Note that under these conditions, in which V_{Ch} and I_D are small, the channel behavior is ohmic and $R_{Ch} = dV_{Ch}/dI_D \propto (V_{GS} - V_T)^{-1}$.

At higher values of V_{Ch} and I_D the free-electron surface charge density Q_{IL} and the electric field E_y both vary along the channel, as shown in Figure 3.3*c*. At $y = 0$, the source end, Q_{IL} is unchanged:

$$Q_{IL}(0) = C_{ox}(V_{GS} - V_T) \tag{3.7}$$

But at the drain end, it is reduced as a consequence of the voltage dropped along the channel:

$$Q_{IL}(l) = C_{ox}(V_{GS} - V_T - V_{Ch}) \tag{3.8}$$

If we take $C_{ox}(V_{GS} - V_T - \frac{1}{2}V_{Ch})$ to be the average surface charge density, the total free-electron charge in the channel is reduced to

$$q_e = C_{ox}wl(V_{GS} - V_T - \tfrac{1}{2}V_{Ch}) \tag{3.9}$$

The average drift velocity remains $v_d = \mu_e V_{Ch}/l$, so the transit time is still given by Equation (3.5). The drain current is thus

$$\begin{aligned} I_D = \frac{q_e}{t_{tr}} &= \mu_e \frac{w}{l} C_{ox}[(V_{GS} - V_T)V_{Ch} - \tfrac{1}{2}V_{Ch}^2] \\ &= k[(V_{GS} - V_T)V_{Ch} - \tfrac{1}{2}V_{Ch}^2] \end{aligned} \tag{3.10}$$

The coefficient $k = \mu_e(w/l)C_{ox}$, is often referred to as the *device transconductance parameter*.

When V_{Ch} is increased to the value $V_{GS} - V_T$, $Q_{IL}(l)$ becomes zero. The channel then *pinches off*, and the inversion layer no longer continues through to the drain end. There a depletion layer forms. The situation is illustrated schematically in Figure 3.3*d*. Further increases in V_{DS} do not change V_{Ch} or I_D but merely serve to expand the depletion layer at the pn junction. The increased voltage appears as the voltage V_{SCL} shown in Figure 3.2, which is dropped across this depletion layer. We discussed in Appendix 1 how heavier doping of the p region than the n^- region minimizes any reduction of the channel length caused by the expansion of the depletion layer. As a result, the conditions in the channel are little affected by increases in V_{DS}. The drain current and channel voltage remain sensibly constant. Substituting $V_{Ch} = V_{GS} - V_T$ into Equation (3.10) gives

$$\begin{aligned} I_D = I_{D(Sat)} &= \tfrac{1}{2}\mu_e \frac{w}{l} C_{ox}(V_{GS} - V_T)^2 \\ &= \tfrac{1}{2}k(V_{GS} - V_T)^2 \end{aligned} \tag{3.11}$$

The characteristics shown in Figure 3.1 make it clear that for VDMOS FETs, $I_{D(Sat)}$ remains quite independent of V_{DS} beyond pinch-off. In other types of MOSFET, where a small increase in $I_{D(Sat)}$ does occur, this can best be represented by

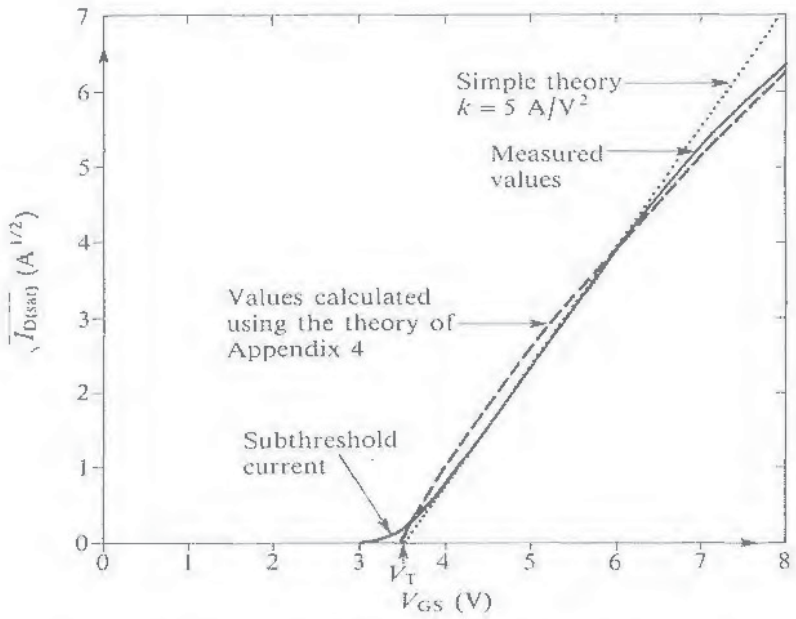


Figure 3.4. A plot of the square root of the saturated drain current versus the gate–source voltage. The data are taken from characteristics like those of Figure 3.1. They show that the behavior of this IRF 540 sample can be represented by $V_T = 3.5$ V, and $k = 5$ A/V^2 for a range of drain currents between 0.2 and 16 A. Deviations occur at higher and lower currents.

introducing a *channel-length modulation parameter*, λ. Then

$$I_D = I_{D(Sat)}(1 + \lambda V_{DS}) \tag{3.12}$$

where $I_{D(Sat)}$ is given by Equation (3.11). For VDMOS devices $\lambda \approx 0$.

In Figure 3.4 a plot of $\sqrt{I_{D(Sat)}}$ versus V_{GS}, taken in part from the curves of Figure 3.1, demonstrates the applicability of Equation (3.11) over a limited range of values. The measured value of k for this device, the IRF 540, is 5 A/V^2. If we assume that $t_{ox} = 100$ nm, so that $C_{ox} = 3.45 \times 10^{-4}$ F/m^2, that the gate width is about 1 m, and that the electron mobility is 0.05 m^2/V s, this value of k implies a channel length in the region of 3.5 μm. This is several times longer than we would expect. Furthermore, the characteristics shown in Figure 3.1 are much steeper near the origin than Equation (3.10) predicts, and they approach the constant-current condition more sharply.

Differentiation of Equation (3.10) with respect to V_{GS} shows that in the linear region, the transconductance $\partial I_D/\partial V_{GS}$ (with V_{Ch} kept constant) should be kV_{Ch}. Differentiating with respect to V_{Ch} (with V_{GS} kept constant) shows that the slope of the I_D versus V_{Ch} characteristics at the origin should be $k(V_{GS} - V_T)$. When allowance is made for $R_{DS(on)}$, the curves of Figure 3.1 indicate a value of k, on this basis, of approximately 24 A/V^2. This implies that the effective channel length is less than 1 μm, which is rather shorter than we would expect.

These discrepancies are not surprising, because there are several assumptions in the simple theory that do not bear close examination. These are discussed in Appendix 4. There it is shown that a physically more self-consistent theory

predicts values for I_D and $I_{D(Sat)}$ that fit the measured device characteristics over a much wider range of values of V_{GS} and V_{DS}. We return to these results shortly, but first we consider what happens when higher gate voltages are applied.

The predicted square-law relationship between $I_{D(Sat)}$ and $V_{GS} - V_T$ comes about because first the total quantity of free-electron charge in the pinched-off channel is proportional to $V_{GS} - V_T$, and secondly the average longitudinal electric field in the channel is proporitional to $V_{Ch} = V_{GS} - V_T$ and thus so too is the carrier drift velocity. With the very short channels used in VDMOS FETS this is something that changes at higher values of $V_{GS} - V_T$ and $I_{D(Sat)}$. The electric field along the channel, whose average value is V_{Ch}/l, becomes so large that it causes the electron drift velocity to saturate. This phenomenon is illustrated in Figure 3.5. Only for electric-field strengths less than about 5×10^5 V/m (0.5 V/μm) is there a well-defined, constant electron mobility ($v_d \propto E$). At higher field strengths the drift velocity increases more slowly until, for fields exceeding about 5×10^6 V/m (5 V/μm), v_d becomes constant at the saturated drift velocity, $v_s \approx 9.2 \times 10^4$ m/s. Unlike the electron mobility, v_s does not appear to depend on the normal electric field E_x at the semiconductor surface [3], but it is expected to decrease with increasing temperature.

For $V_{GS} - V_T > E_s l$, where $E_s = v_s/\mu_e$, the carrier drift velocity approaches v_s over an appreciable proportion of the length of the channel before pinchoff occurs. Remember that pinchoff is defined as the point when $Q_{IL}(l) = 0$. The transit time then approaches the limiting value

$$t_{tr} = l/v_s \tag{3.13}$$

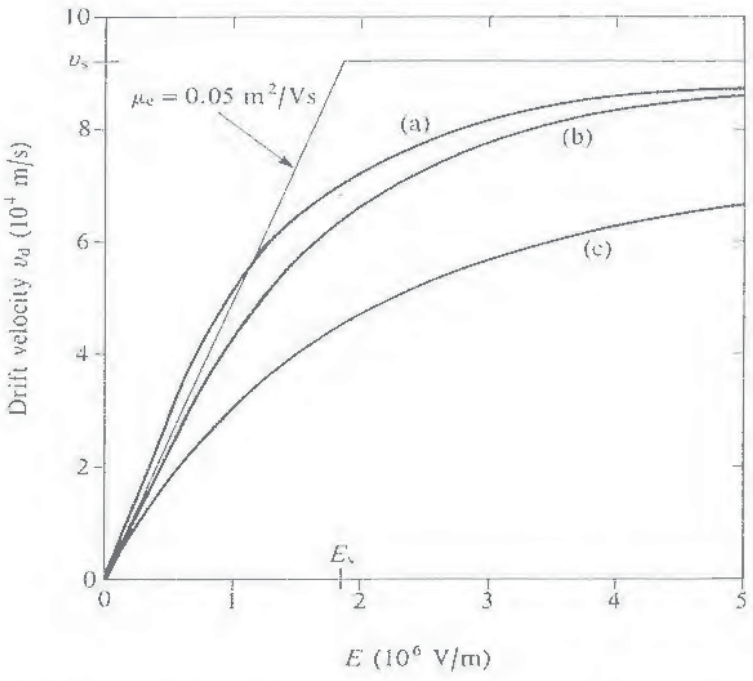


Figure 3.5. Saturation of the carrier drift velocity. Curve (a) shows experimental data based on the results of Ref. 3. Curves (b) and (c) are plots of Equation (A4.22) with $\mu_e = 0.05$ m^2/V s, $v_s = 92{,}300$ m/s, and $\alpha = 2$ and 1, respectively.

At pinchoff, $V_{Ch} = V_{GS} - V_T$, so that combining Equations (3.9) and (3.13) gives

$$I_{D(Sat)} = q_e/t_{tr} = \tfrac{1}{2}C_{ox}w(V_{GS} - V_T)v_s \tag{3.14}$$

At the higher levels of drain current, these effects cause a more even spacing of the I_D-versus-V_{DS} characteristics. This can be seen in the departure from the straight-line relationship in Figure 3.4 at high current levels. When these data are extended and plotted as $I_{D(Sat)}$ versus V_{GS}, the linear relationship expected under velocity saturation conditions is obtained. This is shown in Figure 3.6. Equation (3.14) predicts the gradient of the curve to be $g_{fs} = \frac{1}{2}C_{ox}wv_s$. Putting $v_s = 9.2 \times 10^4$ m/s, and using the values suggested previously as appropriate for the IRF 540 (namely $C_{ox} = 3.45 \times 10^{-4}$ F/m^2 and $w = 1$ m), we should expect this slope to be about 16 S. It can be seen in Figure 3.6 that the measured value is 12 S. Thus, once again, the simple theory predicts the general form of the device characteristics quite well, but not the transition to velocity saturation, nor the absolute value of g_{fs} once this is established.

Taking $l \approx 1$ μm, we expect velocity saturation effects to become apparent over the range $0.5\text{ V} < V_{GS} - V_T < 5\text{ V}$. In fact they are seen in Figure 3.6 to be fully established at a gate voltage of 6 V and a drain current of 16 A. This corresponds to $V_{GS} - V_T = 2.5$ V, and the current is well within the maximum continuous current rating of the IRF 540, which is 27 A.

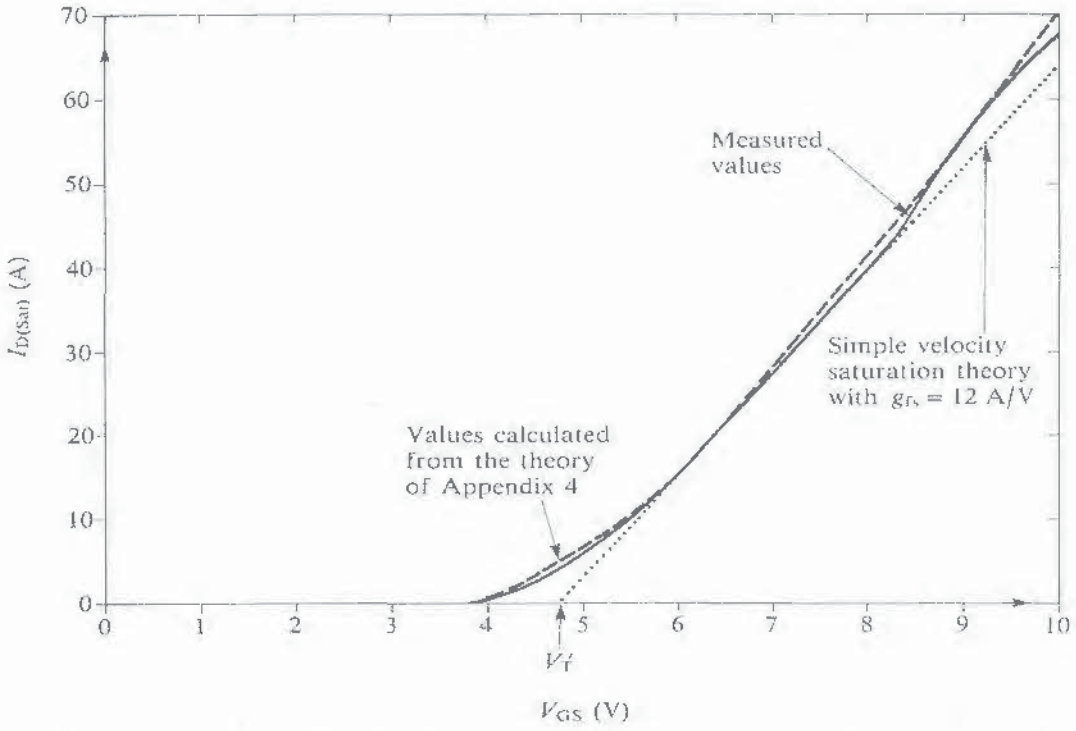


Figure 3.6. A plot of the saturated drain current versus the gate–source voltage. The data are taken from characteristics like those of Figure 3.1. They show that the behavior of this IRF 540 sample can be represented by $g_{fs} = 12$ S and $V_T' = 4.7$ V over the current range 10 to 45 A. Deviations occur at higher and lower currents.

It should be noted that the effective value of the threshold voltage in the velocity saturation region, which we call V'_T, is higher than the effective value, V_T, that applies when the square law [Equation (3.11)] relates I_D and V_{GS}. This can be seen by projecting back the linear parts of the graphs in Figures 3.4 and 3.6 to $I_{D(SAT)} = 0$. This type of behavior would be expected in a more detailed theory, because of the gradual onset of velocity saturation. For purposes of device modelling, V'_T should be used in Equation (3.14), rather than V_T.

Velocity saturation can be a valuable feature of power MOSFETs, even though it causes the drain current to become limited at much lower levels than would otherwise occur. It gives rise to a linear transfer characteristic (g_{fs} = constant) that is beneficial in linear-amplifier applications.

