_2^+ implant (Fig. 5-69a). The n -doping under the p -layer is thus raised, making the junction depth lie further up the tail of the boron distribution (and thereby making it shallower). The arsenic diffuses relatively slowly, and thus will largely remain in the vicinity of where it was implanted. This forms a steep profile at the counter-doping junction, and maintains a shallow junction depth. The simulation shown in Fig. 5-69a indicates that Y_j with such an arsenic implant is $\sim 0.09 \mu\text{m}$. Such an extra implant can also be used with a retrograde n -well.

Two drawbacks of this technique are that since the boron channel implant is compensated by an arsenic implant very good process control must be maintained for it to be successful, and it also requires an extra masking step to prevent the arsenic implant from entering the n -channel transistor regions.

To prevent a shallow implanted-boron layer from growing thicker, it is necessary to use a reduced thermal budget in order to restrict boron diffusion following the implant. Specific steps for restricting boron redistribution include the following:⁷⁶

1. Implanting the boron through the gate oxide. This avoids the oxidation-enhanced diffusion of boron that would occur during the growth of the gate oxide if the implant were performed first. In this case, it is necessary to prevent the gate oxide from becoming contaminated during implant, either by material sputtered by the beam line or by vaporized resist material used as a mask against the implant. To prevent such contamination, a thin layer of polysilicon may be

deposited on the gate oxide prior to the implant (in fact, immediately after the oxide is grown).⁷⁷

2. After the implant has been performed, the remainder of the polysilicon film is deposited. In order to avoid the 900°C thermal cycle that would have to be used if the poly were doped following deposition, this polysilicon is doped *in situ* at ~600°C with phosphorus.

3. A BPSG glass layer is used as the dielectric between the gate and the first level of metal. A significantly lower temperature can be used to flow BPSG than PSG (e.g., 900°C versus 1000°C). A lower temperature cycle can thus be used to smooth the surface topography (flow step) and gently taper the contact holes after etch (reflow step).

5.8.3.2 Shallow p^+ Source/Drain Junctions to Suppress Punchthrough in PMOSFETs. As noted in our discussion of MOSFET scaling in section 5.6.1.2, reducing the source/drain junction depths also decreases the susceptibility to short-channel effects. By reducing the vertical depth of the source/drain junctions the lateral spread of the dopants in these regions beneath the gate also becomes smaller. Hence, a longer effective channel length is possible for a fixed drawn gate length. In addition, the lateral spread of source and drain depletion regions below the surface layer makes a MOSFET more prone to punchthrough. Consequently, shallow source and drains effectively suppress subsurface punchthrough paths.

In NMOSFETs, it is easier to fabricate shallow n^+ source/drain regions (i.e., $\leq 0.2 \mu\text{m}$ in depth) because dopants with heavy ionic masses (usually arsenic) are implanted to form these regions. They yield implant profiles with small projected ranges and little channeling. Furthermore arsenic has a low diffusivity, so that little redistribution occurs during postimplant annealing. In PMOSFETs, however, the only practical p -type dopant for forming the p^+ regions is boron, and junctions formed with boron are much deeper than those formed with arsenic for several reasons. First, the projected range of the low-mass boron is relatively larger, even at low implantation energies. Ion channeling also causes the implanted boron profile to have a long tail, even if the target wafers are tilted to minimize this effect. Thus, the minimum junction depth which results when boron is implanted into a lightly doped n -type crystalline silicon wafer is about $0.35 \mu\text{m}$.

The high diffusivity of boron compounds the problem. That is, in CMOS processes both the n - and p -implants that form the NMOS and PMOS source and drain regions, respectively, must be annealed. The anneal used for the arsenic implant will cause significant ($>0.1 \mu\text{m}$) redistribution of the boron implant. In CMOS technologies with channel lengths of $2 \mu\text{m}$ or longer, this issue does not arise because the boron can be implanted after the arsenic implantation and anneal steps (i.e., allowing the boron to be annealed only for the minimum time and temperature needed for full activation). Thus, boron redistribution after implantation can be minimized. However, in submicron CMOS processes it is generally not possible to perform the boron implant after the n^+ arsenic implant. This is because submicron NMOS devices usually use LDD structures,

in which the arsenic implant is performed after spacers have been formed. The boron p^+ implant, however, must be carried out before the spacers are formed for two reasons. First, this ensures that the boron implant sufficiently counterdopes the n^- implant (used to form the LDD tip region) everywhere in the areas of the PMOS source and drain (i.e., the n^- implant is typically done unmasked to maintain process simplicity). Second, this also guarantees that the p^+ doping will extend under the gate. Thus, in CMOS processes that use LDDs in the NMOS devices the p^+ regions must experience the same anneal cycle as the n^+ regions. Alternative approaches have been proposed to avoid subjecting the implanted boron to the n^+ anneal, as will be described below.

Techniques to decrease the p^+ junction depths fall into two groups: (1) those which set out to reduce the junction depth of the implanted boron profile; and (2) those which attempt to reduce the diffusion of the boron during the post-implantation anneals.

The approaches mentioned earlier when discussing techniques to reduce the depth of the as-implanted boron region in the channel of PMOS devices also apply to reducing the source/drain junction depths. These include use of BF_2^+ instead of atomic boron as the implant species, and implanting through a screen oxide. Pre-amorphizing the wafer surface region into which the junction dopants will be implanted to a depth of several tenths of a micron (with an earlier implant of silicon or germanium) has also been investigated (see Vol. 2, chap. 3). The amorphous layer effectively prevents channeling, and subsequent annealing at a low temperature is sufficient to achieve recrystallization by solid-phase epitaxy. Use of very-low-energy ion implanters (<5 keV) for implanting the boron has also been studied, as have techniques for implanting the boron into silicide layers formed above the source/drain regions and then diffusing the boron into the substrate from these silicide layers.

Several techniques to limit the redistribution of boron after implant have also been suggested. The first involves use of rapid thermal processing instead of a conventional furnace treatment following the implant. Relatively high temperatures are employed to realize adequate dopant activation, while keeping the times short enough so that boron redistribution is minimized. The second makes use of lower temperature flow and reflow steps by employing BPSG as the interlevel dielectric material instead of PSG. The third modifies the process sequence so that the boron implant is done after the arsenic implant. If the spacers required for NMOS LDD formation are in place, the boron implant must be sufficiently redistributed to ensure that the p -regions extend under the p -channel gates.¹⁰⁵ On the other hand, some process sequences have been reported in which the spacers are removed prior to implanting the boron. Other materials than SiO_2 must be used for such removable spacers (e.g., silicon nitride,¹⁰² TiN,¹⁰³ or polysilicon¹⁰⁴) since removing SiO_2 spacers would significantly thin the field oxide. Concerns about damaging the gate oxide at the gate edge have been raised if dry etching is used for such spacer removal. In the reports describing silicon nitride and TiN removable spacers, wet etching is used to remove them.

5.8.3.3 LDD Structures in PMOS to Limit Punchthrough. Another way to achieve a shallow junction in a PMOS device is to use an LDD structure. The junction depth of the lightly-doped LDD tip region is shallower than the heavily doped

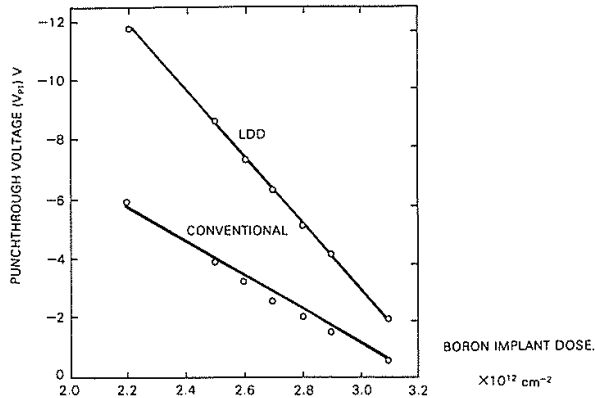


Fig. 5-70 Punchthrough voltage as a function of boron implant dose for a $0.4 \mu\text{m}$ PMOSFET with and without LDD structure.⁷⁵ (© 1986 IEEE).

p^+ source/drain junction depth implanted at the same energy. Figure 5-70 indicates that the punchthrough voltage V_{PT} is increased by 2-3V when an LDD is incorporated into a PMOS device. (The LDD structure, of course, may also be used to increase the hot-carrier resistance of the PMOS devices, especially as the channel length is decreased below about $0.8 \mu\text{m}$.) The LDD, however, will generally reduce I_D due to a rise in the series resistance of the drain region. Thus, a tradeoff analysis between an increase in V_{PT} and a decrease in the drain current must be made.

Adding an LDD structure to a PMOS device in CMOS involves relatively straightforward changes to the process sequence. The gate sidewall spacers are already available from the creation of NMOS LDDs, but there is some increase in process complexity, (especially if the former process used a maskless n^- implant to form the tip regions of the NMOS LDD). That is, to form both PMOS and NMOS LDDs four masking steps are generally needed (n^-, p^+, n^+, p^+), rather than just two. However, a two-mask process, which employs removable polysilicon spacers, has also been reported.¹⁰⁴

5.8.3.4 New PMOS Device Structures Designed to Suppress Punchthrough. New PMOS device structures have also been proposed to minimize punchthrough current. For example, a small segment of an n -type region can be created by ion implantation between the p -type S/D and the channel as a punchthrough stopper (refer to Fig. 5-71). This structure has been termed a double-implanted or *halo* structure. One report of such a halo device with $L = 0.5 \mu\text{m}$ used a phosphorus implant dose of $1 \times 10^{12} \text{ cm}^{-2}$ at 130 keV.⁷⁸ A sidewall spacer was used to form the lateral offset of the n -type region in this device. However, since the p^+ source/drain is offset from the gate by the sidewall spacer, this method of forming the local n -regions degrades the drive current and impairs hot-electron immunity.⁷⁹ Furthermore, as this technique also places a highly doped n -region adjacent to a highly doped p -region, the maximum halo doping

is determined by the onset of low-level avalanche at the chosen supply voltage and channel length.

A large-tilt-angle implant can be used to fabricate halo-type structures without a sidewall spacer, thereby gaining the benefits while avoiding the problems of the structure of Fig. 5-71.⁷⁹ The dopant is placed precisely between the p^+ source and the buried p -channel junction. The lateral penetration and depth of the n -type region are controlled by the implant energy and tilt-angle. The peak concentration is set by the dose. A typical implant sequence to form such a *large-angle tilt implanted punchthrough stopper* (LATIPS) structure as shown in Fig. 5-72 is: (1) implant phosphorus at 90 keV with a dose of $2 \times 10^{13} \text{ cm}^{-2}$ at 25° tilt, in two segments, with the target rotated 180° between segments; then (2) implant BF_2^+ (source/drain) at 40 keV (no tilt) to a dose of $3 \times 10^{15} \text{ cm}^{-2}$.⁸⁰

Devices fabricated using the LATIPS process were compared to conventional PMOSFETs. LATIPS devices showed dramatically improved punchthrough characteristics for physical gate lengths as small as 0.5 microns. As an example, the OFF-state leakage current was eight orders of magnitude smaller (at -3V drain bias) than that of a conventional control device fabricated without the LATIPS structure. This dramatic improvement in punchthrough resistance was achieved without extreme scaling of the source/drain junction depth or the n -well background doping. That is, the p^+ source/drain junction depth was about $0.2 \mu\text{m}$ and the n -well concentration was $1 \times 10^{16} \text{ cm}^{-3}$. Furthermore, the benefits were obtained without significant degradation of device performance. That is, no gate-oxide degradation was observed as a result of using BF_2^+ source/drain implants at 8 keV, even at doses of $5 \times 10^{15} \text{ cm}^{-2}$. Better high-current drive capability than conventional devices was also exhibited, as well as a smaller body (back-gate bias) effect.

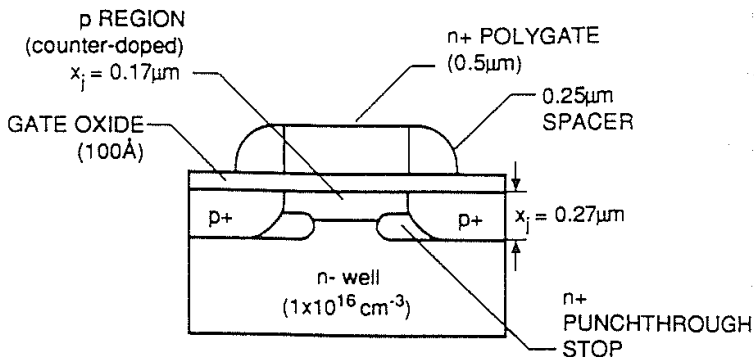


Fig. 5-71 PMOS Punchthrough-stop structure.⁷⁸ (© 1985 IEEE).

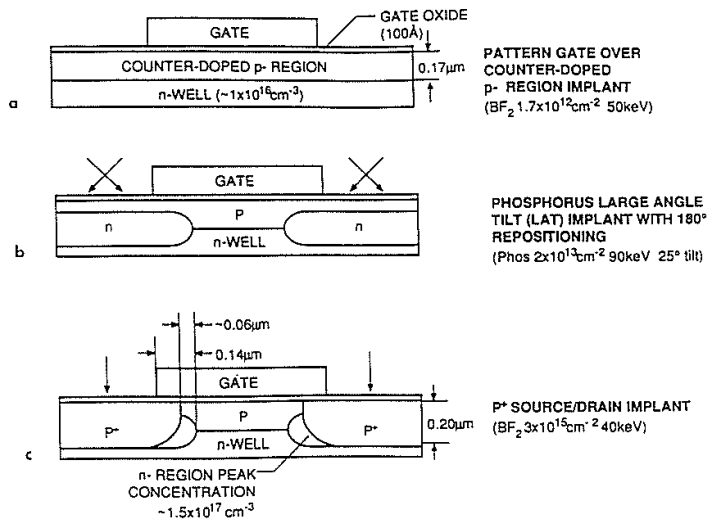


Fig. 5-72 LATIPS process sequence for PMOSFET: (a) Pattern gate over an n -well region that has been counterdoped with a p -type implant (BF₂); (b) Large angle tilt implant with phosphorus; (c) p^+ Source/Drain implant (BF₂).⁷⁹ (© 1988 IEEE).

5.8.4 PMOS Devices with p^+ -Polysilicon Gates

The use of a boron implant to adjust V_{Tp} becomes less feasible as devices use even thinner gate oxides, since larger doses of boron are needed. Y_j thus becomes deeper, and the punchthrough problem worsens. In addition, buried channel devices are subject to more process variability stemming from the difficulty of controlling the compensating implant. Solutions involving the use of gate electrodes other than n^+ polysilicon must therefore be explored. One alternative is to use p^+ polysilicon for PMOS devices. This approach appears to become necessary for CMOS and BiCMOS once L becomes smaller than 0.5 μm.

When p^+ polysilicon is used for the gate material, V_{Tp} is shifted from the values that occur when n^+ polysilicon is used. Figure 3-20 also plots V_{Tp} as a function of substrate doping when p^+ polysilicon is employed (using the scale on the right side of the figure). V_{Tp} can be seen to be less negative than -0.7 V over the substrate doping range of 10¹⁵-10¹⁷ cm⁻³. Thus, V_{Tp} can easily be made more negative through the implantation of phosphorus or arsenic into the channel. If p^+ polysilicon is used throughout the circuit, however, the NMOS device then has to be overcompensated in order for V_{Tn} to be reduced to sufficiently small values. (Note that some reports have been published on the use of p^+ polysilicon alone.^{81,82}) This implies it would be ideal

to use both n^+ and p^+ poly gates on the same chip (with n^+ poly for NMOS devices, and p^+ poly for PMOS devices).⁸³

Such a dual-doped poly approach, however, adds process complexity, and also introduces other problems. One drawback of it is that the sheet resistance of p^+ poly is $\sim 3\times$ higher than that of an n^+ poly film of the same thickness, because the mobility of holes is $\sim 3\times$ smaller than that of electrons in Si. A second problem arises from the need to connect the two types of poly (e.g., at the input of an inverter). Such problems occur when a silicide overlayer (or strap) is used to make the connection between n^+ and p^+ poly (as a method to avoid a separate, space-consuming metal contact). Because the silicide strap provides a very rapid diffusion path for boron and arsenic, one type of poly can be counterdoped by the other when the device is subjected to high-temperature excursions. This counterdoping can occur to the degree that a region of poly can change doping types (i.e., from n^+ to p^+). In such cases, the threshold voltage of devices with counterdoped poly will be shifted from their designed value.

If the processing temperature is limited to 800°C after the two types of poly have been connected by the silicide, such lateral diffusion does not produce significant shifts in V_T .⁸⁴ On the other hand, temperatures of 900°C will produce sufficient counterdoping to significantly shift V_T . Since one of the last high temperature steps in CMOS is activation of the source/drain implants, this would mean either using a lower-temperature activation step, or performing the step prior to formation of the silicide layer.

Another approach is to form a polysilicon electrode with an overlying conductor layer that suppresses such counterdoping. One such structure is a W-TiN-poly electrode structure.⁸⁵ The thin (30-nm) TiN film acts as a diffusion barrier to the dopants in the poly and also prevents reaction between the poly and the W. Very little of the dopant present in the poly is found to diffuse into the W, even following an anneal at 900°C for 1 hour.

Another problem encountered with p^+ poly gates when a thin gate oxide is used is poor V_T process control in the PMOS devices, due to penetration of the boron into the oxide (or further, into the silicon). It is reported that boron will penetrate gate oxides that are ≤ 12.5 nm thick during a 900°C 30-minute post-implant anneal in N_2 .⁸⁶ This implies that if p^+ poly gates are used, a lower processing temperature may be needed. If too low a temperature is used, however, the boron implanted into the polysilicon will not be sufficiently redistributed. The polysilicon dopant concentration at the polysilicon-gate oxide interface could thus end up being less than the mid- $10^{19}/\text{cm}^3$. This would make the work function of the polysilicon different from the desired degenerately doped value, creating V_T control problems in the MOS devices.

It has also been found that the presence of fluorine in the gate oxide worsens the boron penetration problem (Fig. 5-73).^{87,88} Such fluorine can be introduced into the gate oxide if the PMOS source drain regions are implanted using BF_2^+ . Elemental boron is therefore considered inherently superior to BF_2^+ as the implant species for surface-channel PMOS devices in CMOS technologies that use p -doped polysilicon. A

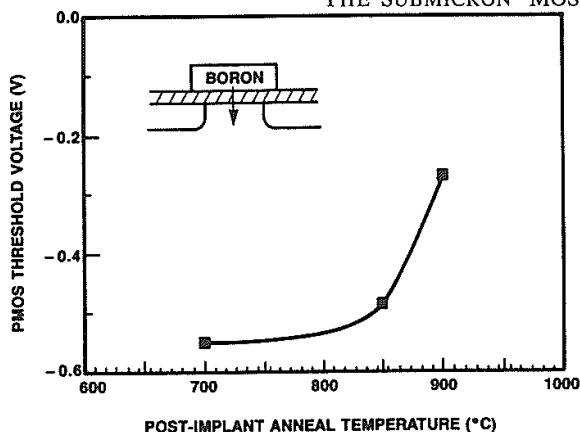


Fig. 5-73 Boron penetration through the gate oxide causes V_T in BF_2 implanted PMOSFETs to shift positive at anneal temperatures above $\sim 800^\circ\text{C}$.⁸⁷ (© IEEE 1989).

study of enhanced boron diffusion through thin SiO_2 layers in a wet oxygen atmosphere is also reported in reference 89.