It has become customary to refer to velocity saturation in power MOS devices as a "short-channel" effect. This is unfortunate because to a VLSI engineer, who is designing small MOSFET devices for very large-scale integration, the *short-channel effect* is something quite different and wholly bad. When channel lengths are reduced in the quest for smaller devices, basic device characteristics can be preserved, as long as all linear dimensions and all voltages are scaled in direct proportion and doping densities are varied in inverse proportion [4]. If this is not done, V_{DS} may modulate the channel length, making λ large. It may even cause variations in the height of the potential barrier at the source end of the channel, thereby preventing effective cutoff at the threshold voltage. Clearly, velocity saturation gives rise to a very different kind of short-channel effect in power MOS devices.

When devices are operated at high current levels, two effects may act to reduce the electron mobility and hence the drain current. This may occur either in normal use, or when device characteristics like those of Figure 3.1 are being measured. The first is simply the rise in junction temperature that results from the dissipation. Unless very short pulses at low duty cycle are used, this can cause an apparent negative differential resistance, with the drain current falling as V_{DS} is increased. The second is the increased electric field normal to the surface, E_x, that results from the higher values of V_{GS} required. This is expected to reduce μ_c, but not v_s. The average value of E_x in the inversion layer may be taken to be $(Q_D + \frac{1}{2}Q_{IL})/\epsilon_0\epsilon_{Si}$. It thus varies from about 15 V/μm at threshold to over 40 V/μm when $V_{GS} = 20$ V. For device modeling the variation of μ_c is usually approximated by putting $\mu_c = \mu_{c0}/[1 + \theta(V_{GS} - V_T)]$. Values of μ_{c0} and θ may be expected to depend on the doping level and the surface conditions. The values normally used for a typical power VDMOS FET are 0.05 m^2/V s and 0.1 V^{-1}, respectively. Such a large value for θ includes effects associated with higher gate voltages in addition to the direct effect on the mobility through the normal field E_x.

The more complete theory for the drain-current characteristics presented in Appendix 4 takes into account several effects neglected in the simple theory. These include:

1. the extra charge contained in the depletion layer at the drain end of the channel because of the higher voltage dropped across it;
2. the reduced electron mobility and eventual velocity saturation at high longitudinal field strengths E_y.

Neither of these alters the predicted drain current at very low levels. Both cause I_D to increase more slowly as V_{Ch} rises and the constant-current condition is approached. It is this that gives rise to the discrepancy between the simple theory of Equations (3.10) and (3.11) and the device characteristics.

There remains one more important effect that is neglected in the theory. That is the gradual reduction in the doping concentration of the channel towards the drain end. This leads to a reduction in the charge density in the depletion layer (Q_D reduced), and hence to an increase in the inversion-layer charge (Q_{IL}). As a result, the longitudinal electric field needed to maintain any given drain current is reduced, and the saturation drain current $I_{D(Sat)}$ is increased. It is this that accounts for the unexpectedly low effective channel lengths implied by the characteristics.

The variation of v_d with E_y assumed in the theory of Appendix 4 takes the

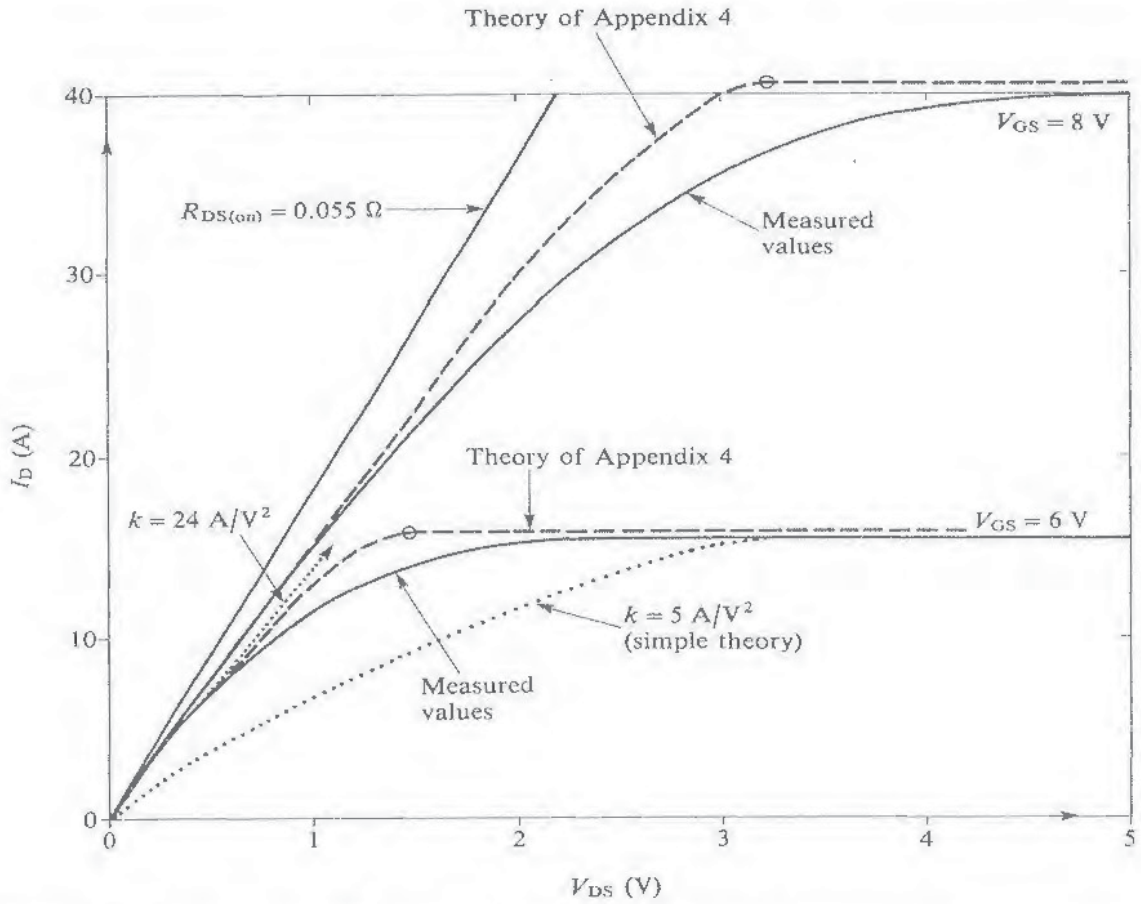


Figure 3.7. Comparison of theoretical and experimental characteristics. Full curves are taken from the experimental characteristics for the IRF 540 shown in Figure 3.1. The dotted curves are plots of Equation (3.10) for $k = 5$ and 24 A/V^2. The lower value fits the constant current region with $V_{GS} = 6$ V; the higher value fits the observed gradient at the origin. The dashed curves are plots of Equation (A4.34), and represent the results of the more sophisticated theory set out in Appendix 4. It can be seen that this theory predicts a sharper approach to the constant-current condition than occurs in practice. In each case, in calculating V_{DS}, a value of 0.055 Ω has been assumed for $R_{DS(on)}$.

form $v_d = \mu_e E_y/(1 + E_y/E_s)$, with μ_e being a function of the gate–source voltage, as has just been mentioned. Then $E_s = v_s/\mu_e$, with $v_s = 9.23 \times 10^4$ m/s. The effect on the theoretical characteristics is simply to increase the effective channel length by V_{Ch}/E_s. Thus k decreases as V_{Ch} and I_D increase. It can be seen in Figure 3.5 that this assumed form for $v_d(E_y)$ underestimates the true drift velocity at any given field.

Results obtained using the theory of Appendix 4 are included in Figures 3.4 and 3.6. The device parameters used previously are assumed: $n_A = 1 \times 10^{23}\ m^{-3}$, $n_D = 1 \times 10^{26}\ m^{-3}$, $t_{ox} = 100$ nm, $Q_{SS} = 4 \times 10^{-4}\ C/m^2$, $w = 1$ m. The effective channel length is taken to be $l = 0.7\ \mu$m, and the mobility parameters used are $\mu_{e0} = 0.06\ m^2/V\,s$ and $\theta = 0.02\ V^{-1}$.

Equation (3.10) predicts a parabolic variation of I_D with V_{Ch}, with the drain current remaining constant once it reaches its maximum value. In Figure 3.7, the results of this simple theory and the theory developed in Appendix 4 are each compared with two of the experimental characteristics taken from Figure 3.1. In estimating V_{DS}, a value of 0.055 Ω has been assumed for $R_{DS(on)}$. This is a typical value for the IRF 540. It is clear that neither theory provides a close fit to the experimental curves. The simple theory, with a value of k that matches the observed constant current, predicts an initial rise of drain current that is much too low, and a channel length that is much too long. The full theory approaches constant current rather too abruptly, but makes no arbitrary assumptions. All the input parameters can be related in some way to the construction of the device.

3.3. THE SUBTHRESHOLD REGION

Figure 3.4 shows that $I_{D(Sat)}$ departs from the square-law relationship with $V_{GS} - V_T$ at low as well as at high currents. It is clear that $I_{D(sat)}$ does not fall to zero as V_{GS} approaches the threshold voltage in the way predicted by Equation (3.11). The threshold is in fact quite "soft", and a significant current flows even when V_{GS} is less than V_T. This is called the *subthreshold* current. A log–linear plot of $I_{D(Sat)}$ versus V_{GS} shows that in the subthreshold region the drain current varies exponentially with V_{GS}. This can be seen in Figure 3.8, which shows the results of measurements made on an IRF 540. The exponential variation is maintained over more than five decades.

In Section 2.5 we discuss how the drain current in the region of threshold is governed by the height of the potential barrier facing the electrons in the source. This is illustrated in Figure 2.13. It is therefore only to be expected that the transition from the blocking to the conducting state should occur over a range of values of V_{GS}, the drain current varying exponentially with the barrier height. In Appendix 3, Figure A3.2 shows that the inversion-layer surface charge density Q_{IL} does indeed vary exponentially with the surface potential, right through the threshold region.

Because of the very wide gate used in power MOSFETs, the subthreshold current is considerable. On all power-MOS data sheets the threshold voltage is specified at a saturated drain current (with $V_{GS} = V_{DS}$) of 0.25 or 1 mA. This puts it well into the subthreshold region for most devices. Threshold voltages defined in this way

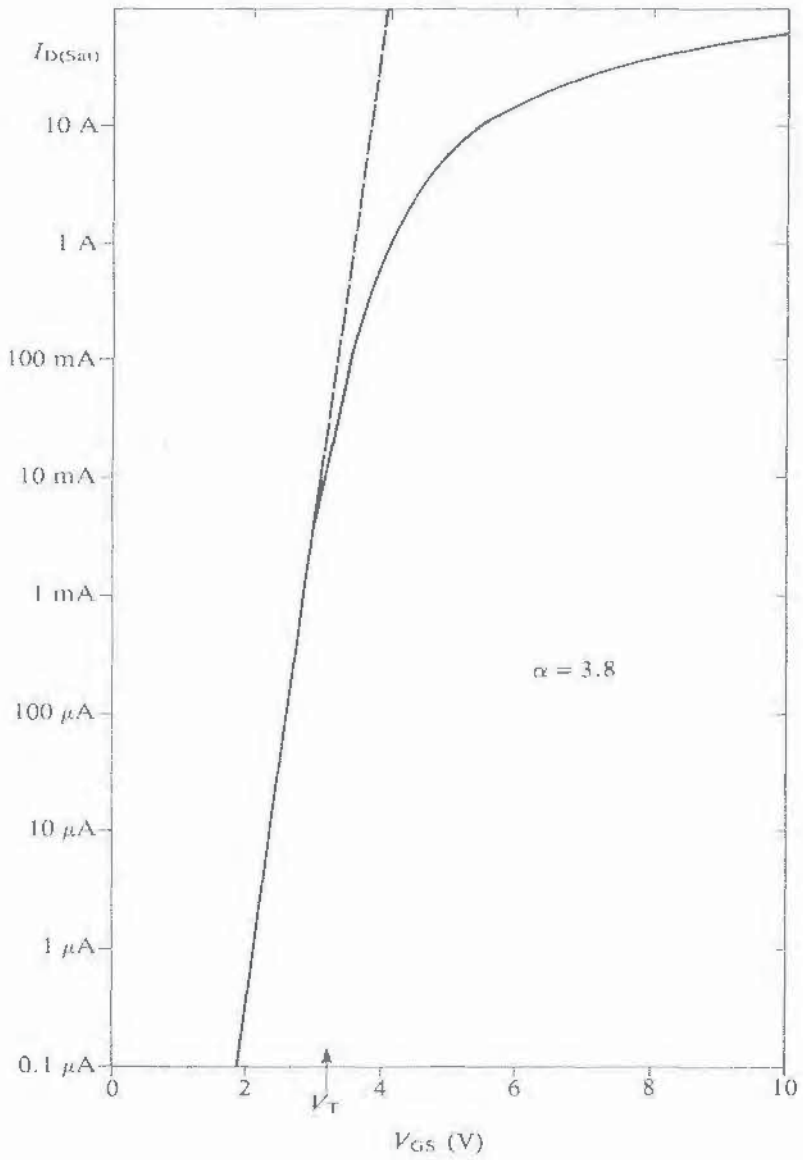


Figure 3.8. A plot of the logarithm of the saturated drain current versus the gate voltage. The data are derived from measured IRF 540 characteristics, including those of Figure 3.1. They demonstrate the exponential variation of the subthreshold current with the gate–source voltage, with $\alpha = 3.8$.

are thus lower than the theoretical values that we define in Chapter 2 and those that are measured from graphs such as Figure 3.4.

It is interesting to estimate the expected saturation drain current, $I_{D(Sat)}$, at threshold. In Appendix 3, it is shown that at threshold, $Q_{IL} \approx 0.1Q_0$, where $Q_0 = (n_A kT\epsilon_0\epsilon_{Si})^{1/2}$. Following the discussion of the last section, let us assume that the average inversion-layer charge density along the channel at pinchoff is half this value. Pinchoff occurs at a channel voltage which is sufficient to reduce Q_{IL} to zero at the drain end of the channel. This requires $V_{Ch} = Q_{IL}/C_{ox} =$

$0.1Q_0/C_{ox}$. Thus, at threshold,

$$I_{D(Sat)} = \frac{q_e}{t_{tr}} = \tfrac{1}{2} \times 0.1Q_0 \frac{w}{l} \mu_e V_{Ch} = \tfrac{1}{2}\mu_e \frac{w}{l} \frac{(0.1Q_0)^2}{C_{ox}} \tag{3.15}$$

As an example, assume parameter values based on the IRF 540, for comparison with Figure 3.8. Let $n_A = 1 \times 10^{23}\ m^{-3}$, so that $Q_0 = 2.08 \times 10^{-4}\ C/m^2$. Let $w = 1$ m, $l = 1\ \mu$m, and $C_{ox} = 3.45 \times 10^{-4}\ m^{-2}$, and take $\mu_e = 0.05\ m^2/V\,s$. Then, at "threshold", $V_{Ch} = 0.06$ V and $I_{D(Sat)} = 30$ mA. It can be seen in Figure 3.8 that the corresponding gate voltage is 3.25 V. This is a little lower (0.25 V) than the value of V_T indicated in Figure 3.4.