It also appears that nitrided oxides and reoxidized nitrided oxides (see chap. 9) are effective in blocking boron diffusion.⁹⁰ For this, and other reasons, they are being extensively studied for use as candidates to replace SiO_2 as the gate insulators in dual-gate CMOS processes. An alternative approach to introducing nitrogen into the gate oxide to block boron penetration is through ion implantation. Recently, ion implantation of nitrogen into the gate p - and n -doped poly electrodes of a $0.25\text{ }\mu\text{m}$ CMOS process has been reported. The heavy nitrogen dose ($1 \times 10^{15}\text{ cm}^{-2}$) resulted in the gate oxide incorporating high concentrations of nitrogen. This had the effect of suppressing boron penetration of the gate oxide.¹¹⁰

Hot carrier effects in p^+ -poly $0.5\text{-}\mu\text{m}$ PMOSFETs have also been examined. Stress experiments showed significantly more electron trapping in short stress times than is observed in NMOSFETs. The device degradation due such trapping can be severe enough that it may pose a problem in such devices.¹¹⁸

REFERENCES

1. J.E. Chung et al., "Performance and Reliability Design Issues of Deep Submicron MOSFETs," *IEEE Trans. Electron Devices* **ED-38**, March 1991, p. 545.
2. J. Greenfield and R. Dutton, "Nonplanar VLSI Device Analysis Using the Solution of Poisson's Equation," *IEEE Trans. Electron Dev.*, **ED-27**, p. 1520 (1980).
3. L.D. Yau, "A Simple Theory to Predict the Threshold Voltage in Short-Channel IGFETs," *Solid-State Electronics*, **17**, p. 1059, (1974).

4. W. Fichtner and H.W. Poetzel, "MOS Modelling by Analytical Approximations - Subthreshold Current and Subthreshold Voltage," *Int. J. Electronics*, **46**, p. 33, Jan. 1979.
5. C.R. Viswanathan, B.C. Burkery, G. Lubberts, and T.J. Treadwell, *IEEE Trans. Electron Dev.*, **ED-32**, p. 932, (1985).
6. R.R. Troutman, "VLSI Limitations from Drain-Induced Barrier Lowering," *IEEE Trans. Electron Dev.*, **ED-26**, p. 461, (1979).
7. F.-C. Hsu et al., *IEEE Trans Electron Dev.*, **ED-30**, p. 571 (1983).
8. S.C. Jain and P. Balk, "A Unified Analytical Model for DIBL ... in Short-Channel MOSFETs," *Solid-State Electronics*, **30** (5), p. 503, May 1987.
9. T. Skotnicki et al., *IEEE Electron Dev. Letts.*, **EDL-9**, p. 109, March 1988.
10. G.W. Taylor, *Solid-State Electronics*, **22**, p. 701, (1979).
11. B.J. Sheu, D.L. Scharfetter, P.K. Ko, and M.-C. Jeng, "BSIM: Berkeley Short-Channel IGFET Model for MOS Transistors," *IEEE J. of Solid State Circuits*, August 1987, p. 558.
12. G. Merkel, *Process and Device Modeling for IC Design*, (F. Van de Wield et al., Eds.), Noordhoff, Leyden, (1977), p. 705.
13. L. A. Akers et al., "A Model of a Narrow-Width MOSFET including Tapered Oxide and Doping Encroachment," *IEEE Trans. Electron Dev.* **ED-28**, p. 1490, 1981.
14. C.R. Ji and C.T. Sah, "Two-Dimensional Analysis of the Narrow-Gate Effect in MOSFET," *IEEE Trans. Electron Dev.*, **ED-30**, P. 635, June 1983.
15. T. Furukawa and J.A. Mandelman, "Process and Device Simulation of Trench Isolation Corner Parasitic Device," *Ext Abs. Mtg. Electrochem. Soc.*, Fall 1988, p. 329.
16. L. A. Akers, "The Inverse Narrow-Width Effect," *IEEE Electron Dev. Letts.*, **EDL-7** (7), p. 419, July 1986.
17. E.H. Li et al., "The Narrow-Channel Effect in MOSFETs with Semi-Recessed Oxide Structures," *IEEE Trans. Electron Dev.*, **ED-37** (3) p. 692, March 1990.
18. S. S. Chung and T.-C. Li, "An Analytical Threshold-Voltage Model of Trench-Isolated MOS Devices with Nonuniformly Doped Substrates," *IEEE Trans. on Electron Dev.*, March 1992, p. 614.
19. K. Ohe et al., "Narrow-Width Effects of Shallow Trench-Isolated CMOS with n^+ -Polysilicon Gate," *IEEE Trans. Electron Dev.*, **ED-36** (6), p. 1110, June 1989.
20. G. Merkel, "A Simple Model of the Threshold Voltage of Short and Narrow Channel MOSFETs," *Solid-State Electronics.*, **23**, p. 1207, December 1983.
21. S. Selberherr et al., "MINIMOS - A Two-Dimensional MOS Transistor Analyzer," *IEEE J. Solid State Circuits*, **SC-15**, p. 605 (1980).
22. Z.-H. Liu et al., "Threshold Voltage Model for Deep-Submicrometer MOSFETs," *IEEE Trans. Electron Dev.*, **ED-40**, January 1993, p. 86.
23. J.R. Brews, W. Fichtner, E.H. Nicollian, and S.M. Sze, "Generalized Guide for MOSFET Miniaturization," *IEEE Electron Dev. Letts.*, **EDL-1**, p. 2, 1980.
24. C. Hu, "Future CMOS Scaling and Reliability," *Proceedings of the IEEE*, **81**, p.682, May 1993.
25. F.M. Klaasen, "Design and Performance of Micron-Sized Devices," *Solid-State Electron.*, **21**(3), p. 565 (1978).
26. C.F. Codella and S. Ogura, "Halo Doping Effects in Submicron DI-LDD Device Design," *Tech. Dig. IEDM*, p. 230, 1985.

27. A. Hori et al., "A Self-Aligned Pocket Implantation (SPI) Technology for 0.2 μm -Dual Gate CMOS," *Tech. Dig. IEDM*, p. 641, 1991.
28. H. Nihara et al., "Anomalous Drain Current in NMOS and its Suppression by Deep Ion Implantation," *Tech. Dig. IEDM*, p. 487 (1978).
29. L.C. Parrillo, "CMOS Active and Field Device Fabrication," *Semicond. International*, April 1988, p. 64.
30. T. Skotnicki, G. Merkel, and T. Pedron, "A New Punchthrough Current Model Based on the Voltage-Doping Transformation," *IEEE Trans. on Electron Dev.*, **ED-35**, July 1988, p. 1076.
31. C. Duvvury, "A Guide to Short-Channel Effects in MOSFETs," *IEEE Circuits and Systems Magazine*, November 1986, p. 6.
32. J. A. Cooper and D.F. Nelson, "Measurement of the high-field drift velocity of electrons in the inversion layer in silicon," *IEEE Electron Dev. Lett.*, **EDL-2**, July 1983, p. 169.
33. R.H. Crawford, *MOSFETs in Circuit Design*, McGraw-Hill, New York, 1967.
34. D. Frohman-Bentchkowsky, "On the effect of mobility in inversion and accumulation layers on thermally oxidized silicon surfaces," *Proc. IEEE*, **vol. 56**, p. 217, 1968.)
35. A.G. Sabnis and J.T. Clemens, "Characterization of Electron Velocity in the Inverted $\langle 100 \rangle$ Si Surface," *Tech. Dig. - IEDM*, p. 19, 1979.
36. M.S. Liang et al., "Inversion layer capacitance and mobility of very thin gate oxide MOSFETs," *IEEE Trans. Electron Dev.*, **ED-33**, March 1986, p. 409.
37. K.K. Hung, *Electrical characterization of the Si-SiO₂ interface for thin oxides*, Doctoral Thesis, Hong Kong University, (1987).
38. C.G. Sodini, T. Ekstedt, and J. Moll, "Charge Accumulation and Mobility in Thin Dielectric MOS Transistors," *Solid-State Electronics*, **25**, p. 833, (1982).
39. K. Lee et al., "Physical Understanding of Low Field Carrier Mobility in Silicon Inversion Layer," *IEEE Trans. Electron Devices*, **ED-38**, August 1991, p. 1905.
40. S.R. Hofstein and G. Warfield, "Carrier mobility and current saturation in the MOS transistor," *IEEE Trans. Electron Dev.*, **ED-12**, p. 129, (1965).
41. F.N. Trofimenkoff, "Field-dependent mobility analysis of the field-effect transistor," *Proc. IEEE*, **53**, p. 1765, (1965).
42. D.M. Caughey and R.E. Thomas, "Carrier mobilities in silicon empirically related to doping and field," *Proc. IEEE*, **55**, p. 2192 (1967).
43. C.G. Sodini, P.K. Ko, and J.L. Moll, "The effect of high fields on MOS device and circuit performance," *IEEE Trans. Electron Dev.*, **ED-31**, October 1984, p. 1386.
44. E.A. Talkhan, I.R. Mansour, and A.I. Barboor, "Investigation of the effect of drift-field-dependent mobility on MOSFET characteristics," Parts I and II. *IEEE Trans. Electron Dev.* **ED-19**, August 1972, p. 899.
45. P.K. Ko, "Approaches to Scaling," Ch. 1, in *Advanced MOS Device Physics*, N.G. Einspruch and G. Gildenblatt, Eds. Academic Press, Orlando, FL., 1989, p. 1
46. K.Y. Toh, P.K. Ko, and R.G. Meyer, "An engineering model for short-channel MOS devices," *IEEE J. Solid-State Circuits*, August 1988, p. 950.
47. M.C. Jeng et al., "Performance and reliability of deep-submicron MOSFETs," *Tech. Dig. IEDM*, 1987, p. 710.