With $V_{GS} < V_T$, we expect an exponential variation of the drain current with the barrier height, and hence, as observed, with the gate–source voltage. Thus

$$I_{D(Sat)} \propto \exp\left(\frac{e\,\Delta V}{kT}\right) \propto \exp\left(\frac{eV_{GS}}{\alpha kT}\right)$$

The coefficient α relates V_{GS} to the surface potential ΔV at the source end of the channel, where the potential minimum occurs. Its theoretical value can be obtained by applying Equation (2.7):

$$V_{GB} = V_{GS} = V_{ox} + \Delta V - V_{CP} \tag{2.5}$$

$$= \frac{(2e\epsilon_0\epsilon_{Si}n_A\,\Delta V)^{1/2} - Q_{SS}}{C_{ox}} + \Delta V - V_{CP} \tag{3.16}$$

where we have substituted for V_{ox} using Equations (A2.6) and (A2.7). Thus,

$$\alpha = \frac{dV_{GS}}{d(\Delta V)} = \frac{(e\epsilon_0\epsilon_{Si}n_A/2\,\Delta V)^{1/2}}{C_{ox}} + 1$$

$$= \frac{C_D}{C_{ox}} + 1 \tag{3.17}$$

Here, C_D may be thought of as the depletion-layer capacitance,

$$C_D = \frac{dQ_D}{d(\Delta V)} = \left(\frac{e\epsilon_0\epsilon_{Si}n_A}{2\,\Delta V}\right)^{1/2} \tag{3.18}$$

With $n_A = 1 \times 10^{23}\ m^{-3}$, we have $\Delta V \approx \Delta V_{Th} \approx 0.81$ V and $C_D = 10.2 \times 10^{-4}\ F/m^2$. With $t_{ox} = 100$ nm, $C_{ox} = 3.45 \times 10^{-4}\ F/m^2$. Together these give $\alpha = 4.0$.

The simplest way of measuring α is to take the change in V_{GS} required to cause a factor-of-ten change in $I_{D(Sat)}$. Call this $(\Delta V_{GS})_{Dec}$. Then, at room temperature,

$$\alpha = \frac{(\Delta V_{GS})_{Dec}}{(kT/e)\ln 10} = 16.9(\Delta V_{GS})_{Dec} \tag{3.19}$$

The graph of Figure 3.8 shows that $(\Delta V_{GS})_{Dec} = 0.225\,V$, so the measured value of α is 3.8. This implies an effective doping concentration at the source end of the channel of $0.9 \times 10^{23}\,m^{-3}$. However, in some instances the presence of interface trapping levels may cause Q_{SS} to be a function of ΔV. This would modify the value of α slightly, but the measured value justifies our estimate of n_A very precisely.

The departure from the exponential variation of I_D with V_{GS}, shown in Figure 3.8, starts at 10 mA, when $V_{GS} = 3.05\,V$. This is the point at which the inversion-layer charge density becomes big enough to influence the potential at the semiconductor surface significantly.

3.4. DEVICE TRANSCONDUCTANCE

A device parameter of great importance is the transconductance g_{fs} which is defined as:

$$g_{fs} = \left(\frac{\partial I_D}{\partial V_{GS}}\right)_{V_{DS}=\text{constant}} \tag{3.20}$$

This measures the sensitivity of the drain current to changes in the gate–source voltage. It determines the current-carrying capacity of the device.

According to the theory of Section 3.2 we should expect g_{fs} to take on values as follows:

1. Below pinchoff, when $V_{Ch} < V_{GS} - V_T$, differentiation of Equation (3.6) yields

$$g_{fs} = \mu_e \frac{w}{l} C_{ox} V_{Ch} = k V_{Ch} \tag{3.21}$$

2. Above pinchoff but before velocity saturation, that is, for $E_s l > V_{Ch} \approx V_{GS} - V_T$, Equation (3.11) leads to

$$g_{fs} = \mu_e \frac{w}{l} C_{ox}(V_{GS} - V_T) = k(V_{GS} - V_T) \tag{3.22}$$

3. After velocity saturation, $V_{Ch} > E_s l$, Equation (3.14) applies and

$$g_{fs} = \tfrac{1}{2} w C_{ox} v_s \tag{3.23}$$

For practical purposes the specification of g_{fs} for a given device, as on the device data sheet, requires that certain prescribed values of V_{GS} and V_{DS} must be chosen. Usually, V_{DS} is large enough to ensure that the drain current is in the constant-current region, $I_{D(Sat)}$. The choice of V_{GS} usually ensures that this value of $I_{D(Sat)}$ is about half the maximum current rating of the device. Data sheets also display graphs of g_{fs} versus I_D. A typical plot is shown in Figure 6 of the data

sheet presented in Appendix 7. These curves demonstrate the increase of g_{fs} with I_D at low values, and its independence of I_D once velocity saturation is established.

For a given value of V_{GS}, the three factors that influence g_{fs} are the gate width w, the channel length l, and the gate oxide thickness t_{ox}. In saying this we are presuming that by choosing silicon as the semiconductor and silicon dioxide as the gate insulator, the values of μ_e (or v_s) and ϵ_{ox} are prescribed.

Decreasing t_{ox} increases g_{fs} through its effect on C_{ox}. The limit is set by the need to preserve the uniformity and integrity of the gate oxide, and for it to support the rated voltage V_{GS}, which is usually ± 20 V. The design of the transistor has to ensure that no greater voltage appears across the gate oxide in any permitted operating condition. With $t_{ox} = 100$ nm (1000 Å), the breakdown voltage of good-quality gate oxide averages about 55 V. Technological improvements have reduced the spread around this value, thereby increasing the margin for reliability.

Decreasing l also increases g_{fs}. The minimum value of l is set by the level of technology, in this case the ability to control the double-diffusion process that determines l. What matters is not what can be achieved in ideal circumstances, but the manufacturability of the device with a reasonable yield within the parameter tolerances specified. This becomes increasingly difficult with $l < 2\ \mu$m.

There is greater scope for increasing g_{fs} by increasing the gate width w, because w varies in direct proportion with the active device area. The parameter determined by technology is the gate width per unit area, w/A. There is an advantage in using a cellular rather than a linear pattern, and several different cellular designs have been proposed and used. Some are illustrated in Figure 3.9. The quantitative theoretical advantages of one pattern over those of another are marginal. They are discussed in Appendix 5. They have been the subject of a number of theoretical studies (e.g. [5]) involving a consideration of the ON-state resistance $R_{DS(on)}$, as well as the transconductance g_{fs}. It has recently been shown [6] that a circular cell geometry offers the best compromise, provided that the full bulk breakdown voltage is realized. The hexagonal structure shown in Figure 3.9*b* approximates this optimum design most closely. Again, though, what really matters is the manufacturability and robustness of the device made by a particular process.

What is clear from theory is that w/A increases as the cell lattice spacing decreases. The limit to this is set in two ways: again by the level of technology, but also by the need to maintain a minimum spacing b between the channel diffusions. This region, Ⓒ in Figure 2.1, acts as a throat which constricts the drain current. The limit on b results from the need to minimize the contribution of this region to $R_{DS(on)}$, bearing in mind that the current density is higher here than in the rest of the drain region. This is discussed in the next section. The problem is exacerbated by the presence here of a parasitic JFET, a matter that is taken up in Section 3.6.

In higher-voltage devices the epitaxial layer is more lightly doped and so has a higher resistivity. This not only increases the resistance of the epilayer, but also causes the channel depletion layers to expand further into the drain drift region, thereby enhancing the JFET action. This can be offset to some extent by increasing

the drain doping concentration towards the surface, but the minimum acceptable value of b remains higher in high-voltage devices.

The technological limit is determined by the line width and registration tolerances of the photolithographic processing steps. A key factor is the need to obtain a low-resistance metal contact to the source. The contact area must be sufficient to ensure this after making provision for registration tolerances in the masking processes, which may cause the metal contact to be offset from the center of the cell. The chip area can most readily be filled by a regularly repeated lattice of square or hexagonal cells. Such patterns dominate in commercial devices.

Another limiting factor is the need to radius the channel at each cell apex in order to maintain its length l constant. The more acute the angle of the apex, the more difficult this is, which confers another benefit on the hexagonal cell.

The first range of devices to become widely established commercially in the early 1980s used a cell lattice spacing of 38 μm, giving over 500,000 cells per square inch (approximately 800 per mm^2). The resulting value of w/A is about

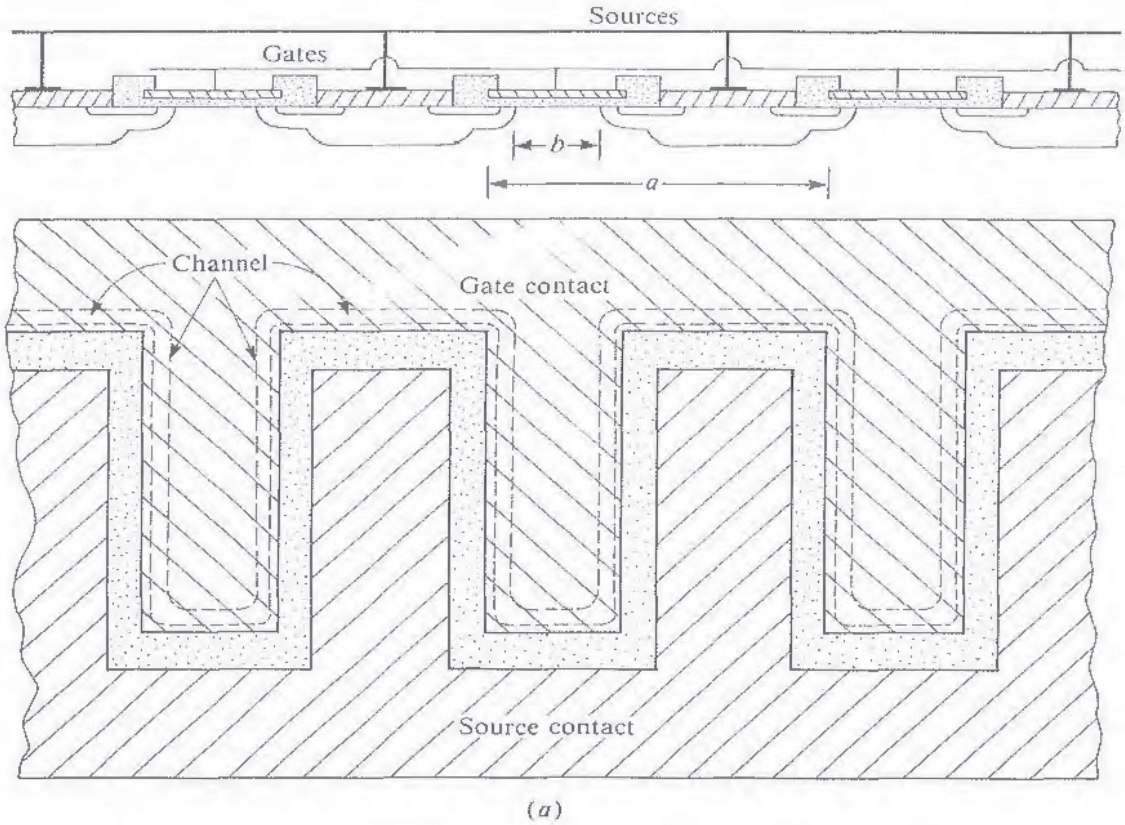


Figure 3.9. Examples of gate patterns: (*a*) Interdigitated. (*b*) Hexagons on hexagonal lattice. The plan view of the poly-Si layer shows the channel positions and the source contact areas. A section through the active region is also given. (*c*) Hexagons on a square lattice. (*d*) Squares on a square lattice. (*e*) Squares on an offset square lattice. (*f*) Squares on a hexagonal lattice. (*g*) Triangles on a hexagonal lattice.

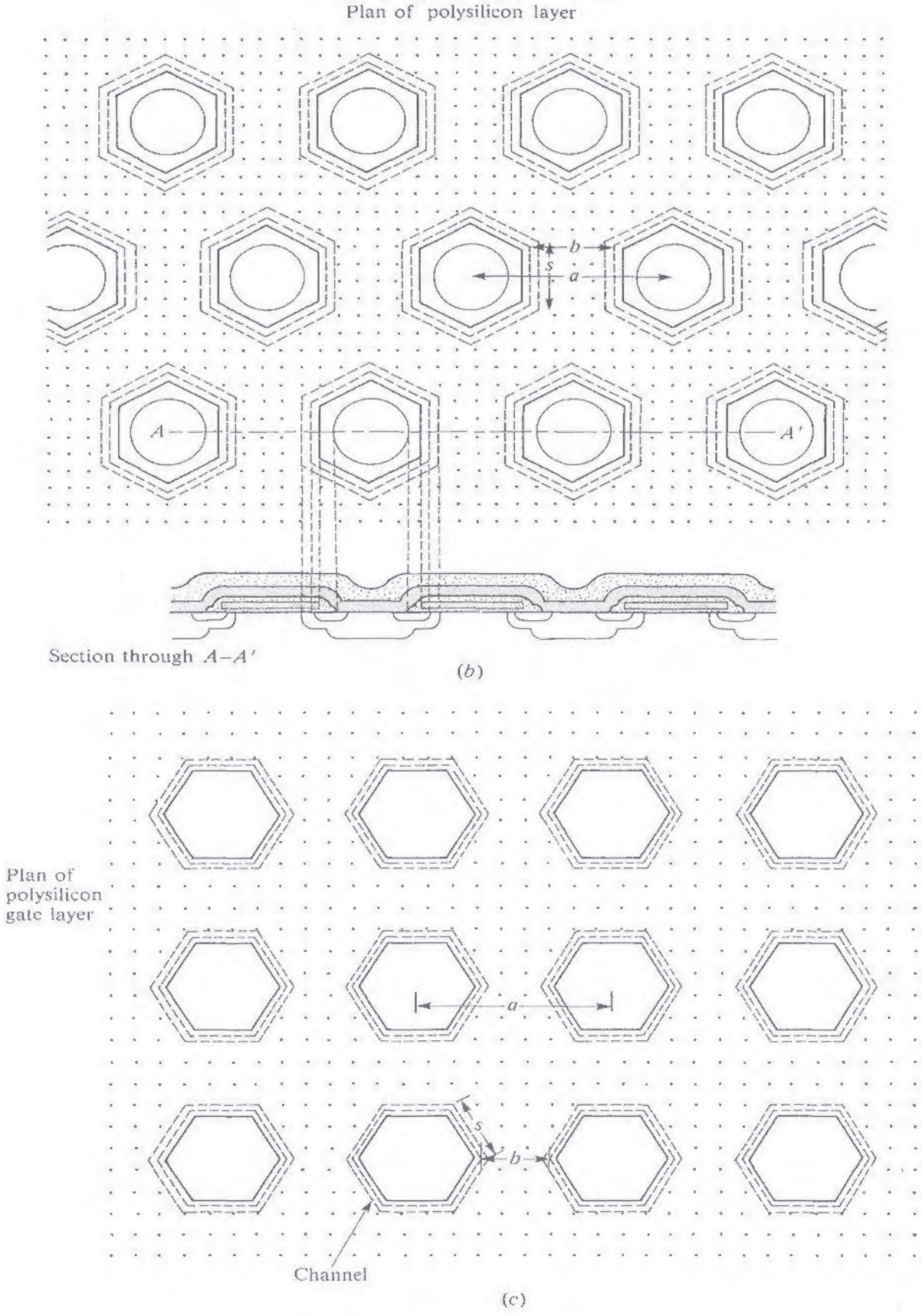
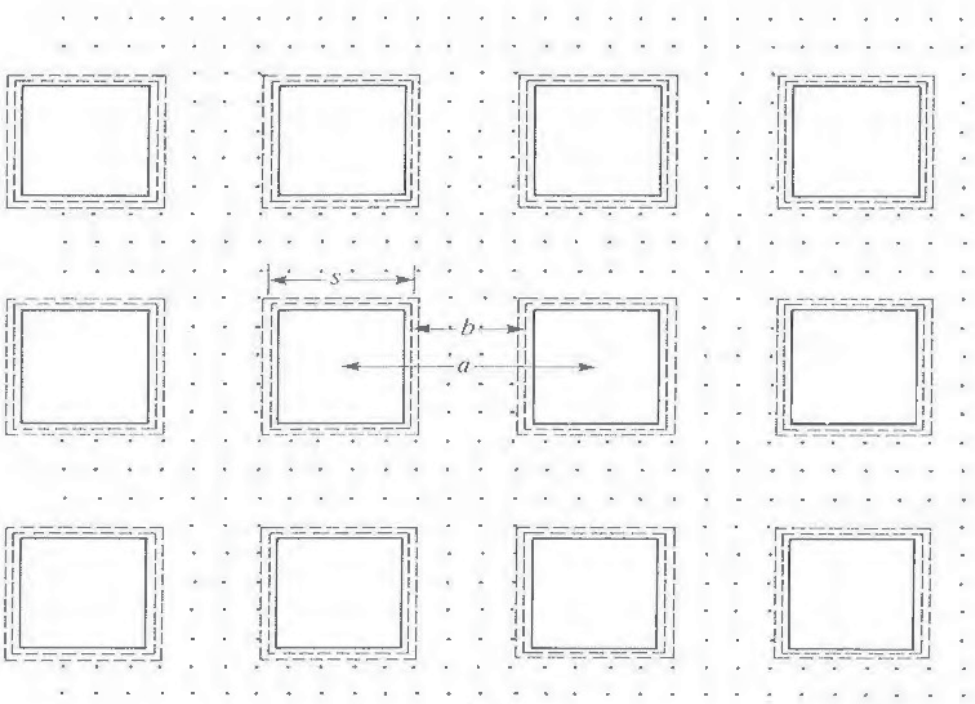
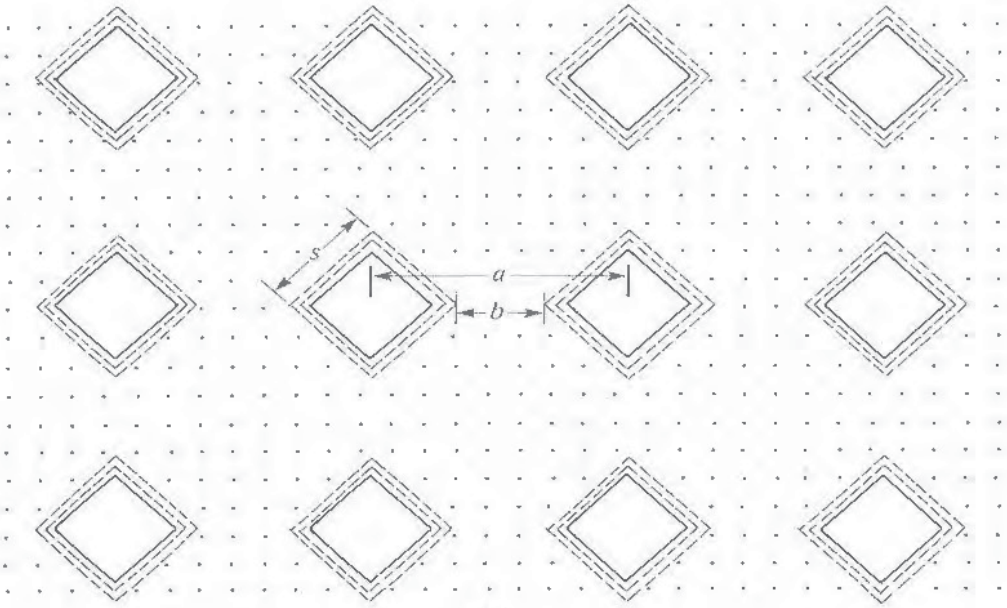


Figure 3.9. (*Continued*)



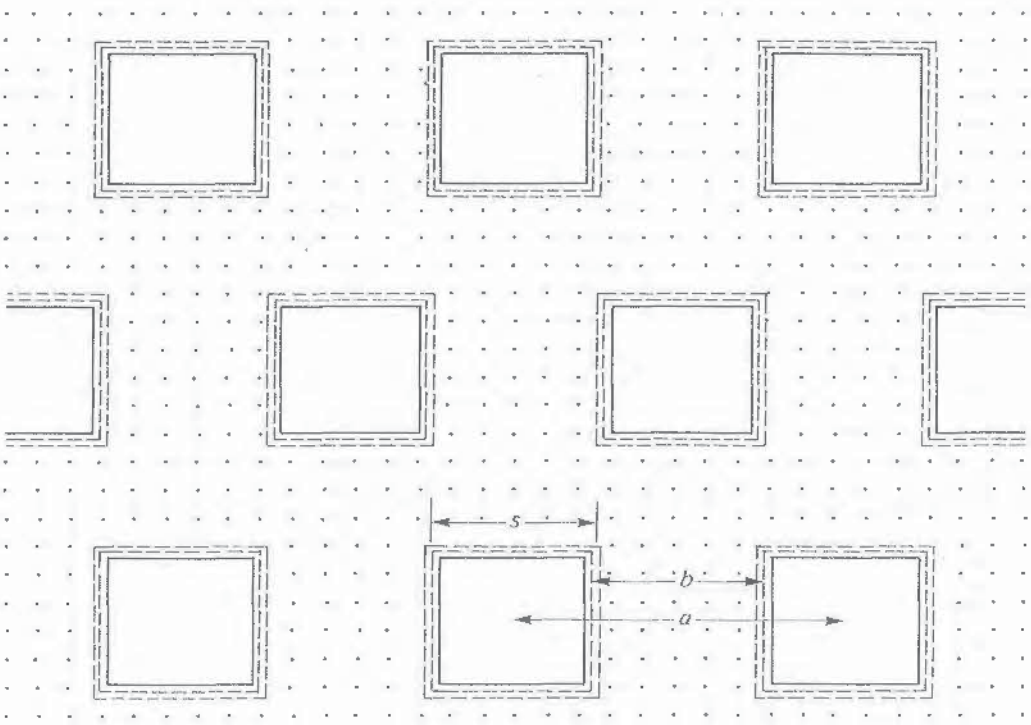
(*d*)



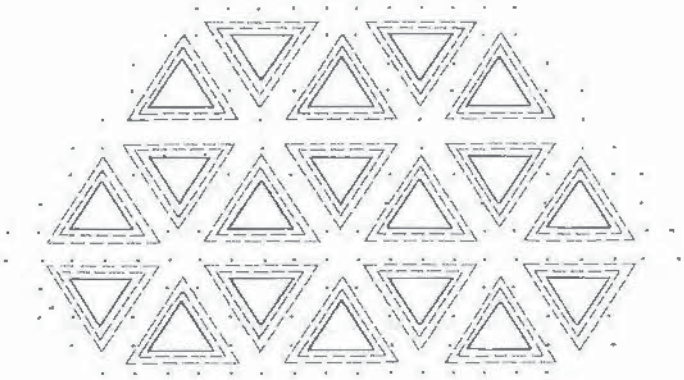
(*e*)

Figure 3.9. (*Continued*)

0.07 m/mm^2. This remains a perfectly satisfactory standard for higher-voltage devices. However, low-voltage MOSFETs have benefited from progressive reductions in the cell lattice spacing to 20 μm and below, with cell densities increasing first to 1,600,000 and more recently to 3,500,000 per square inch (approximately 2500 and 5400 per mm^2, respectively). Corresponding values of w/A are in the



(*f*)

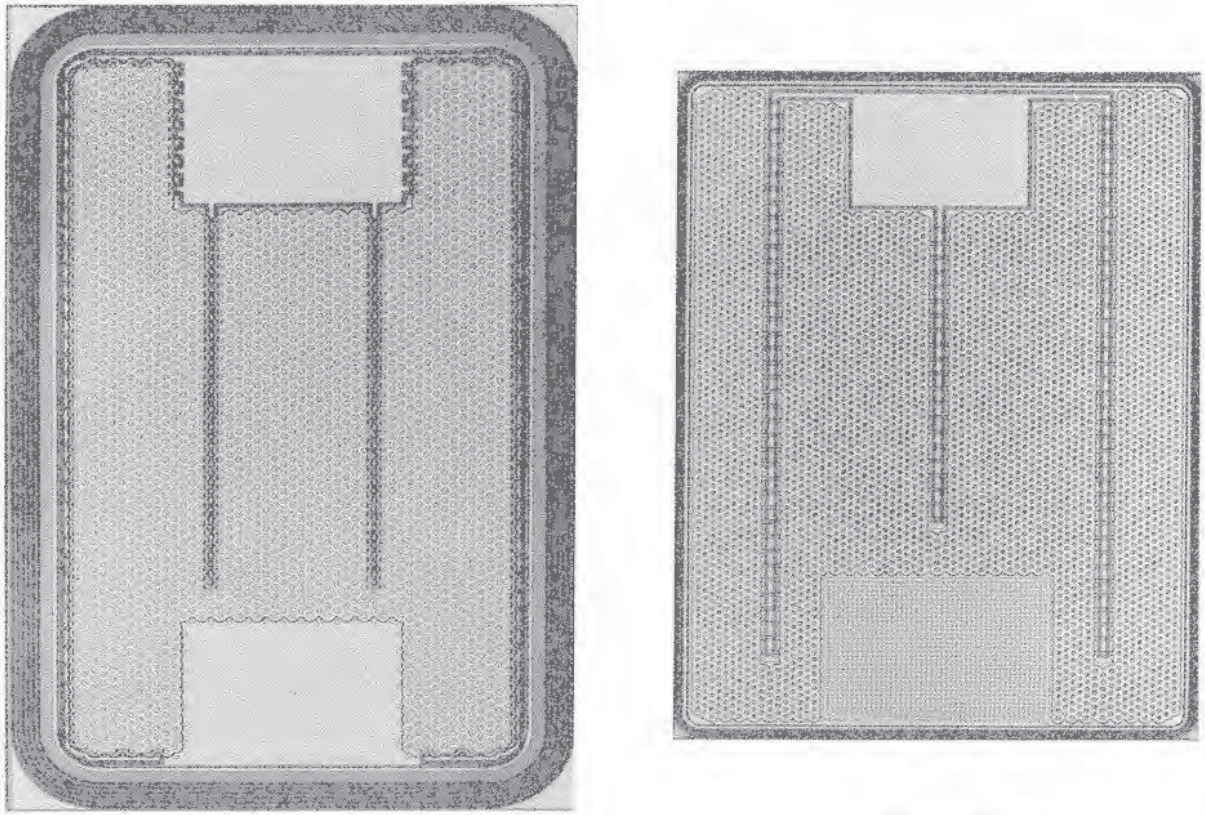


(*g*)

Figure 3.9. (*Continued*)

region of 0.12 and 0.17 m/mm^2, respectively. As a result, these devices can support a given current rating on a chip of smaller area. Since the silicon area required is a major element in the cost of a component, obtaining the maximum performance from a given size of chip is extremely important.

Even so, the proportion of the chip available as active device area A_A may be less than 50% of the total area A_{Ch}. For reasons of reliability the source and gate bonding pads are not often laid over active material. The gate vias that distribute the gate current over the area of polysilicon and the field rings used around the edge of the chip to prevent peripheral voltage breakdow · all diminish the area



(*a*) HEXFET I (*b*) HEXFET III

Figure 3.10. Photographs of HEXFET™ chips. (*a*) Earlier design, HEXFET™ I; (*b*) more recent design, HEXFET™ III. These plan views show the hexagonal cell pattern, the peripheral field rings, the gate vias, and the bonding pads. In the earlier design, the active area occupies 60% of the total chip area. In the later design, the size of the field rings and the gate bonding pad have been reduced, giving better chip utilization. (Courtesy International Rectifier Corp.)

that can be used for transistor action. The plan views of HEXFET™ chips shown in Figure 3.10 illustrate the type of layout used. Approximately 60% of the chip area is active in the earlier design shown in Figure 3.10*a*. This is increased in the more recent design shown in Figure 3.10*b*.

3.5. THE DRAIN DRIFT REGION

The drain drift region is particularly critical to the design of a power MOSFET. Its principal function is to block and support the full forward voltage held off by the transistor in its turned-OFF state. However, it has also to carry the full forward current in the ON state. It is thus most important that its contribution R_{EPI} to the total ON-state resistance $R_{DS(on)}$ be kept as small as possible. This minimizes dissipation and maximizes the current rating of the device. Thus, the epilayer should be no thicker than it has to be in order to support the maximum blocking

voltage. Values range from about 5 μm for a 50-V device to about 50 μm for a 1000-V device.

For purposes of discussion we have subdivided the drain drift region into three parts. In Figure 2.1 these are identified as the surface layer (B); the part bordered by the source and channel diffusions, (C); and the main part below the diffused regions, (D). In the ON state the current should be evenly distributed all across the epilayer. Clearly, this is not possible in the region between the diffusions, where the electron current emerges from the channel regions. The area between the diffusions is a little less than the gate area. Let it be a fraction g of the total active chip area. The even distribution of current here is assisted by the presence of the accumulation layer in the epilayer immediately under the gate oxide, in the turned-on state, and indeed as long as $V_{DS} < V_{GS}$. Under the gate oxide, then, the current density is higher by the factor $1/g$. It is further increased by the presence of the depletion layers at the pn junctions. These expand as the local potential increases at high current levels, and give rise to an effect characteristic of a series junction field effect transistor. This has already been mentioned and is discussed again in Section 3.6.

In high-voltage devices the epilayer is sufficiently thick for the current to be able to spread out to cover the whole of the active chip area fairly evenly by the time it enters the substrate. This is shown in Figure 3.11*a*. With low-voltage devices the chip area can be fully utilized only by further reducing the lateral cell dimensions, as illustrated by Figures 3.11*b* and *c*. This gives an additional incentive for a reduction in the cell pitch a, over and above the increase in w/A that results, as discussed in the last section.

When R_{EPI} is low, as in low-voltage devices, the effect of the other parasitic

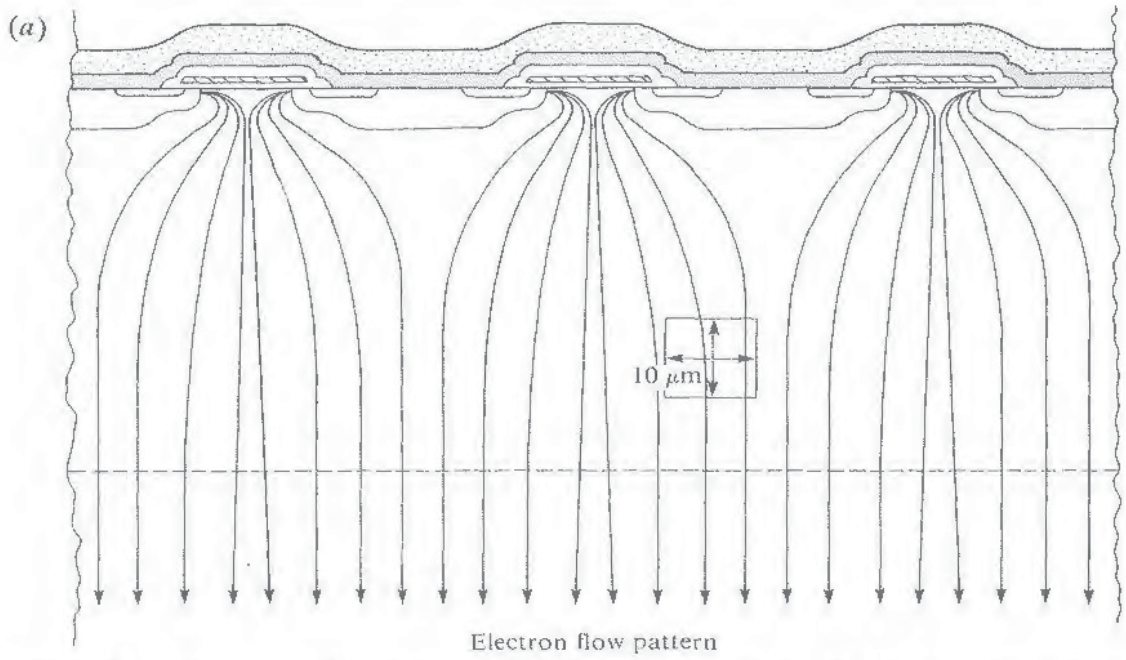


Figure 3.11. Current distribution. (*a*) High-voltage device; (*b*) low-voltage device with same cell pitch; (*c*) low-voltage device with reduced cell pitch.