48. V.G. Reddi and C.T. Sah, "Source to Drain Resistance Beyond Pinchoff in Metal-Oxide-Semiconductor Transistor (MOST)," *IEEE Trans. Electron Dev.*, **ED-12**, p. 108, (1969).
49. D. Frohman-Bentchkowsky and A.S. Grove, *IEEE Trans. Electron Devices* **ED-16**, p. 108, (1969).
50. G. Baum and H. Beneking, "Drift Velocity Saturation in MOS Transistors," *IEEE Trans. Electron Devices*, **ED-17**, June 1972, p. 481.
51. Y.A. Elmansy and A.R. Boothroyd, "A simple two-dimensional model for IGFET operation in the saturation region," *IEEE Trans. Electron Dev.*, **ED-24**(3), p. 254-62, (1977).
52. P.K. Ko, R.S. Muller, and C. Hu, "A Unified Model for Hot-Electron Currents in MOSFETs," *Tech. Dig. IEDM*, p. 600, 1981.
53. P.K. Ko, *Hot-electron effects in MOSFETs*, " Doctoral Thesis, University of California, Berkeley, (1982).
54. R.S. Muller and T.I. Kamins, *Device Electronics for Integrated Circuits*, John Wiley & Sons, New York, 1986, p. 491.
55. T.Y. Chan, P.K. Ko, and C. Hu, "Dependence of Channel Electrical Field on Device Scaling," *IEEE Electron Dev. Lett.*, **EDL-6**, October, 1985, p. 531.
56. B. J. Sheu et al., "BSIM: Berkeley Short-Channel IGFET Model for MOS Transistors," *IEEE J. Solid-State Circuits*, Aug. 1987, p. 558.
57. B. J. Sheu, D. L. Scharfetter, and H. C. Poon, "Compact short-channel IGFET Model," Univ. of Calif., Berkeley, CA Memo. UCB/ERL M84/20, March 1984.
58. M.-C. Jeng, *Design and Modeling of Deep-Submicron MOSFETs*, ERL Memo No. ERL M90/90, Electronics Research Laboratory, University of California, Berkeley, October 1990.
59. A.H.-C. Fung, "A Subthreshold Conduction Model for BSIM," Univ. of Calif., Berkeley, Memo, UCB/ERL M65/22, March 1985.
60. B. Eitan and D. Frohman-Bentchkowsky, "Surface Conduction in short-channel MOS Devices as a limitation to VLSI scaling," *IEEE Trans. Electron Dev.* vol. ED-29, p. 254, Feb. 1982.
61. R. H. Dennard, et al., "Design of II MOSFETs with Very Small Physical Dimensions," *IEEE J. Solid-State Circuits*, **SC-9** (5), 256, (1974).
62. G. Bacarani, M. R. Wordeman, and R. H. Dennard, "Generalized Scaling Theory and its Application to a 0.25 μm MOSFET Design," *IEEE Trans. Electron Dev.*, **ED-31**(4), 452 (1984).
63. J.R. Brews et al., "Generalized Guide to MOSFET Miniaturization," *IEEE Electron Dev. Letts.*, **EDL-1**, January 1980, p. 2.
64. F.M. Klaassen, "Design and Performance of Micron-Size Devices," *Solid-State Electronics*, **21**, March 1978, p. 565.
65. J.R. Brews, K.K. Ng, and R.K. Watts, in "The Submicron Silicon MOSFET," Ch. 1, p. 19, *Submicron Integrated Circuits*, Ed. R.K. Watts, Wiley Interscience, New York, 1989.
66. *ibid.* pp. 23- 37.
67. R.A. Haken et al., "BiCMOS Process Technology," Chap. 3, p. 85, in *BiCMOS Technology and Applications*, A. Alvarez Ed. Kiewer, Norwell MA, 1989.

68. A. Alvarez et al., "Application of Statistical Design and Response Surface Methods to Computer-Aided VLSI Device Design," *IEEE Trans. Computer-Aided Design, CAD-7*, February 1988, p. 272.
69. *ibid.* ref. 65, p. 24.
70. S.M. Sze, Ed. *VLSI Technology*, 2nd edition, chap. 11, McGraw-Hill, New York, 1988.
71. T. Ohzone et al., *IEEE Trans. Electron Dev.*, **ED-32**, 1789, (1980).
72. G.J. Hu et al., *Tech. Dig. IEDM*, 1982, p. 710.
73. K.M. Cham et al., *Computer-Aided Design and VLSI Device Development*, Boston, Kluwer Academic Publishers, 1986, p. 182.
74. J. Zhu, R.A. Martin, and J.Y. Chen,, "Punchthrough Current for Submicrometer MOSFETs in CMOS VLSI," *IEEE Trans. Electron Dev.*, February 1988, p. 145.
75. A. Schmitz and J. Y. Chen, "Design, Modelling, and Fabrication of Submicron CMOS Transistors," *IEEE Trans. Electron Dev.*, **ED-33**, p. 148, (1986).
76. L.C. Parrillo, "CMOS Active and Field Device Fabrication," *Semicond. Internat.*, April 1988, p. 64.
77. R. A. Chapman et al., "A 0.8 μm CMOS Technology," *Tech. Dig. IEDM*, 1987, p. 362.
78. S. Odanaka et al., "A New Half-Micron P-Channel MOSFET with Efficient Punchthrough Stops," *1985 Symp. VLSI Tech., Digest of Tech. Papers*, p. 62.
79. T. Hori et al., "A New P-Channel MOSFET with Large Tilt Angle Implanted Punch-through Stopper (LATIPS)," *IEEE Electron Dev. Letts.*, **EDL-9**, Dec. 1988, p. 641.
80. M. Stinson et al., "Effects of Ion Implantation on Deep-Submicrometer Drain-engineered MOSFET Technologies," *IEEE Trans. Electron Dev.*, **ED-38**, March 1991, p. 487.
81. L. C. Parrillo et al., *Tech Dig. IEDM*, 1984, p. 418.
82. K. M. Cham et al., *IEEE Electron Dev. Lett.*, January 1986, p. 49.
83. C.Y. Wong et al., *Tech. Dig. IEDM*, 1988, p. 238.
84. S. J. Hillenius et al., "A Symmetric Submicron CMOS Technology," *Tech. Dig. IEDM*, 1986, p. 252.
85. P.-H. Pan, J.G. Ryan, and M.A. Lavoie, *Ext. Abs. of the Electrochem. Soc. Meeting*, Spring 1989, Abs. 138, p. 193.
86. M.L. Chen et al., "Constraints in *p*-Channel Device Engineering for Submicron CMOS Technologies," *Tech. Dig. IEDM*, 1988, p. 390.
87. F.K. Baker, et al., *Tech. Dig. IEDM*, 1989, p. 443.
88. J.M. Sung, *Tech. Dig. IEDM*, 1989, p. 447.
89. Y. Sato, K. Ehara, and K. Saito, *J. Electrochem. Soc.*, June 1989, p. 1777.
90. H. Fang et al., "Low-Temperature Furnace-Grown Reoxidized Nitrided Gate Dielectrics as a Barrier to Boron Penetration," *IEEE Electron Dev. Letts.*, **EDL-13** (4) p. 217 April 1992.
91. M. Nishida and H. Onodera, "An anomalous increase of threshold voltages with shortening of the channel lengths fro deeply boron-implanted *n*-channel MOSFETs," *IEEE Trans. Electron Dev.*, **ED-28**, p. 1101, 1981.
92. M. Orlowski, C. Mazure, and F. Lau, "Submicron short-channel effects due to gate re-oxidation induced lateral diffusion," *IEDM Tech. Dig.* 1987, p. 632.
93. C.Y. Lu and J.M. Sung, "Reverse Short-Channel Effects on Threshold Voltage in Submicron Salicide Devices," *IEEE Electron Dev. Lett.*, **EDL-10**, p. 446, October, 1989.

94. N.D. Arora and M.S. Sharma, "Modeling the Anomalous Threshold Voltage Behavior of Submicrometer MOSFETs," *IEEE Electron Dev. Lett.*, **EDL-13**, p. 92, February 1992.
95. H. Hanafi et al., "A Model for Anomalous Short-Channel Behavior in MOSFETs," *IEEE Electron Dev. Lett.*, **EDL-14**, p. 575, December 1993.
96. D. Sadana et al., "Enhanced Short-Channel Effects in NMOSFETs due to Boron Redistribution Induced by Arsenic Source and Drain Implant," *IEDM Tech. Dig.* 1992, p. 849.
97. H. Jacobs et al., "MOSFET Reverse Short Channel Effect Due to Silicon Interstitial Capture in Gate Oxide," *IEDM Tech. Dig.* 1993, p. 307.
98. C.S. Rafferty et al., "Explanation of Reverse Short Channel Effect by Defect Gradients," *IEDM Tech. Dig.* 1993, p. 311.
99. M.D. Giles, *Appl. Phys. Lett.*, **62**(2), p. 1940, (1993).
100. J.Y. Chen, "An n -well CMOS with self-aligned channel stops," *Tech. Dig. IEDM* 1983, p. 526.
101. T.M. Liu and W.G. Oldham, *IEEE Electron Dev. Letts.*, **EDL-5**, 299 (1984).
102. J.R. Pfister et al., "A Poly-Framed LDD Sub-half-Micrometer CMOS Technology," *IEEE Electron Dev. Letts.* **EDL-11**, November 1990, p. 529.
103. J.R. Pfister et al., "An Integrated 0.5 μ m CMOS Disposable TiN LDD/Salicide Spacer Technology," *Tech. Dig. IEDM*, 1989, p. 781.
104. L.C. Parrillo, "CMOS Active and Field Device Fabrication," *Semiconductor International*, April 1988, p. 64.
105. S. Meguro et al., *Tech. Dig. IEDM*, 1984, p. 59.
106. J.S.T. Huang, J.W. Schrankler, and J.S. Kueng, *IEEE Trans. Electron Dev.*, **ED-31**, 1889, (1984).
107. S. Ogura et al., *IEEE Trans. Electron Dev.*, **ED-27**, 1359 (1980).
108. Y. Omura and K. Ohwada, *Solid-State Electronics*, **24**, 301 (1981).
109. N. Kotani and S. Kawazu, *Solid-State Electronics*, **22**, p. 63, (1979).
110. T. Kuroi et al., "Novel NICE (Nitrogen Implantation into CMOS Gate Electrode and Source-Drain) Structure for High Reliability and High Performance 0.25 μ m Dual Gate CMOS," *Tech. Dig. IEDM*, 1993, p. 325.
111. R.R. Troutman, *IEEE J. Solid-State Circuits*, **SC-9**, p. 55 (April 1974).
112. J.Y. Chen et al., "A Fully Recessed Field Isolation Technology Using Photo-CVD Oxide," *Tech. Dig. IEDM*, 1982, p. 233.
113. T. Shibata et al., "An optimally designed process for submicrometer MOSFETs," *IEEE Trans. Electron Dev.* **ED-29**, p. 531, (1982).
114. G.J. Hu and R.H. Bruce, "Design Tradeoffs Between Surface and Buried-Channel FETs," *IEEE Trans. Electron Dev.*, **ED-32**, March 1985, p. 584.
115. K.M. Cham and S.Y. Chiang, "Device design for the Submicrometer p -channel FET with n^+ polysilicon gate," *IEEE Trans. Electron Dev.*, **ED-31**, p. 964 (1984).
116. K.M. Cham et al., "Characterization and modeling of the trench surface inversion problem for the trench isolated CMOS technology," in *IEDM Tech. Dig.*, p. 23, 1983.
117. T. Skotnicki, G. Merckel, and T. Pedron, "Analytical Study of Punchthrough in Buried-Channel PMOSFETs," *IEEE Trans. Electron Dev.*, April 1989, p. 690.
118. Y. Hiruta et al., "Impact of Hot Electron Trapping of Half-Micron PMOSFETs with p^+ Poly-Si Gate," *Tech. Dig. IEDM*, 1986, p. 718.