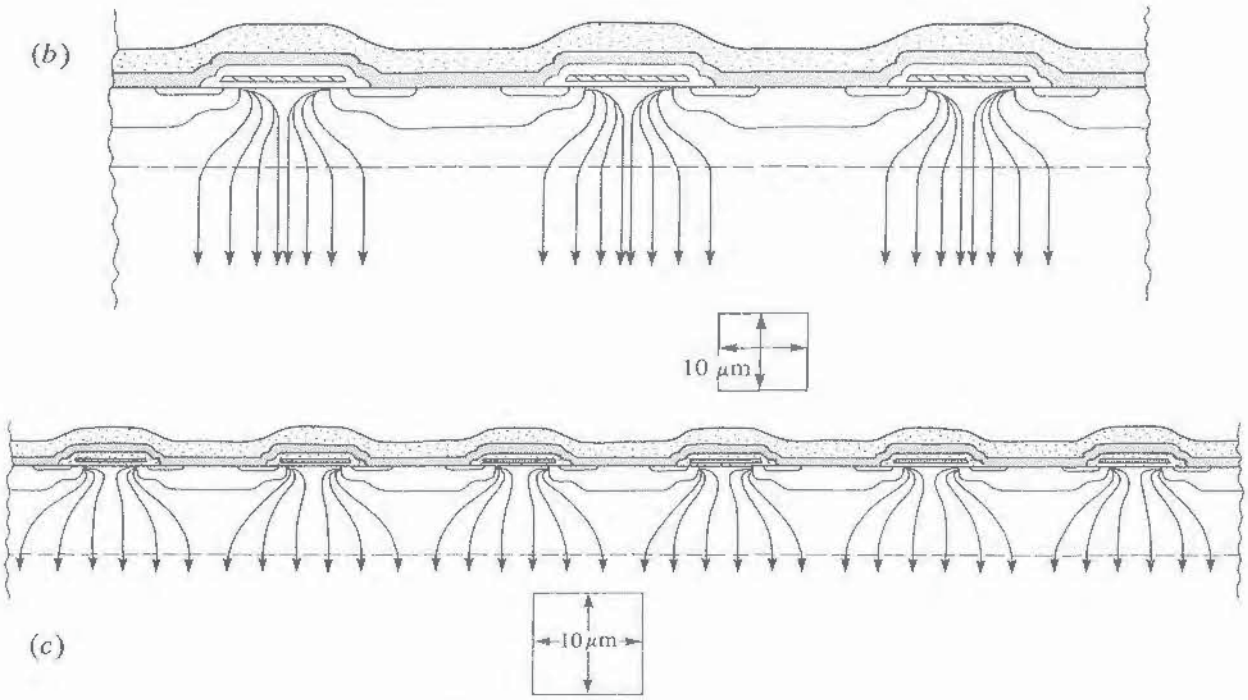


Figure 3.11. (*Continued*)

sources of resistance, R_P, is proportionately greater. A schematic illustration of the relative magnitudes of the different factors contributing to $R_{DS(on)}$ is given in Figure 3.12.

A theory specifying the optimum doping profile and thickness of the epilayer is presented in Appendix 6. It is based on a highly idealized plane-parallel geometry, but it does illustrate the principle of the problem, and it does set into

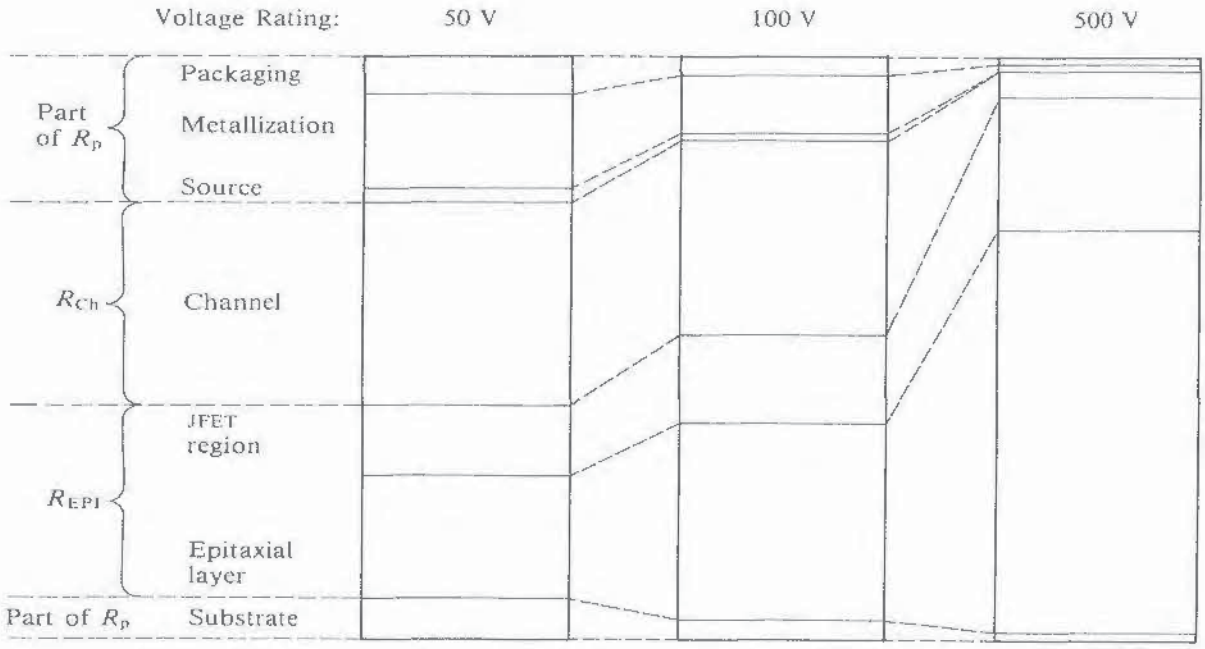


Figure 3.12. Relative contributions to $R_{DS(on)}$ in devices of different voltage rating.

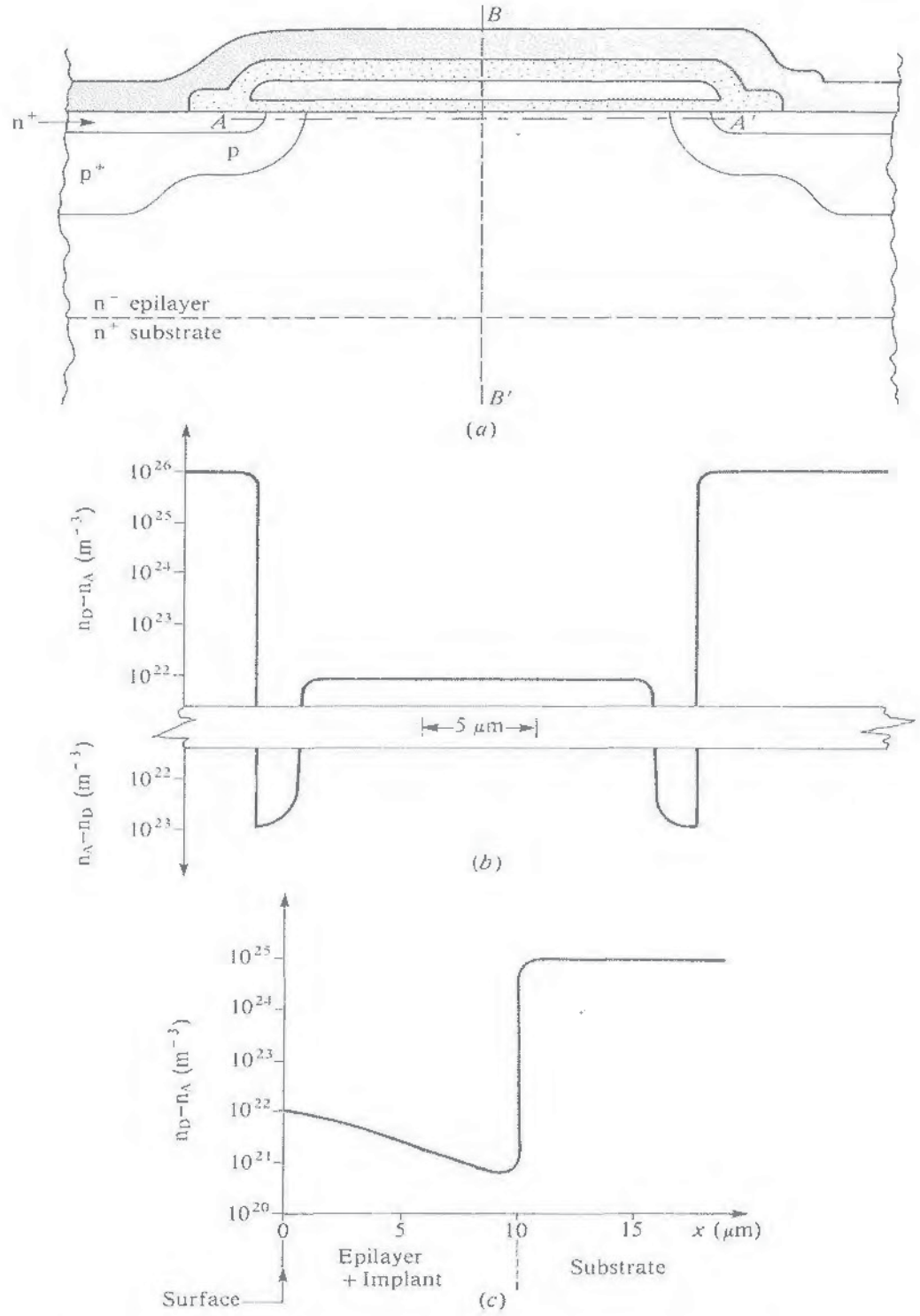


Figure 3.13. Variation of the doping concentration. (*a*) Section through the active region; (*b*) variation parallel to the surface along AA'; (*c*) variation perpendicular to the surface along BB'.

context the much-quoted relationship between the ON resistance and the junction breakdown voltage:

$$R_{DS(on)} \propto V_B^{2.6} \tag{3.24}$$

It is the breakdown voltage V_B that partly determines the voltage rating of the device. The resistance calculated in the theory (R in Appendix 6) is the optimum value of the component R_{EPI} of the ON-state resistance when there is a uniform distribution of current. The lateral nonuniformities of the real device mean that the proposed doping profile of Equation (A6.26) is not appropriate in practice.

Ideally [7], the doping level would follow Equation (A6.26) initially, decreasing from the n^+n^- junction into the epilayer. It would then go through a minimum and increase progressively towards the surface of the chip between the p diffusions. An increase in the epilayer doping concentration in regions (B) and (C) is desirable for two reasons. It helps to minimize the extra contribution to R_{EPI} caused by the higher current density, and more importantly, it reduces the effect of the parasitic JFET. It is found that the breakdown voltage of a well-designed device is not adversely affected. Most benefit is obtained when the doping concentration has lateral uniformity [8], as in the profile shown in Figure 3.13.

Field enhancement would normally occur at the edges and corners of the channel diffusion in the forward-blocking state. But this is reduced if the depletion layers, expanding from the junctions on either side of the throats in the drain drift region beneath the gate oxide, eventually overlap. The breakdown voltage then approaches that of the plane junction between the epilayer and the p^+ diffusion in the centre of the body region under the source contact. In modern devices this junction is designed to withstand avalanche breakdown in a controlled way. These matters are discussed further in the next section.

In Figure 3.14, rated values of $R_{DS(on)}$, normalized against device active area, are plotted against V_B (BV_{DSS}) for some commercially available power MOSFETs. Equations (A6.17) and (A6.28) are also shown. It can be seen that there are significant departures from the theory, especially at low values of V_B. From the earlier discussion this is not surprising, but the reduction of the cell pitch in the more modern designs of low-voltage devices has resulted in their coming nearer to the ideal curves. However, even at the higher values of V_B, the value of $R_{DS(on)}A_A$ is still several times higher than the predicted optimum value RA. One reason is the difficulty of achieving a high fraction of the theoretical value of V_B at the periphery of the chip. Techniques for achieving this are discussed in Chapter 5.

The need to minimize $R_{DS(on)}$ further emphasizes the importance of minimizing the silicon "real-estate overheads" discussed at the end of the previous section. This inevitably leads to compromises in design. Thus, extending the system of field rings at the periphery may help to ensure higher reliability at high voltage, but this is bought only at the expense of a reduced current-handling capacity and an increase in $R_{DS(on)}$. If, in a particular application, this causes the device to run at a higher temperature, its reliability will be degraded.

Some manufacturers have not allowed the bonding pads to overlie active parts of the device. However, an increasing number now have sufficient confidence in

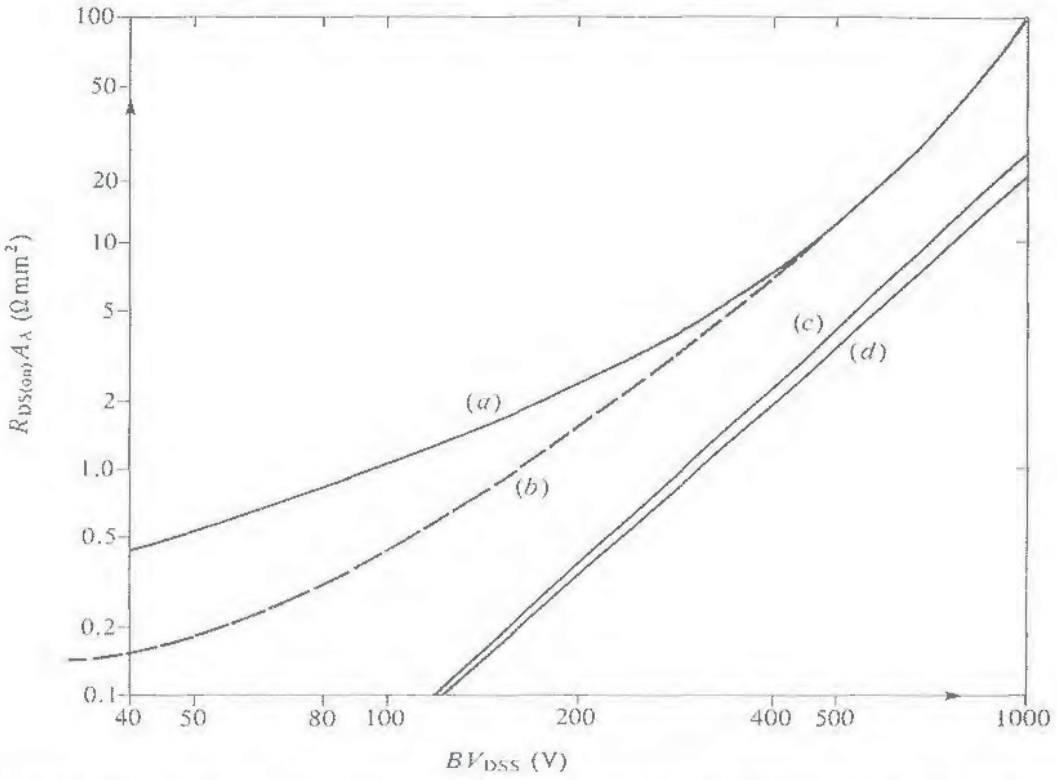


Figure 3.14. Normalized ON resistance versus breakdown voltage. Values of $R_{DS(on)}$ for commercially available devices are compared with the theoretical curves of Appendix 6. Curve (a) shows values typical of standard devices; curve (b) shows values typical of modern low-voltage devices having a high cell density. Curves (c) and (d) represent Equations (A6.17) and (A6.28).

their ability to maintain the integrity of the structure under the bonding pads that they are prepared to extend them over active regions. A final way in which most manufacturers accept an increase in $R_{DS(on)}$ in exchange for extra reliability in the breakdown performance arises from the fact that in general the depletion layer is not permitted to reach through to the substrate. The epilayer is made thick enough to ensure that the field at the n^-n^+ junction is low. Then, for any given voltage, the peak field occurring at the p–n junction is lower than it would be, were reach-through permitted.