119. J. Cooper and D. Nelson, "High-Field Drift Velocity of Electrons at the Si-SiO₂ Interface as Determined by the Time-of-Flight Technique," *J. Appl. Phys.* **54**, 1445 (1983).
120. T.E. Hendrikson, "A Simplified Model for Subpinchoff Condition in Depletion Mode IGFETs," *IEEE Trans. Electron Dev.*, **ED-25**, p. 435 (1978).

PROBLEMS

1. Calculate the threshold voltage (a) in 1 Ω -cm p -type silicon and (b) in 1 Ω -cm p -type silicon. The MOS devices for each case have : (i) n^+ -polysilicon gate; (ii) $t_{ox} = 100$ nm; and (iii) the oxide is free of charge except for a surface density of $Q_f = 5 \times 10^{10}$ cm⁻².
2. For exercise 1, what are the threshold voltages if p^+ polysilicon is used for both the n - and p -channel devices.
3. Describe punchthrough current and subthreshold current and explain the difference between them.
4. An n -channel MOSFET has a t_{ox} of 20 nm, an $L_{eff} = 1$ μ m, $r_j = 0.2$ μ m, and $V_T = 0.6$ V. If the device is biased at $V_{GS} = 3$ V and $V_{DS} = 5$ V, calculate the saturation voltage V_{DSSat} and the maximum electric field.
5. Use one-dimensional junction theory to estimate the punchthrough voltage of a MOSFET with a channel length of 1 μ m. Assume a substrate doping of 3×10^{16} /cm³, a substrate bias of 0 V, and a uniformly doped substrate (no V_T -adjust implant).
6. Using Eq. 5-97, calculate the buried-channel S_t value for $x_1 = 500$ Å, $t_{ox} = 250$ Å, $x_2 < x_3 = 3000$ Å and compare it with the S_t value calculated for a surface-channel MOSFET with the same t_{ox} and a channel depletion width of $d = x_3$.
7. How does a MOSFET's threshold change when constant field scaling rules are applied? Does it scale proportionately?
8. Consider a long-channel MOSFET with $L = 3$ μ m, $Z = 21$ μ m, $N_A = 5 \times 10^{15}$ cm⁻³, $C_{ox} = 1.5 \times 10^{-7}$ F/cm², and $V_T = 1.5$ V. Calculate V_{DSSat} for the case when $V_{GS} = 4$ V. If a constant scaling factor is used to reduce the channel length to 1 μ m, find the following scaled-down parameters: Z , C_{ox} , and I_{Dsat} .
9. A MOSFET with gate oxide thickness of 20 nm is to be fabricated on a p -type substrate of doping $N_A = 10^{16}$ cm⁻³. The source/drain regions have a depth of 0.2 μ m and $N_D = 10^{19}$ cm⁻³. The device is to be operated with $V_{BS} = 0$ V, but V_{DS} could be as large as 3 V. From Eq. 5-84 estimate the minimum channel length this MOSFET can have while still exhibiting electrically long-channel behavior. (The gate length is the channel length plus twice the junction depth.) A smaller value of minimum channel length could be obtained if a larger N_A was used. Suggest some undesirable side effects that such a choice could produce.
10. Qualitatively describe the differences between a surface-channel and a buried-channel MOSFET.
11. We have stated in this chapter that the saturation of electron velocity can limit the drain current in a MOSFET with a channel length of 1 μ m or less operating from a 5-V supply. Examine this effect by using SPICE to generate the I_D - V_{DS} characteristic of an NMOSFET with $L = 1$ μ m formed on a substrate of doping $N_A = 10^{16}$ cm⁻³ with a 50 nm thick gate oxide. Set $V_{GS} = 5$ V and sweep V_{DS} from 0 to 5 V using the Level 2 model. First run a simulation in which V_{MAX} is not specified, and then repeat the simulation with $V_{MAX} = 8 \times 10^6$ cm sec⁻¹, which corresponds roughly to the scatter-limited velocity of electrons in a MOSFET channel.

320 SILICON PROCESSING FOR THE VLSI ERA - VOLUME III

12. Qualitatively summarize why a surface channel NMOSFET is prone to suffering punchthrough only after the punchthrough voltage is exceeded, but that a buried-channel PMOSFET exhibits punchthrough current behavior even when drain voltages are very small.
13. Why does the subthreshold slope of the I_D - V_{GS} curve appear to increase at the onset of subsurface punchthrough?

CHAPTER 8

WELL FORMATION IN CMOS

Both n - and p -channel transistors must be fabricated on the same wafer in CMOS technologies. Obviously, only one type of device can be fabricated on a given starting substrate. To accommodate the device type that cannot be built on this substrate, regions of a doping type opposite to that present in the starting material must be formed. These regions are called *wells* (or sometimes *tubs*) and are usually (but not always) the first features to be defined on a starting wafer. The well-formation process starts with a uniform, lightly doped p - or n -type substrate. Excess dopants of the opposite type are then selectively implanted or diffused to attain the proper well depth and doping profile in the well regions. The doping type in the wells becomes the identifying characteristic of the CMOS technology (e.g., p -well or n -well CMOS). Conventional well-implant doses range from 2×10^{12} – $8 \times 10^{12}/\text{cm}^2$ at energies from 100–200 keV. The term "conventional" in this sense refers to wells formed by low-energy (<200 keV) ion implantation and a thermal drive-in. Alternative well-formation techniques that do not require a drive-in will also be discussed separately. Epitaxial n -on- n^+ or p -on- p^+ starting substrates have also come into widespread use, since they help reduce the susceptibility of CMOS circuits to latchup.

As well as being of the correct doping type, the well and substrate must be electrically isolated from each other under every operating condition. As noted in chap. 6, to ensure that pn junctions are not forward-biased during circuit operation, if a p -type substrate is used as the starting material of an IC, the substrate must be connected to the most negative circuit voltage. If an n -type substrate is used, it must be connected to the most positive circuit voltage. When wells are present, however, the well regions must also be connected to the appropriate circuit voltages to prevent forward-biasing of the junctions within the well (and between the well and substrate). Because the wells are totally junction-isolated from the rest of the wafer, it is especially important that provisions for well connections be made at the top surface of the wafer. That is, in non-well technologies, it may still be possible to contact the substrate from the backside of the wafer even if no provision is made for a substrate contact from the top surface. Obviously, such backside connections cannot be established to the well regions.

The dopant concentration near the surface of the well is one of the factors that impact V_T , carrier mobility, source/drain capacitance, and field isolation. The well concentration deeper beneath the surface affects other characteristics of MOSFET behavior such as latch-up and vertical punchthrough (i.e., between drain and well-substrate junctions).*

* To ensure that vertical punchthrough will not occur, the depletion regions of the drain/well and well/substrate junctions must be prevented from merging. That is, the depth of the well/substrate junction must be deep enough (typically $>1.5 \mu\text{m}$) so that it is greater than the sum of the widths of the two depletion regions in the well, even when a reverse bias of V_{DD} is

8.1 *p*-WELL, *n*-WELL, and TWIN-WELL TECHNOLOGIES

There are many trade-offs involved in the optimization of a CMOS process. The choices revolve around the highly interrelated parameters of circuit performance, layout density, fabrication cost, and susceptibility to latchup. As described in previous chapters, obtaining the best circuit performance from an MOS device involves maximizing the drive current and minimizing junction capacitances and body effect - all characteristics which favor lower doping concentrations in the device body. Optimizing density, however, requires raising these same doping concentrations (to avoid punchthrough and to achieve high values of threshold voltage in the parasitic field-effect transistors). Higher density is thus achieved by allowing closer packing of adjacent *n*- and *p*-channel transistors. Conversely, latchup tolerance can also be improved by keeping *n*- and *p*-channel transistors farther apart (see chap. 6, Vol. 2), but this in turn lowers circuit density.

These complex interacting tradeoffs converge on several processing configurations that are determined at the outset of the processing sequence through the selection of the type of well doping.^{1,2} It is therefore useful to consider the ramifications of the well-type choice with respect to its impact on circuit applications. Because at least one of the device types must be located in the well, the selection of well architecture influences device and circuit behavior. In particular, the performance of devices in the well for the most part suffer as a consequence of the higher well doping. That is, they exhibit higher junction capacitance, increased sensitivity to body effect, and decreased transconductance (due to reduced carrier mobility). Furthermore, substrate current (e.g., due to hot-carrier effects) is harder to collect from the well regions. The issue thus becomes one of deciding which device type (if any) should be subjected to such performance degradation.