3.6. PARASITIC COMPONENTS

3.6.1. Introduction

It will be clear from what has been said already that there are a number of parasitic components inherently associated with the VDMOS FET structure. In addition to the normal resistance of the bulk material and the contacts, the inductance of the leads, and the capacitance of junctions and connections, these parasitic elements include active transistors. Their effects are most pronounced under the transient conditions discussed in Chapter 4. They may severely limit high-frequency operation.

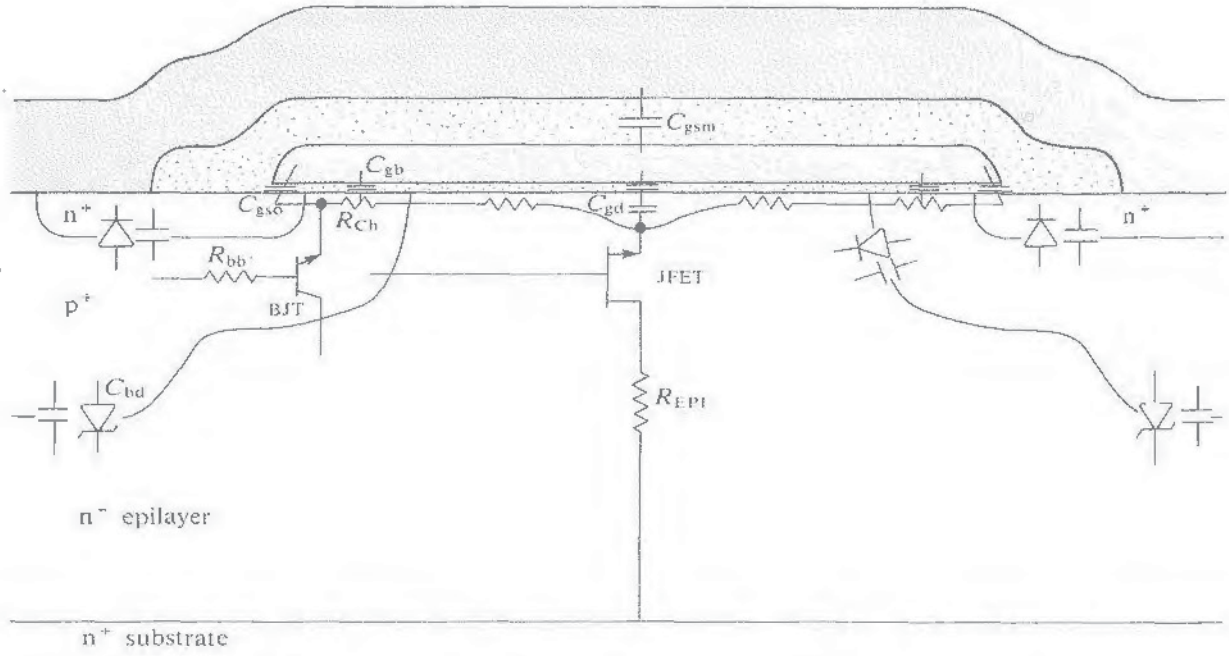


Figure 3.15. The origin of parasitic components within the device.

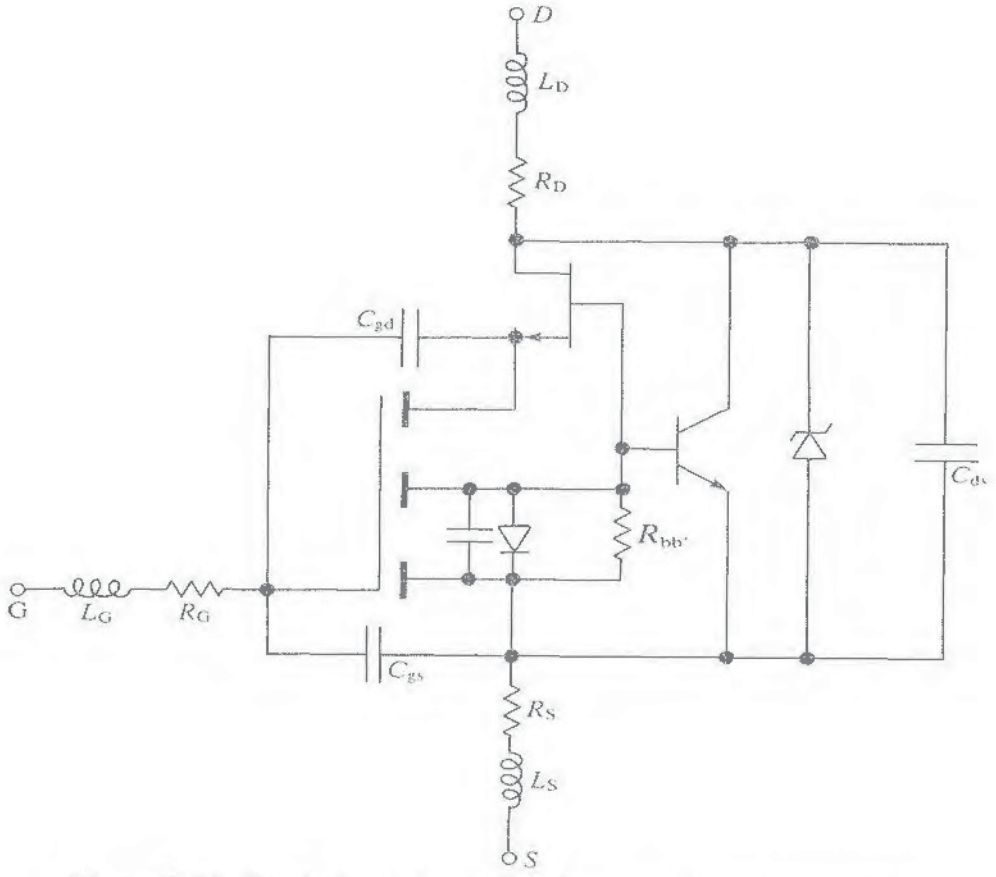


Figure 3.16. Equivalent circuit showing parasitic components.

In Figure 3.15 the various internal parasitic components of a VDMOS FET are shown, drawn onto the usual section through the active region. An equivalent circuit incorporating these and other parasitic elements is shown in Figure 3.16. Passive components include the following:

1. The various resistive elements contributing to $R_{DS(on)}$.
2. The inductance of each of the pin connections.
3. The capacitances between the gate and the source metallization, and between the gate and the source diffusion.
4. The capacitances between the gate and the body region, and between the gate and the epilayer.
5. The capacitances associated with the various pn junctions.
6. The pn junctions show, in addition, their normal diode behavior.

The main active components are:

7. The n-channel junction field-effect transistor (JFET) that forms in the epilayer, in between the channel diffusions.
8. The npn bipolar junction transistor (BJT) formed between tbe source, the body, and the drain. Under static conditions the base and emitter of the BJT are shorted, leaving only the body–drain diode effective. However, under transient conditions and in avalanche breakdown, the transistor may be activated, and this can seriously degrade the overall performance of the MOSFET.

We consider each of these groups of parasitic components in turn.

3.6.2. Resistance, Inductance and the Parasitic JFET

The causes of parasitic resistance in the path of the drain current have been discussed at length already in this chapter. We will not pursue them further, except in so far as they are influenced by the parasitic JFET, which we deal with next. The potential of this component to affect, adversely, many of the MOSFET static characteristics influences a number of design decisions, and has caused us to mention it several times already. The JFET action occurs in the region between the p diffusions. The p-doped body regions act as the gate, and the throat in the n-type drain drift region as the channel. The effect is more severe in high-voltage devices, because the epilayer is more lightly doped and the depletion layers that form at the body-drain junction extend further into the drain.

In Figure 3.12 the additional resistance caused by the JFET can be seen to account for 25% of the total $R_{DS(on)}$ of a high voltage device. At high ON-state currents, the increased voltage in the drain drift region causes an expansion of the depletion layers and a further increase in $R_{DS(on)}$. This is shown in Figure 3.17. In a poorly designed device, it is possible for pinchoff to occur in the JFET rather than in the MOSFET, under some conditions.

The acceptable spacing b between the channel diffusions is mainly determined by the extent, in the ON state, of these depletion layers. As a result, the maximum

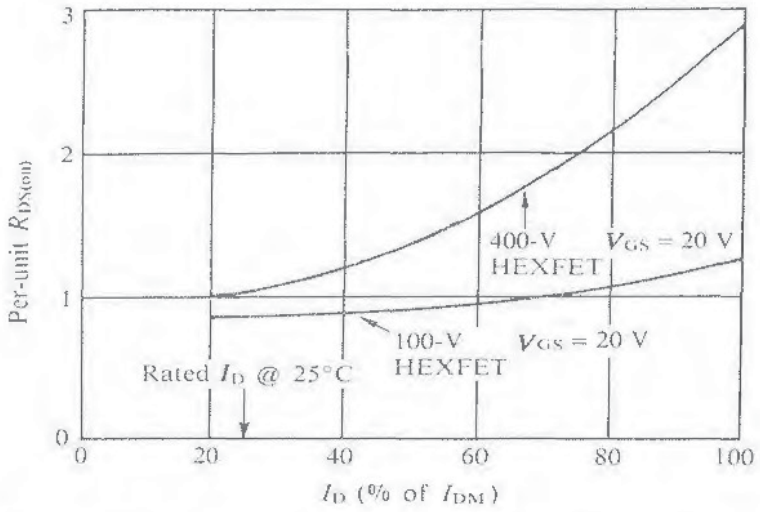
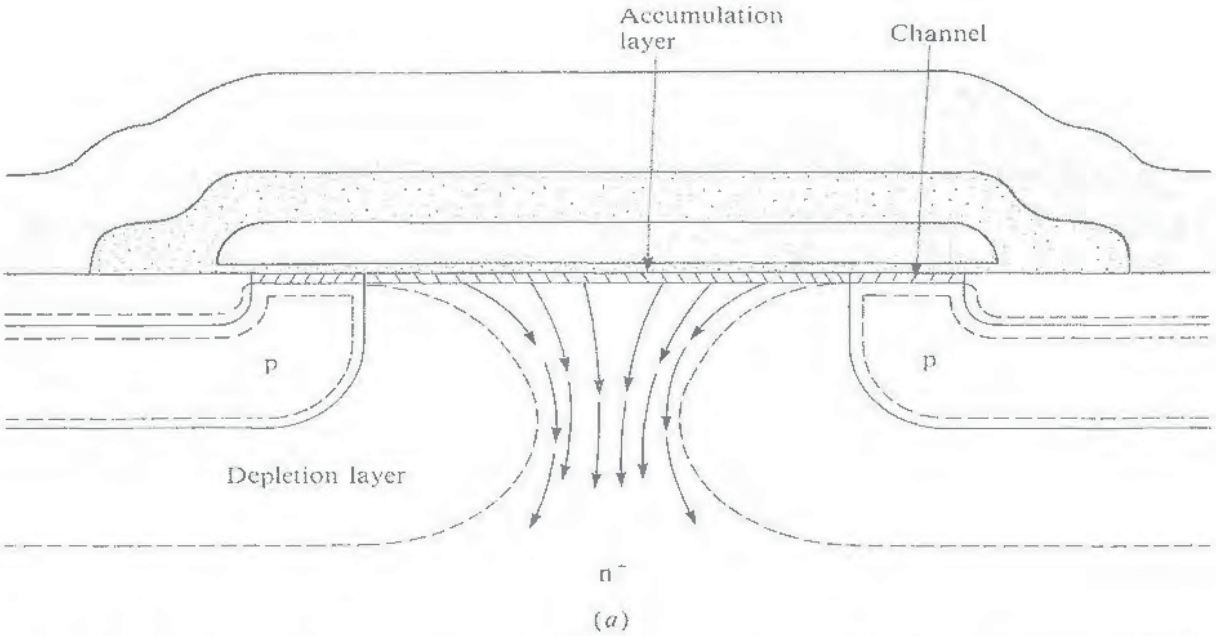
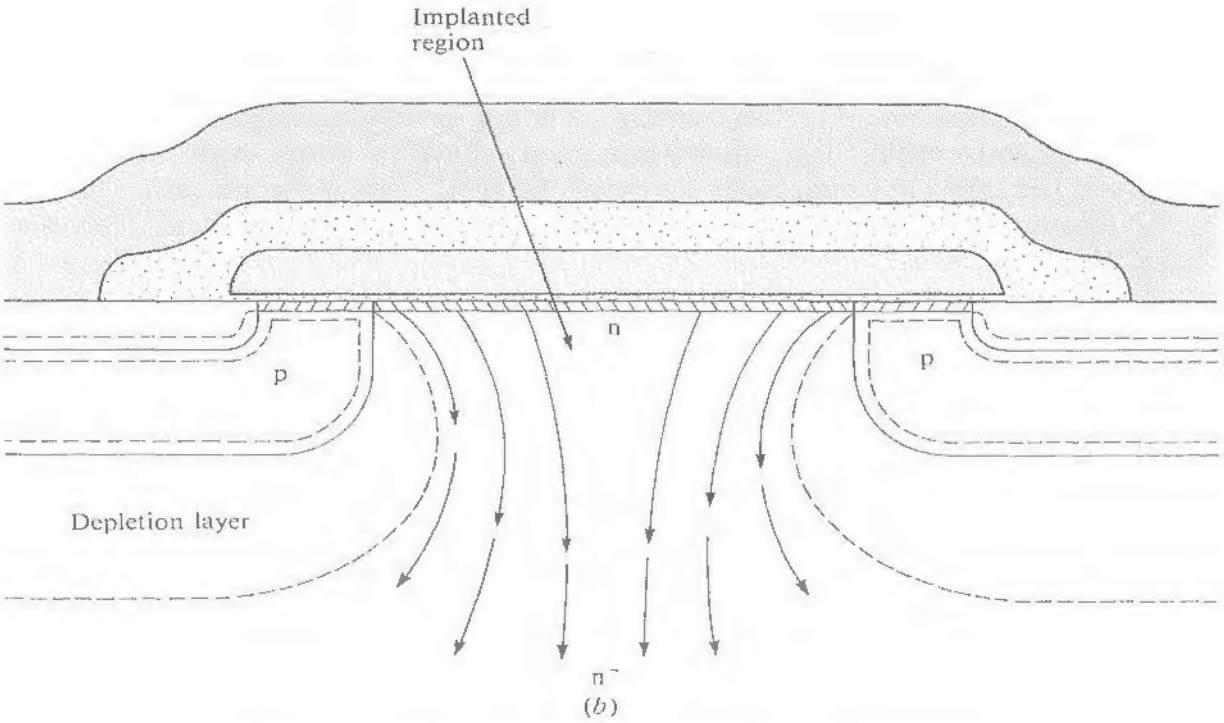


Figure 3.17. Increase in $R_{DS(on)}$ at high drain currents. The current is normalized to I_{DM}, the maximum rated pulse current, which is higher for the lower-voltage device (e.g. 108 A compared to 40 A). The increase in $R_{DS(on)}$ at higher currents is mainly the result of JFET action in the throat region as V_{DS} increases. It can be seen that the increase is much greater in the case of the higher-voltage device, because the epilayer is more lightly doped and the depletion regions expand further into it.