Figure 8-1 shows the five basic well structures employed in CMOS technologies. Figure 8-1a illustrates the *single p-well structure*, in which *p*-wells are fabricated in an *n*-type starting substrate. NMOS devices are made in the well, and PMOS transistors in the substrate. If the dopant types of the *p*-well technology are reversed, a *single n-well structure* is obtained (Fig. 8-1b). The structures in Figs. 8-1c to 8-1e involve approaches in which *both p*- and *n*-wells are formed. CMOS technologies employing such structures are called *twin-well*, *dual-well*, or *twin-tub CMOS*. The first two of these (Figs. 8-1c and 8-1d) are designed to be fabricated on a particular substrate type, like the single *p*- and *n*-well structures. The third of these twin-well technologies (Fig. 8-1e), however, is independent of starting material type.

In the next sections we will outline the pros and cons of both the single *p*-well and *n*-well configurations, as well as those of the more complex well configurations that have been developed (e.g., twin-well and retrograde-well CMOS).

* (cont.) applied to the drain (see example 8-1 in this chapter.). In conventional-well CMOS structures, this depth is achieved by a long, high-temperature drive-in (up to 1200°C) following the well implants. For example, in one twin-well process the well-implants are followed by a drive-in at 1100°C for 500 minutes (8 1/3 hours).¹³

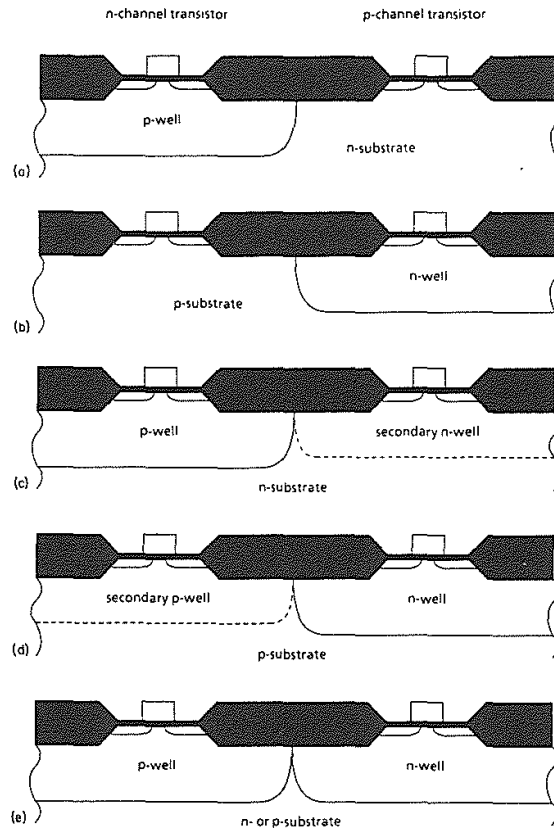


Fig. 8-1 Well structures available for CMOS circuit fabrication: (a) *p*-well CMOS; (b) *n*-well CMOS; (c) *p*-well with secondary *n*-well (twin-well constrained to *n*-substrate); (d) *n*-well with secondary *p*-well (twin-well constrained to *p*-substrate); (e) twin-well CMOS on *n* or *p*-substrate. A.G. Lewis and J.Y. Chen, "Current Trends in MOS Process Integration," in N.G. Einspruch and G. Gildenblatt, Eds. *Advanced MOS Device Physics*, Academic Press, San Diego CA, 1989, Reprinted with Permission.

8.1.1 *p*-Well CMOS

The *p*-well process, illustrated in Fig. 8-1a, involves the creation of *p*-regions in an *n*-type substrate to permit the fabrication of NMOS devices. The *p*-wells are formed by implanting a *p*-type dopant into an *n*-substrate at a high enough concentration to overcompensate for the substrate doping and to give adequate control over the *p*-type

doping in the well. The starting n -type substrate, however, must also have sufficient doping to ensure that the characteristics of devices fabricated in the substrate regions are adequate (a minimum doping concentration of 3×10^{14} - $1 \times 10^{15}/\text{cm}^3$ is required). If the p -well is doped too highly, however, the performance of the n -channel devices will be degraded through lower carrier mobility, increased source/drain to p -well capacitance, and increased sensitivity to body-biasing effects. The p -well doping is therefore usually selected to be about five to ten times higher than the doping in the n -substrate.

p -well CMOS was the first type of CMOS that could be practically manufactured, and such circuits were introduced in the late 1960s. The first companies to commercially offer CMOS components produced many designs in p -well technology. As this experience was spread throughout the industry, p -well CMOS became widely established.

There were several reasons for choosing p -well CMOS at that time. First, the earliest MOS ICs were based on a metal-gate PMOS process which utilizes an n -substrate. The work-function of aluminum is such that enhancement-mode PMOS devices can be made in a lightly doped n -type substrate without the need for additional channel doping. Thus, when CMOS circuits were invented (to take advantage of the reduced power consumption possible with CMOS), it was simplest to manufacture them merely by adding a p -well to the basic PMOS process. Second, higher channel doping levels were needed to permit sufficiently large V_T values to be established in the enhancement-mode NMOS transistors. Such higher doping was also needed in the field regions of the NMOS sections of the circuit to ensure adequate isolation (see chap. 6). These requirements are compatible with the need to have well surface concentrations five to ten times larger than the substrate doping level. Finally, it was relatively simple to manufacture p -wells with the existing process fabrication technology. That is, in the late 1960s ion implantation was not yet available and it was necessary to introduce the well dopants by chemical diffusion. At the doping levels required in the wells it was much easier to achieve adequate doping profile control using boron rather than phosphorus diffusion processes.

There are also several circuit and system advantages of p -well over n -well CMOS. First, p -well technology may be the better choice for pure-static logic, in which a good balance between the performance of both MOS device types is beneficial. (The performance of the PMOS and NMOS devices are "more balanced" because the carrier mobility reduction associated with higher well doping is compensated by forming the NMOS devices in the p -well; i.e., the electrons have higher mobility.) Second, it is attractive for applications that require an isolated p -region (such as those using an $n\text{pn}$ bipolar transistor as an on-chip driver or an n -channel FETs for an analog input). Third, p -well CMOS is less susceptible to field-inversion problems than n -well CMOS (which uses a lightly doped p -substrate), and can thus be slightly easier to fabricate. (We will describe later how p -well CMOS can use the inherent substrate and well doping as the channel stops, whereas n -well CMOS must use a separate substrate channel-stop process.) Fourth, if the so-called retrograde-well process (rather than a diffusion of a shallow implant) is used to form the wells, p -well technology is more feasible. It is easier to form a p -retrograde well than an n -retrograde well, since boron ions penetrate deeper than arsenic or phosphorus ions at the same implant energy. Finally, p -well

CMOS may be better for fabricating SRAMs. Since the alpha-particle-induced soft-error rate (SER) becomes significant even in SRAMs if feature sizes are scaled to submicron dimensions, the cells should be made inside a well. The sensing of the state of an SRAM cell depends on the current provided by the cell. As a result, high-gain NMOS devices are more desirable for the pass gates and drivers in the cell, and must be built in a p -well. Furthermore, it facilitates the use of polysilicon load resistors in 4-transistor memory cells and buried contacts.

Some of the disadvantages of p -well CMOS are the following: (1) Since NMOS transistors are fabricated in the p -well, their characteristics will be degraded compared to those made in a lightly doped substrate. In CMOS circuits that rely heavily on optimum NMOSFET behavior, this can lead to reduced circuit performance; (2) Substrate currents, the result of impact-ionization generated carriers in short-channel devices (see chaps. 7 and 9), are much larger in NMOS than PMOS devices. It is more difficult to collect these carriers from a well than from the substrate. Since the NMOSFETs in p -well CMOS are in the well, this represents a disadvantage of submicron CMOS p -well technology.

8.1.2 n -Well CMOS

By the late 1970s, the technological barriers that prevented enhancement-mode NMOS transistors from being manufactured were overcome. Primarily, the introduction of ion implantation made V_T control in NMOSFETs a simple, straightforward procedure. As a result, NMOS technologies ousted PMOS in high-speed, high-packing-density ICs. By the mid-1980s, however, the power dissipation in VLSI NMOS ICs became intolerably high (see chap. 6, Vol. 2), and CMOS development was pursued as a mainstream replacement technology for NMOS. For those companies that had been producing NMOS as their main IC product, n -well CMOS appeared to be an attractive approach for making the switch to CMOS.

In the n -well process, shown in Fig. 8-1b, the p -channel devices are formed in the more heavily doped n -well, while the NMOS devices are made in the substrate. Since it was now feasible to build NMOS devices in lightly doped substrates, this made it possible to make n -channel devices in an n -well CMOS process that were identical to those made by NMOS technology. Thus, virtually all of the experience that had been amassed in fabricating high-performance NMOS could be transferred to an n -well CMOS process. Furthermore, CMOS circuits could be made that exploited existing NMOS device design. The availability of ion implantation for introducing well dopants into the substrate (instead of having to rely on diffusion), now made dopant control of n -wells much more feasible. This led to the successful development of n -well CMOS technology.

As a result, virtually all EPROMs, microprocessors, and dynamic RAM designs in the generations of technology built with 1.25-2.0 μm dimensions were implemented with n -well CMOS. Mixed CMOS-NMOS designs were also made, in which the high noise immunity and negligible stand-by-power dissipation of CMOS was exploited in the input/output sections, and faster NMOS devices were used internally.

However, it should be noted that n -well technology also has some disadvantages. First, as mentioned earlier, it is more sensitive to field-inversion problems than p -well CMOS. Second, it may be more difficult to build pure-static, high-performance logic circuits with n -well CMOS.