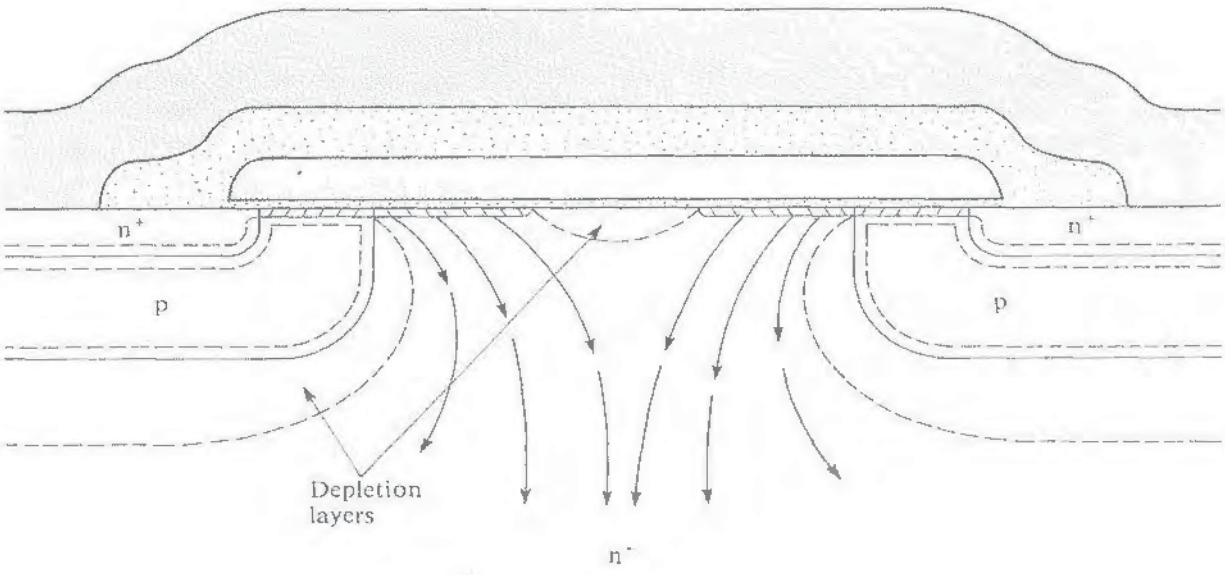


(*a*)

Figure 3.18. Effect of drain implant in reducing JFET action. (*a*) Depletion layers pinch off the channel through the drain neck region in a uniformly doped epilayer. The MOSFET is fully ON. (*b*) Depletion-layer thickness is reduced when the doping level under the gate oxide is increased. (*c*) Under some conditions the potential at the semiconductor surface in the middle of the throat region may exceed V_{GS}. Then a depletion layer forms there too, further constricting the current path.



(*b*)



(*c*)

Figure 3.18. (*Continued*)

usable cell density is strictly limited, as discussed in Appendix 5. Increasing the donor doping concentration in the drain throats, as shown in Figure 3.13, has two beneficial effects. The conductivity is improved, and the widths of the depletion layers are reduced, so increasing the cross-sectional area available for conduction. This is illustrated in Figure 3.18. The consequence is that the cell size can be reduced and the resulting benefits of increased g_{fs} and lower $R_{DS(on)}$ can be

realized. Of course, the level of the donor concentration in the drain immediately under the gate oxide has to be finely judged. If it is too high, the avalanche breakdown voltage is reduced and channel-length modulation may occur.

Of the lead inductances, that of the gate is not usually significant, but those of the drain, and particularly the source, may have a considerable effect on switching behavior and high-frequency performance. This is discussed in Chapter 4. Values can be minimized by careful encapsulation and design. They lie, typically, in the range 5 to 15 nH.

3.6.3. Capacitance

The physical capacitances shown in Figure 3.15 each contribute to the equivalent-circuit capacitances shown in Figure 3.16. We need to relate them to the terminal capacitances, C_{iss}, C_{oss}, and C_{rss}, that are specified on device data sheets. These are normally measured at 1 MHz, using the circuits shown in Figures A9.4, A9.5, and A9.6 of Appendix 9. Each is a function of the dc drain–source voltage, as shown in Figure 10 of the data sheet of Appendix 7. Values are usually specified for $V_{DS} = 25$ V and $V_{GS} = 0$ V. The three capacitances are related as follows:

1. The input capacitance, C_{iss}, is the parallel sum of C_{gd} and C_{gs}. Thus,

$$C_{iss} = C_{gd} + C_{gs} \tag{3.25}$$

2. The output capacitance, C_{oss}, represents the total capacitance between drain and source with the gate and source shorted together, that is,

$$C_{oss} = C_{ds} + C_{gd} \tag{3.26}$$

   On many early data sheets, an incorrect relationship is given, implying that C_{oss} is the sum of C_{ds} and the series combination of C_{gs} and C_{gd}. The published curves, however, were always obtained using the circuit of Figure A9.5, and are thus correct.

3. The reverse transfer or Miller capacitance, C_{rss}, is simply the gate–drain capacitance with $V_{GS} = 0$. That is,

$$C_{rss} = C_{gd} \tag{3.27}$$

From the discussion in Section 2.4, we can see that C_{gd} is made up of the series combination of the gate oxide capacitance and the capacitance of the drain depletion layer beneath the gate oxide. This latter component exists only when the potential of the drain surface region (B) is more positive than that of the gate. When the gate is more positive, an accumulation layer forms at the semiconductor surface and $C_{gd} \approx C_{ox}A_G$, where $C_{ox} = \epsilon_0\epsilon_{ox}/t_{ox}$, as before, and A_G is the area of the gate. As the drain potential rises above that of the gate, the value of C_{gd} falls quite sharply. It continues to fall progressively as the depletion layer extends deeper into the epilayer. This is shown in Figure 2.11. Depletion-layer capacitance is discussed in Appendix 1. Its variation with the dc bias voltage is shown to

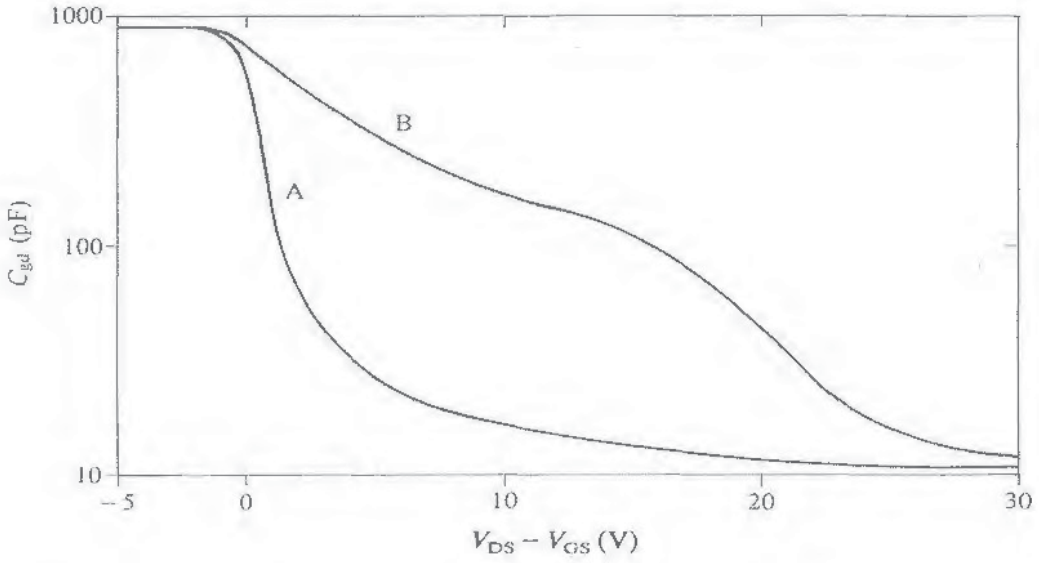


Figure 3.19. Effect of drain implant on the variation of C_{gd} with V_{DS}. Curve A represents a device with a uniformly doped epitaxial layer. Curve B is from one in which the conductivity under the gate oxide has been increased by an implant of 10^{16} phosphorus atoms per m^2 (10^{12} cm^{-2}).

be a function of the doping profile. The detailed variation of C_{gd} with V_{DS}, given that $V_{GS} = 0$, is therefore a function of the impurity concentration profile in the epilayer. This can be seen in Figure 3.19. Really we are applying the well-known capacitance–voltage (C–V) profiling technique to the whole device.

Physically, the gate–source capacitance C_{gs} comprises the overlap capacitance, the gate–channel capacitance, and the capacitance through the field oxide of the source overlay. It is largely independent of bias.

The principal component of C_{ds} is the capacitance of the junction between the p diffusion and the epilayer. This varies inversely with the square root of the drain–source voltage, as indicated by Equation (A1.18). Thus, at low drain voltages, particularly when $V_{GS} > V_{DS}$, $C_{oss} = C_{ds} + C_{gd}$ becomes quite large.

Because of their considerable variation with the biasing conditions, knowledge of C_{iss}, C_{oss}, and C_{rss} at any one state of bias is of limited value to the circuit designer. Graphs of the three capacitances versus V_{DS} are normally provided, but, unless these are extended to negative values of V_{GD}, they do not show the full extent of the capacitance variation. To overcome this difficulty, it has become normal practice to specify the total amount of charge needed to change the gate voltage from 0 to 10 V while the device switches a prescribed current. The current chosen is normally somewhat greater than the maximum dc rating. The details and implications of this are discussed further in Chapter 4.

3.6.4. The Body–Drain Diode and the Bipolar Junction Transistor

Finally, we come to the bipolar junction transistor structure that is formed by the n^+ source, the p body, and the n^- epilayer. Under normal, static conditions the

emitter and the base of the BJT (the source and the body region) are shorted, leaving only the effect of the body–drain diode.

When the MOSFET is reverse biased, the body–drain diode is set into forward conduction. Its current rating is comparable to that of the MOSFET itself. So that this should not lead to excessive dissipation, or to problems in external circuits, its forward voltage ($V_{SD(on)}$) needs to be kept low. Typical values at the full rated current are from 2 to 2.5 V. This reverse-conduction property is valuable in a power switching device. It is particularly necessary in circuits such as dc-to-ac inverters and dc-to-dc choppers, when they are used in applications such as motor speed control, as discussed in Chapter 12.

As well as a low on-state voltage, the body–drain diode must exhibit a rapid, but not too abrupt, recovery waveform, and low reverse recovered charge Q_{RR}. Occasionally, heavy-metal doping (gold or platinum) or electron bombardment have been used to speed up the recovery of the body–drain diode. These are normal techniques used with other types of device. They introduce recombination centers by creating trapping levels in the middle of the band gap, and so kill the carrier lifetime. Unfortunately, this increases both $V_{SD(on)}$ and $R_{DS(on)}$ at room temperature. At elevated temperatures, the trapping levels are no longer effective, and both the lifetime and the ON resistance revert to their normal values. This means that the temperature coefficient of $R_{DS(on)}$ is lower than that of a normal device. The effect on Q_{RR} is discussed in Chapter 13; see Figure 13.22, in particular.

It has been shown [9] that a careful adjustment of the energy and dose of electron irradiation can give valuable reductions in the recovery time and the recovered charge (and so reduce the transient dissipation), without degrading the other parameters significantly. In Ref. 9, 8 Mrad at 3 MeV was reported to be effective. Another adverse effect is a reduction caused to the threshold voltage by the creation of positive charge in the gate oxide. To reduce this charge, it is necessary to anneal the device after irradiation for several hours at about 140°C. This normally restores the threshold voltage to a reasonable value, but leakage currents remain appreciably higher than before irradiation.

The reverse avalanche-breakdown characteristics of the body–drain diode determine the forward voltage rating of the MOSFET, BV_{DSS}. In modern devices, the diode junction is designed to withstand the passage of substantial current under conditions of controlled avalanche breakdown. This enables forward transients to be quenched, and inductive energy to be dissipated, without the use of additional, external clamping and snubber components. To achieve this it is necessary for the breakdown to be very evenly distributed over the p^+ region of the junction and, in particular, to be evenly distributed from cell to cell. This is illustrated in Figure 3.20.

Many devices now have single-shot and repetitive pulsed ratings for the energy (charge $\times$ BV_{DSS}) that they can safely dissipate in this way. An example of a device being driven into controlled breakdown by an inductive load is shown in Figure 3.21. The device under test is an IRF 460, rated at 500 V, 24 A. Once the MOSFET is turned on, the current rises approximately linearly with time at a rate determined by the voltage across the inductance (17 A/ms, that is, 100 V across 6 mH). When the transistor is turned off, the voltage across it rises to the breakdown voltage V_B, which is 550 V. This causes the current in the inductance

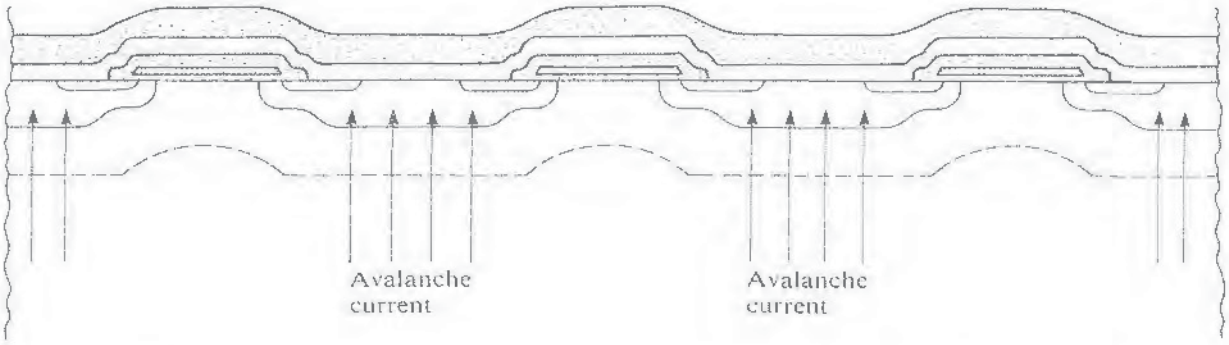


Figure 3.20. Avalanche breakdown of the body–drain diode. The diagram shows the uniform breakdown of the p^+n junction required in a good device. The shielding effect of adjacent cells mitigates the field concentration at the edges and corners of the p diffusion.

to fall at the rate (550 V)/(6 mH) = 90 A/ms. Thus the energy stored in the inductance at the moment of turnoff is dissipated in the transistor during the fall of current. In this case it amounts to about 1.2 J.

The body region has a lateral resistance from the contact to the shoulder that is known as the "pinch-base" resistance, $R_{bb'}$. Current flowing laterally through $R_{bb'}$, as shown in Figure 3.22, causes an ohmic voltage drop. If this voltage exceeds about 0.6 V, it will initiate transistor action in the parasitic BJT. This is because it forward-biases the source–body junction and causes the injection of electrons from the source into the body region. These then become the emitter and base of the parasitic bipolar transistor.

The value of $R_{bb'}$ is minimized by reducing the length of the shoulder linking the channel region to the p^+ diffusion. The depth of the shoulder is determined by the relative diffusion depths of the body and the source, which in turn determine the channel length. With a channel length of about 1 μm, the depth of the p diffusion is rather less than 2 μm. With a maximum boron doping density of about $1 \times 10^{23}\ m^{-3}$ ($1 \times 10^{17}\ cm^{-3}$), the body-region resistivity has a minimum value of about $2 \times 10^{-3}\ \Omega$ m (0.2 Ω cm). The surface resistance of the shoulder is

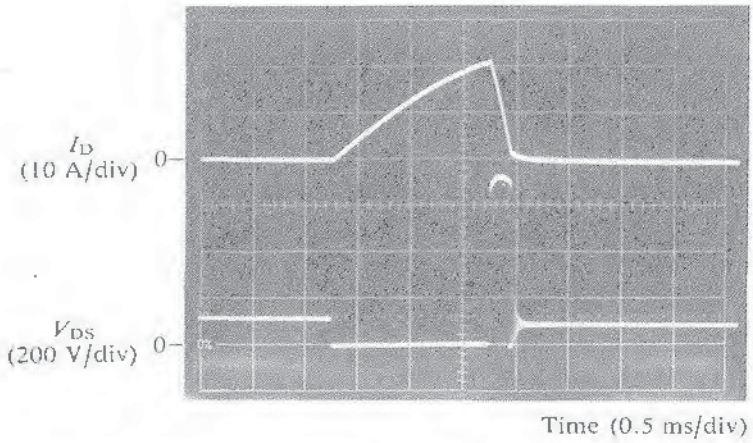


Figure 3.21. Example of controlled avalanche breakdown immediately following turn-off.