8.1.3 CMOS on Epitaxial Substrates

As described in vol. 2, chap. 6, heavily doped substrates with a more lightly doped surface epitaxial layer have been utilized to suppress latchup in CMOS.⁴ When such starting material is used with single-well CMOS, the epitaxial layer is doped to a concentration equal to that of the substrate in a nonepitaxial wafer used for that process. If a twin-well CMOS process is used, the epitaxial layer is doped to a level significantly lower than that required for building either the p - or n -channel MOSFETs (see section 8.1.4).

The epitaxial layer is made thicker than the well depth, since the dopants in the heavily doped substrate under the epi layer diffuse toward the surface as the well dopants are diffused toward the bulk (Fig. 8-2). Thus, some of the epitaxial layer becomes more heavily doped during the CMOS process flow. The process is designed so that the bottom of the well is eventually adjacent to the heavily doped substrate region.

Either n -epi on n^+ substrates or p -epi on p^+ substrates can be used, with each method having advantages and drawbacks. Because the problems with n -epi on n^+ are more serious, p -epi on p^+ is more widely used. The major limitation of the latter

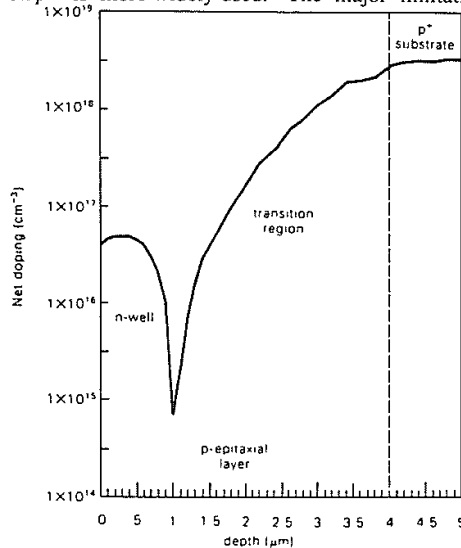


Fig. 8-2 A p -on- p^+ -epi doping profile and diffused n -well profile measured after the entire CMOS process has been completed. The as grown epi thickness is 4 μm and the final lightly doped epi thickness is about 2 μm .³ (© 1987 IEEE).

approach is that the outdiffusion of boron from the p^+ is much more severe than is the outdiffusion of As or Sb in n -epi on n^+ . (The reason for this is that boron diffuses much more rapidly than antimony, which is the most widely-used n^+ dopant material.) Thus, a thicker p^- -epitaxial layer must be used.

In addition, the transition region between the p^+ substrate and the p^- -epi layer is thicker, producing a larger series resistance (R_{sub}), which in turn reduces latchup immunity. On the other hand, the p -on- p^+ material is less sensitive to process-induced defects, and the p -type substrate provides higher conductivity under NMOS devices. Such additional conductivity is desirable, since it reduces the voltage drops caused by the substrate currents (generated as a result of the hot-carrier effects in short-channel devices). It is especially important in the regions containing NMOS devices, since the hot-electron substrate current is much higher in such devices than in PMOS devices.

The n -on- n^+ epitaxial substrates also offer some advantages. First, SRAMs are often built on n -type substrates, because the p -well-to- n -substrate junction provides protection from radiation-induced discharging of the memory's n -type storage nodes in the p -well.⁵ Second, retrograde p -wells are easier to implement because their implantation energy requirements are much lower. Third, the n -to- n^+ transition region is smaller than that in p -on- p^+ substrates, and the smaller value of R_{sub} provides improved latchup protection.

The limitations of n -epi on n^+ involve the process by which the heavily doped substrate is grown. Antimony (Sb) is the n -type dopant used, both because its diffusion coefficient is so low and because it exhibits much less lateral autodoping than arsenic (the other slow-diffusing n -type dopant). The problem with Sb is that its segregation coefficient, k_0 , is very small (i.e., $k_{0,\text{Sb}} = 0.023$; see, Vol. 1, chap. 1), and a large quantity of Sb must therefore be put into the Si melt to ensure that a sufficiently heavily doped ingot will be produced. Furthermore, in even the most highly refined Sb there are high concentrations of unwanted heavy metals, which become incorporated into the growing silicon crystal. Finally, the oxygen content of the Sb-doped crystal is relatively small, due in part to the special growing conditions used when the Sb-doped ingot is pulled.^{6,7} As a result, intrinsic gettering techniques that would immobilize the metals in the substrate are not as effective as they are in p -on- p^+ epi. For these reasons, n -on- n^+ substrates are less frequently chosen when epi-CMOS is implemented.⁴

A problem that exists with both types of epitaxial substrates for CMOS is that the wells cannot be made too deep, since the lateral diffusion would then take up too much area. On the other hand, if the wells are too shallow, vertical punchthrough (between the substrate and the source/drain regions of devices in the well) will ensue.⁸ In this case, appreciable leakage current can flow along this vertical path. As noted, punchthrough can occur if the depletion regions of the p^+/n -well junction and the n -well/ p^+ -substrate junction merge. The problem is even more severe in the case of the heavily doped p^+ substrate, due to the high degree of boron outdiffusion. A major limitation is thus imposed on the minimum epitaxial-layer thickness.

Another problem of epi-CMOS is back-surface autodoping. For example p -type 20 $\Omega\text{-cm}$ (7×10^{14} boron/cm³) epi on a 0.005 $\Omega\text{-cm}$ (2×10^{19} boron/cm³) substrate is representative of epi for CMOS devices. If the back surface of the substrate is not

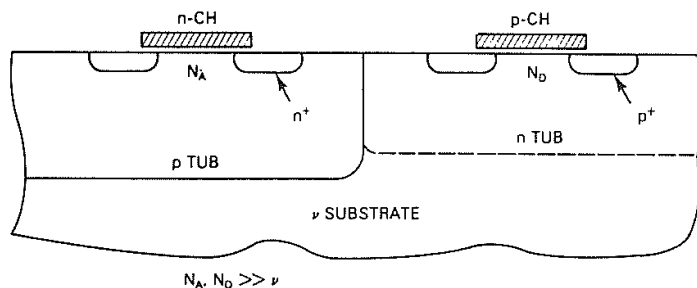


Fig. 8-3 Cross section of a twin-well 1.2- μm CMOS structure.⁵ Reprinted with permission of Semiconductor International.

sealed (e.g., by using a sealing layer such as undoped silicon dioxide or silicon nitride), boron evaporation from the substrate can contribute to autodoping on the front surface during the entire epi deposition cycle. This widens the epi/substrate interface and may even prevent the epi from reaching the 20 $\Omega\text{-cm}$ specified resistivity. Note that if silicon nitride is used as a sealing layer it should only be used in thin layers (e.g., less than 100 nm thick) since its high intrinsic stress causes it to bow the silicon.⁹

8.1.4 Twin-Well CMOS

With the twin-tub approach, two separate wells are formed for n - and p -channel transistors on a lightly doped substrate (Figs. 8-1c and 8-3). The substrate may be either a lightly doped wafer of n or p material, or a thin, lightly doped epitaxial layer on a heavily doped substrate. In either case, the level of the well surface doping is significantly lower than that required for building either the p - or n -channel MOSFETs. Each of the well dopants is implanted separately into the lightly doped surface region and is then driven in to the desired depth.

In twin-well CMOS, the doping profiles of each of the device types can be set independently, since the constraint of single-well CMOS does not exist (i.e., that the well doping must always be higher than the doping of the substrate in which one type of device is made). This was originally cited as an advantage of twin-well CMOS over single-well CMOS, with the argument made that both device types could thus be optimized.¹² This benefit for 1-2 μm CMOS has been questioned by Chen,² who points out that in modern single-well CMOS processes, an additional implant is used to prevent punchthrough without the need to raise the entire substrate-doping concentration. By incorporating this additional implant, it is possible to build higher-performance devices than can be achieved with the twin-well approach, in terms of junction capacitance and sensitivity to body effect.

Twin-well CMOS does offer some significant benefits for devices with submicron-channel lengths (although these advantages are gained at the cost of greater process

complexity). The first, and most important advantage arises when devices with submicron channel lengths are fabricated. Since it is recognized that the two device types perform similarly as channel lengths approach $0.5\ \mu\text{m}$, it is useful to provide symmetrical n - and p -channel devices. Furthermore, at submicron dimensions the body doping of both transistor types must be raised significantly to prevent punchthrough and to maintain adequate threshold-voltage levels. Thus, the advantage of having one type of MOS transistor in a lightly doped region (to optimize its performance at the expense of the other) disappears. It is instead more beneficial to produce two types of active device wells, each with its own optimized doping profile (i.e., formed by separate implants into a lightly doped substrate).*

The second advantage of the twin-well process is that it is compatible with the technologies of either isolation by selective epitaxial growth (SEG) or trench isolation.¹ Both approaches restrict the lateral diffusion of the dopants in each of the wells. In addition, sidewall inversion along the trench is less likely when both device sidewalls are butted against a trench that has been formed in a more highly doped well.¹⁰ Finally, when deep trenches are used with a thin epitaxial layer on a heavily doped substrate, latchup can be eliminated. The combined use of the twin-well process and advanced isolation techniques allows n^+ to p^+ spacing to be dramatically reduced in comparison to single-well technologies.

A third benefit is that the combination of epitaxial substrates and the twin-well process allows either substrate type to be chosen with no effects on transistor performance and essentially no change in process flow. Such flexibility is useful since some applications are best met with n^+ substrates and others with p^+ substrates. This advantage may be important if a single process is needed for implementing designs with different applications.

Finally, self-aligned channel stops can be easily implemented in the twin-well approach, allowing the spacing between n - and p -channel devices to be reduced. Although this spacing reduction is not as great as it would be if trenches or SEG were used, the process is much less complex with the twin-well method.

A twin-retrograde-well $0.7\text{-}\mu\text{m}$ -CMOS process for fabricating 1-Mbit SRAMs has been reported.¹¹ The high energy implants also allow a restricted thermal budget to be used, thus reducing the up-diffusion from the p^+ substrate. This permits the use of a thinner epi layer (which also helps prevent latch-up), and allows implantation of the channel-stop dopants after the field oxide has been grown and planarized (consequently, minimizing lateral diffusion of the boron channel-stop dopants). We discuss retrograde-well CMOS in more detail in section 8.3.

* It is not useful to start with a substrate doped to the optimum level needed for just one of the submicron devices. If this were done, a single well of much higher doping would have to be established for the other type of submicron device, which would unnecessarily degrade the device performance. It is possible, with the twin-well approach, to have both types placed in a well of optimum doping profile.

8.1.5 Summary of CMOS Well-Technology Issues

The selection of a particular CMOS process depends largely upon circuit applications, and to a lesser degree on technology evolution. For VLSI circuits with 1-2 μm design rules *n*-well CMOS has been the more widely used single-well technology, both because *n*-channel devices can be optimized in the *p*-substrate and because most circuits use more NMOS than PMOS devices. In circuits that use transistor channel lengths smaller than 1 μm , twin-well and retrograde-well technology are more attractive. As noted, the price paid for their benefits is increased process complexity.

The optimum process-integration design strategy is to first select the well technology (and to decide whether or not to use epitaxial substrates), and then to select the isolation method. Once these decisions have been made, the well depth and doping profile can be determined. The well depth impacts the lateral-diffusion distance of the well and the vertical-punchthrough voltage. The doping profile affects transconductance, threshold voltage, source/drain punchthrough, junction capacitance, carrier mobility, source/drain-to-substrate breakdown, sensitivity to body effect, and hot-electron effects.

As previously discussed, early CMOS used lightly doped *n*-type wafers and diffused *p*-wells (in which the NMOSFETs were fabricated). The *p*-well approach was initially adopted because NMOSFETs could only be reliably produced in more heavily doped regions until ion implantation was developed in the mid 1970's (see chaps. 5 and 6, Vol. 2). Eventually, however, *n*-well as well as *p*-well single-well CMOS processes were implemented. Later twin-well processes were introduced, and these now dominate the newest CMOS fabrication lines (although older *n*- and *p*-well single-well technologies are still used to fabricate a variety of CMOS products).

The well-formation process was also one that quickly exploited the technology of ion implantation. The doping precision required in the well gave implantation a significant edge over diffusion as a method for introducing the dopants into the substrate. Advances in implantation technology are allowing development of processes involving multiple implants and higher energies to avoid the long, high-temperature drive-in and to form retrograde well profiles (as will be discussed in sections 8.3 and 8.4).

8.2 CONVENTIONAL FORMATION OF WELLS AND CHANNEL STOPS

There are several issues associated with conventional-well and channel-stop formation that are covered in this section, including: (1) alignment of subsequent patterns to the well locations; (2) process sequences for forming twin-well and channel-stop regions, and; (3) novel well-formation schemes that reduce mask count and process complexity.

8.2.1 Aligning Subsequent Patterns to the Wells

The wells are often the first features to be formed in the CMOS process flow. However, the implant process that introduces the well dopants does not produce visible features on the wafer surface. Thus, it is necessary to provide some indirect indication of the well locations so that subsequent layers can be aligned to them. Here we illustrate some

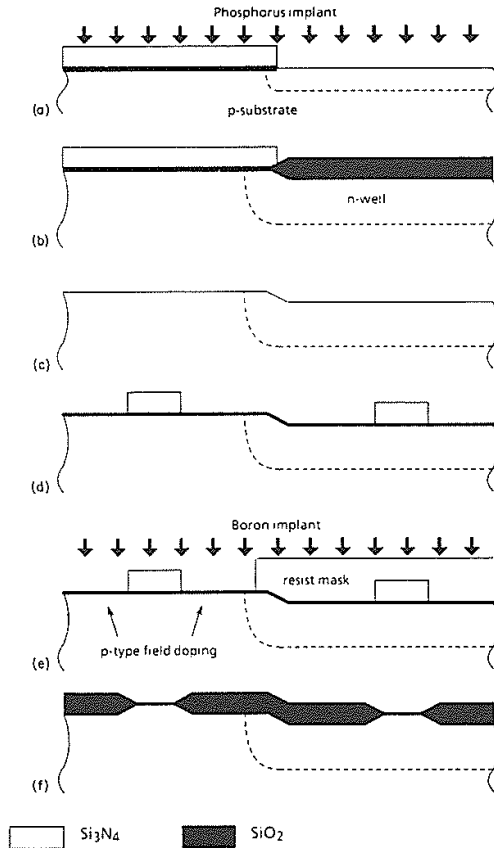


Fig. 8-4 Conventional *n*-well architecture with *n*-well as first mask. (a) Pad oxide/nitride well mask and well implant. (b) Well oxidation and drive-in. (c) Nitride/oxide strip. (d) LOCOS pad oxide/nitride mask and active area definition. (e) NMOS field implant. (f) LOCOS field oxidation. A.G. Lewis and J.Y. Chen, "Current Trends in MOS Process Integration," in N.G. Einspruch and G. Gildenblatt, Eds. *Advanced MOS Device Physics*, Academic Press, San Diego CA, 1989, Reprinted with Permission.

methods that have been developed to provide such alignment marks. For simplicity we utilize an *n*-well CMOS structure in depicting the process flows, but later we show how these methods can be incorporated into twin-well CMOS processes.

In the first approach the wells are the initial features created on the wafer. Here a nitride mask is employed to both define the well locations and to serve as an oxidation mask (see Fig. 8-4a and 8-4b). That is, the patterned nitride/resist layer blocks the phosphorus n -well-implant from penetrating to the p -substrate (or p -well) regions. Part of the drive-in step after the implant is carried out in an oxidizing ambient with the nitride mask layer left in place. This causes a thermal oxide layer to be grown selectively over the n -well regions. At the conclusion of the drive-in process, this oxide is stripped. Since the oxide growth consumes some silicon in the n -regions and since no such silicon loss occurs in the regions covered by the nitride mask, a step is produced on wafer surface when the oxide is stripped (Fig. 8-4c). This step is also self-aligned to the well edge, therefore providing a visible indication of the well boundaries that allows the active area features to be aligned to them. (Subsequent layers, however, are usually aligned to the active area.) The disadvantage of this approach is the step that is created on the wafer surface. In submicron processing such nonplanar topography is undesirable.

The second approach is a simpler process and it also avoids the creation of the step produced in the previous method (see Fig. 8-5). Here, however, the first features to be defined are the active areas (Mask #1). That is, a nitride mask stack layer is patterned first and all subsequent layers are aligned to the active area pattern. Then the well mask (Mask #2) is aligned to the active area pattern, and the well implant is performed. Note that the implanted dopants must penetrate the nitride mask stack. This does not present a serious difficulty, but it does mean that the n -well tends to be shallower and slightly less-heavily doped beneath active regions than under the field areas. Both the first and second approaches require three masking steps (including the boron channel-stop implant that is performed after the well implant - but prior to the LOCOS field oxidation process). However, the second approach offers the benefit of needing fewer oxidations and depositions.

The first method can be applied to a twin-well process by growing an oxide after the n -well implant (Mask #1), but not doing much of the drive-in step during this oxidation. Then the nitride mask is stripped and a second p -well mask is aligned to the n -well oxide (Mask #2). After doing the p -well implant the well drive-in process is performed. An even easier (i.e., one mask) procedure would be to use the n -well oxide to mask the p -well implant (instead of using the second mask). The second method can also be applied to a twin-well process but two masks must be used to create the twin wells. However, if a shallow p -well is acceptable, the same resist mask that is used to form the channel-stop region in Fig. 8-4e (Mask #3) can be used to form the shallow p -well. In this case the p -well implant must be performed at a higher energy than the field implant in order to penetrate the thin nitride mask layer.

The third method utilizes an initial alignment mark etched into the silicon surface prior to forming any device features. This mark is also used as an automatic alignment target by the wafer stepper. Figure 8-6 shows an example of such an alignment pattern.¹⁶ The use of such initial alignment marks is likely to increase with the proliferation of automatic stepper alignment.

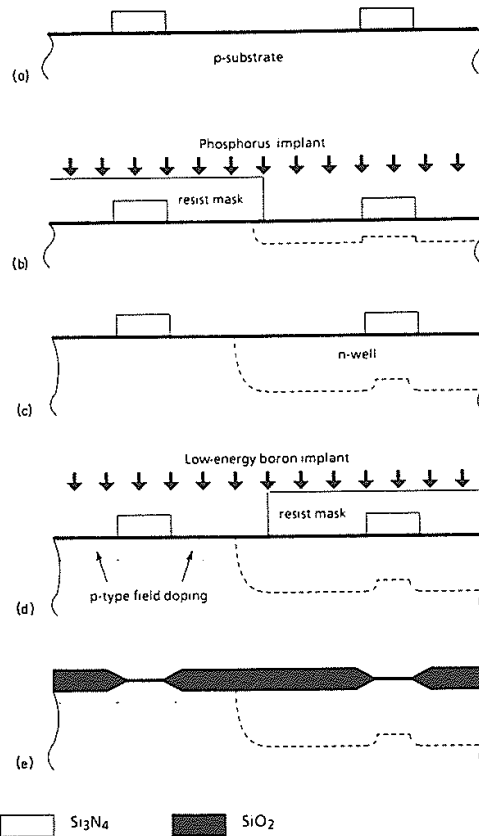


Fig. 8-5 Conventional *n*-well architecture with active area as first mask. (a) Active area definition using LOCOS nitride mask. (b) *n*-well mask and implant. (c) *n*-well drive-in. (d) boron field implant, and (e) LOCOS field oxidation. A.G. Lewis and J.Y. Chen, "Current Trends in MOS Process Integration," in N.G. Einspruch and G. Gildenblatt, Eds. *Advanced MOS Device Physics*, Academic Press, San Diego CA, 1989, Reprinted with Permission.

8.2.2 Twin-Well and Channel-Stop Formation.

The most obvious method of forming twin wells is to use two masking steps, each of which blocks one of the well implants. A procedure for forming twin-wells with a single masking step, however, has also been suggested (see Fig. 8-7).¹² In this method,