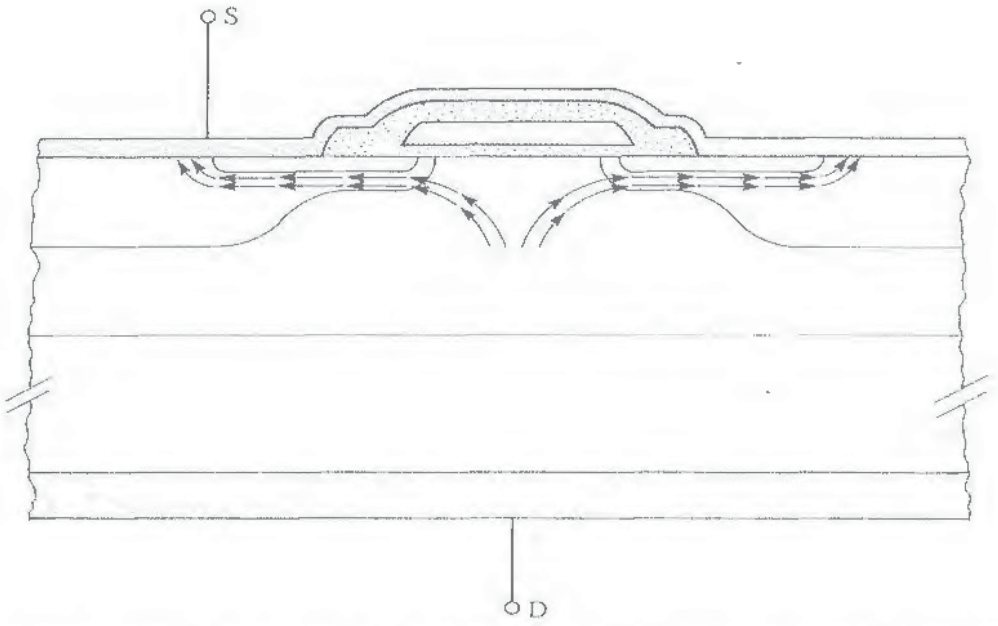
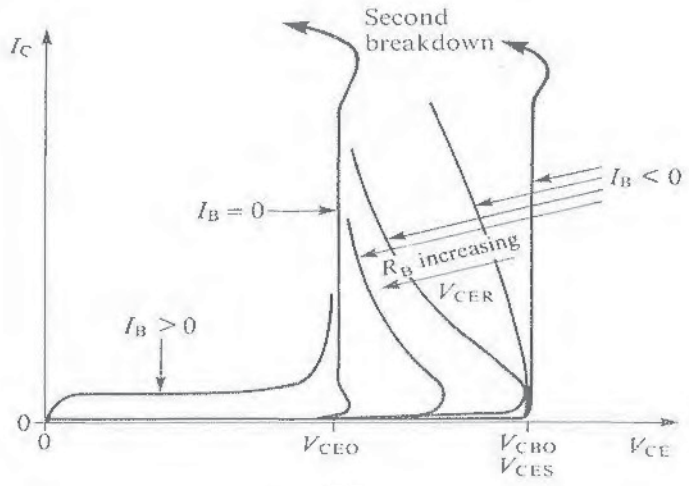


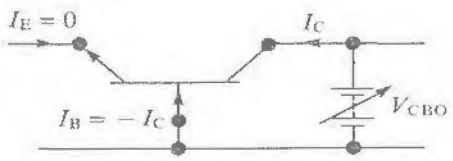
Figure 3.22. Current flow through the pinch-base resistance. This creates the danger that the parasitic BJT may be activated.

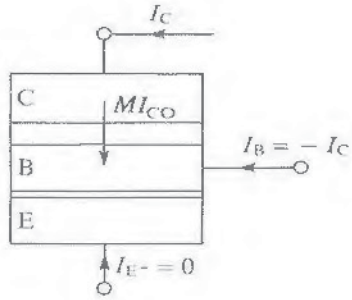
Figure 3.23. The avalanche-breakdown behavior of a bipolar junction transistor. (*a*) Typical characteristics; (*b*) common-base configuration (V_{CBO}); (*c*) common-emitter configuration (V_{CEO}); (*d*) emitter and base connected (V_{CER}). Avalanche breakdown is a regenerative process, because some of the electrons and holes created by impact ionization pass back through the high-field region and themselves create more carriers. The breakdown voltage of a junction, V_B, is the voltage at which the carrier avalanche multiplication factor $M = \infty$. As the temperature rises, M decreases, so V_B increases. In a BJT, the breakdown voltage V_{CBO}, is that of the collector–base junction with the emitter open-circuited. A similar value is obtained when the base and emitter are shorted, but it is then called V_{CES}. The breakdown voltage V_{CEO} is that between the collector and the emitter with the base open-circuited. It is significantly lower than V_{CBO}, because of a further process of regeneration. When holes created in the multiplication process at the collector junction pass back across the base and enter the emitter, they may cause the injection of further electrons into the base. A proportion of these reach the collector junction, where they in turn suffer multiplication. Let a fraction α_0 of the current crossing the emitter junction be electrons which reach the collector junction and are multiplied. The current in the transistor can increase without limit when $1 - \alpha_0 M = 0$, that is, when $M = 1/\alpha_0$. With $V_{BE} < 0.6$ V, $\alpha_0 \approx 0$, but the buildup of carriers in the base, once avalanche multiplication becomes appreciable causes the base–emitter junction to become forward biased. Transistor action is initiated and α_0 increases rapidly. The minimum value of V_{CEO} corresponds to the maximum value of α_0, and is known as $V_{CEO(sus)}$. When a finite resistance connects base and emitter, the breakdown curve (V_{CER}) lies between these two extremes. At low currents α_0 is low and V_{CER} approximates to V_{CBO}. As the current increases, α_0 increases and V_{CER} approaches V_{CEO}. Only if the emitter is open-circuited or shorted to the base can $V_{CBO} = V_{CES}$ be sustained. Further increase in current, if maintained, eventually leads to second breakdown and device destruction.



(*a*)

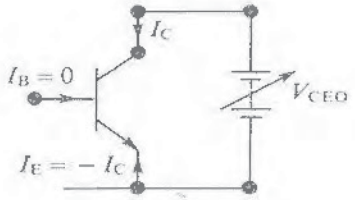
(*b*)

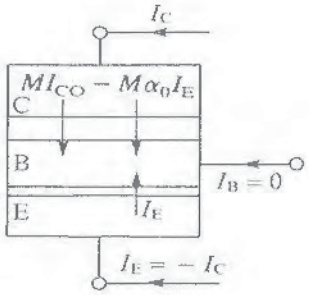




$$I_C = MI_{CO} = -I_B$$

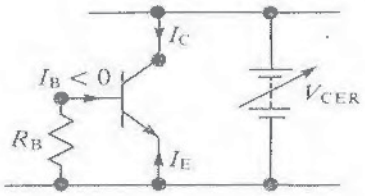
(*c*)

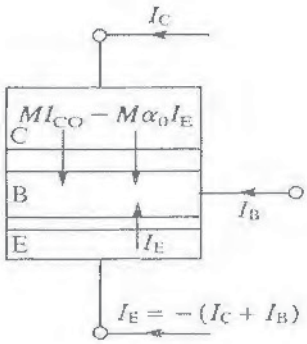




$$I_C = \frac{MI_{CO}}{1 - \alpha_0 M} = -I_E$$

(*d*)





$$I_C = \frac{M(\alpha_0 I_B + I_{CO})}{1 - \alpha_0 M}$$

thus rather more than 1 kΩ/square. With the length of the shoulder about 6 μm and a gate width of about 1 m, the overall value of $R_{bb'}$ is in the region of 6 mΩ, say 6 to 10 mΩ. It would thus require a uniformly distributed current of 60 to 100 A to activate the bipolar transistor. However, any mask misalignment that causes the p^+ diffusion to be displaced also increases the length of the shoulder on one side and so increases $R_{bb'}$ locally. It is important to understand that it only requires the BJT to he turned on in one part of one cell for this region to hog all the current passing through the device. The rise of temperature caused by the resulting dissipation may induce current crowding and lead on to thermal runaway and second breakdown failure. It is a comment on their design quality and process controllability that manufacturers can now guarantee ratings at which such failures do not occur.

The use of the p^+ diffusion to connect the body region back to the source contact can be seen to be an important part of the VDMOS FET design. It reduces $R_{bb'}$ and determines the breakdown voltage of the body–drain diode. Its vertical and lateral doping profiles have to be carefully controlled so that no adverse effects on the breakdown voltage or the properties of the channel region occur.

The presence of the parasitic bipolar junction transistor, with its base and emitter connected by $R_{bb'}$, lowers the breakdown voltage of the body–source junction. This effect is characteristic of bipolar junction transistors in general. It is illustrated in Figure 3.23. There it can be seen that electrons injected into the base (body) from the emitter (source) give rise to an additional regenerative effect. This is serious because field concentration at tbe edges and corners of the shoulder of the body region would normally reduce the breakdown voltage there anyway. Fortunately, as was mentioned in the previous section and illustrated in Figure 3.20, the proximity of adjacent cells helps to minimize this field concentration. If, as a result of current flowing through $R_{bb'}$, the breakdown voltage is locally reduced to V_{CEO}, and this is below the supply voltage, the MOSFET will be destroyed.

Under transient conditions, the parasitic BJT may be activated in two further ways. First, changes in the potential difference across the body–drain diode require its self-capacitance to be charged and discharged, and the resulting capacitative currents flow through $R_{bb'}$. This could, in principle lead to a forward dV/dt failure. In practice, even taking $C_{ds} = 1$ nF, which would be typical for V_{DS} small, the rate of change of V_{DS} would have to be in the region of 60 to 100 V/ns for $R_{bb'} C_{ds}\, dV_{DS}/dt$ to exceed 0.6 V. This is not a problem with the high degree of cell uniformity achieved in modern devices. The second way of activating the BJT also follows turnoff. Excess carriers left from the previous conduction phase are swept out from the regions either side of the body–drain junction, as the depletion layers expand. This adds greatly to the normal capacitance current. Such carriers are most likely to remain in the region under the gate oxide. From there they are swept into the shoulder of the body region and have to pass the whole of $R_{bb'}$ to reach the source contact. They are, of course, blocked by the source–body junction, until this becomes forward biased—the effect we are seeking to avoid! Even greater care has to be taken over regions at the periphery of the chip and adjacent to the gate vias. Unless special precautions are taken, carriers tend to build up in these "dead spaces" and remain there longer than in the more confined regions of the active parts of the device. Figure 13.24, in

Section 13.13, shows an IRF 530 turning off in an inductive switching circuit. Following the recovery of the body–drain diode, the device is subjected to a fast forward dV_{DS}/dt transient. This is safely clamped at 130 V by a controlled breakdown. More detailed discussions of the origin of the waveforms observed during turn-on and turn-off are given in Chapter 4 and in later chapters in the context of specific applications.

3.7. TEMPERATURE EFFECTS

In Chapter 2 we showed that the threshold voltage would be expected to decrease with an increase in temperature. Carrier mobility and saturation velocity also decrease as the temperature rises, and in the bulk semiconductor regions the resistivity increases. For a given device, in which V_{GS} is kept constant, the net consequence of a rise in temperature resulting from these various effects is as follows:

1. In the subthreshold region ($V < V_T$) the drain current after pinchoff increases. Before pinchoff the increase in the various parasitic resistances tends to reduce the slope of the I_D-versus-V_{DS} characteristic.
2. In the "normal" operating region described by Equations (3.10) and (3.11) in the simple theory, or by their extensions discussed in Appendix 4, the decrease of the carrier mobility more than offsets the consequences of the lowering of V_T. As a result, I_D decreases above and below pinchoff.
3. In the velocity saturation region $I_{D(Sat)}$ again decreases, because of the fall in v_s.

The combination of all these effects on the static characteristics is shown in Figure 3.24. In particular, the transfer characteristics like those of Figure 2 in Appendix 7, which are given in most data sheets, demonstrate how $I_{D(Sat)}$ increases with temperature below a certain current level and decreases with increasing temperature at higher currents. This is important in the bias stabilization of linear amplifiers, and is discussed further in Section 15.2.2. In practice, the normal operating currents of LDMOS FETs lie much further above the threshold voltage than those of VDMOS FETs. The result is that LDMOS FETs are less sensitive to the variation of V_T, and the crossover current occurs at a lower level. This is shown in Figure 15.7, in Section 15.2.2.

The effect of temperature on g_{fs} is shown by Figure 6 of Appendix 7. The reduction in the high-current region, when g_{fs} is constant, is brought about by the lowering of v_s as the temperature rises. The increase in $R_{DS(on)}$ with temperature is shown in Figure 9 of Appendix 7. The effect of this is compounded, because the increase in dissipation that results causes a further temperature rise, and so on.

Operating at elevated temperature increases the junction breakdown voltage, by about 1% for each 10°C rise. It also is likely to reduce the expected device lifetime, as described in Chapters 5 and 17. The device capacitances discussed in the previous section are largely unaffected by changes in temperature.

At higher temperatures, a lower forward bias voltage is needed to support a given current flowing across any of the pn junctions. Thus, $V_{SD(on)}$ is lower.

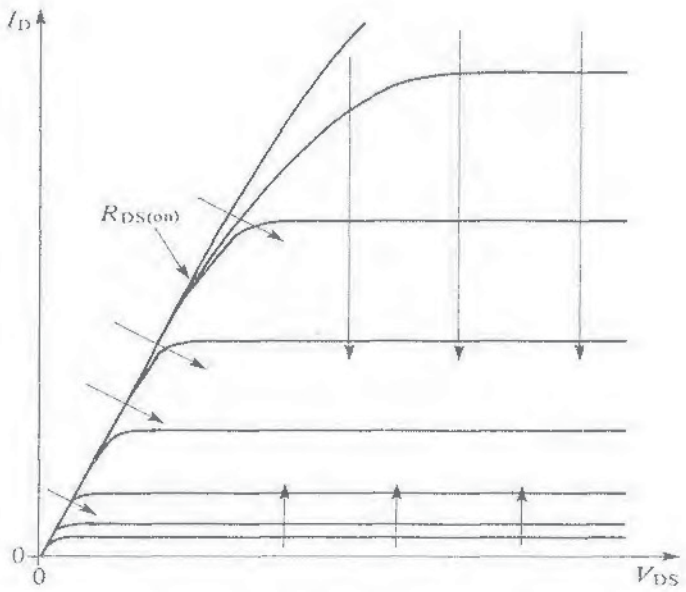


Figure 3.24. Effect of temperature on device characteristics. Arrows indicate the shift of the different parts of the characteristics with an increase in temperature.

Observation of $V_{SD(on)}$, V_T, or $R_{DS(on)}$ permits an indication of the temperature in the active parts of the transistor to be obtained, as discussed in Section 17.2.4.

The parasitic bipolar junction transistor is more easily turned on at high temperatures, because the pinch–base resistance $R_{bb'}$, is increased, and the voltage needed to forward-bias the emitter–base junction is reduced. At the same time the BJT is more rapidly sent into second breakdown failure.

3.8. p-CHANNEL AND OTHER DEVICES

Often considerable advantage in circuit simplicity results when complementary p-channel and n-channel devices are both available. This is as true for power devices as it is for CMOS (in comparison with NMOS) integrated circuits. In silicon, holes have lower mohility as well as a lower saturation drift velocity. As a result, p-channel devices show a higher $R_{DS(on)}$ (about a factor of 2 higher) and a lower value of g_{fs} than n-channel devices having the same silicon area. These important considerations apart, p-channel devices can be made simply by reversing the dopant types, as illustrated in Figure 3.25. Typical characteristics are shown in Figure 3.26.

One advantage of the p-channel device is that the body region is n-type and so has a lower resistivity for a given doping concentration. The pinch–base resistance, $R_{bb'}$, is lower in consequence, and the avalanche characteristics are better because the pnp transistor is less prone to turn on. A further disadvantage, though, is the difficulty of making a good ohmic contact between the source metallization and the body diffusion. Because of the constraints on the body

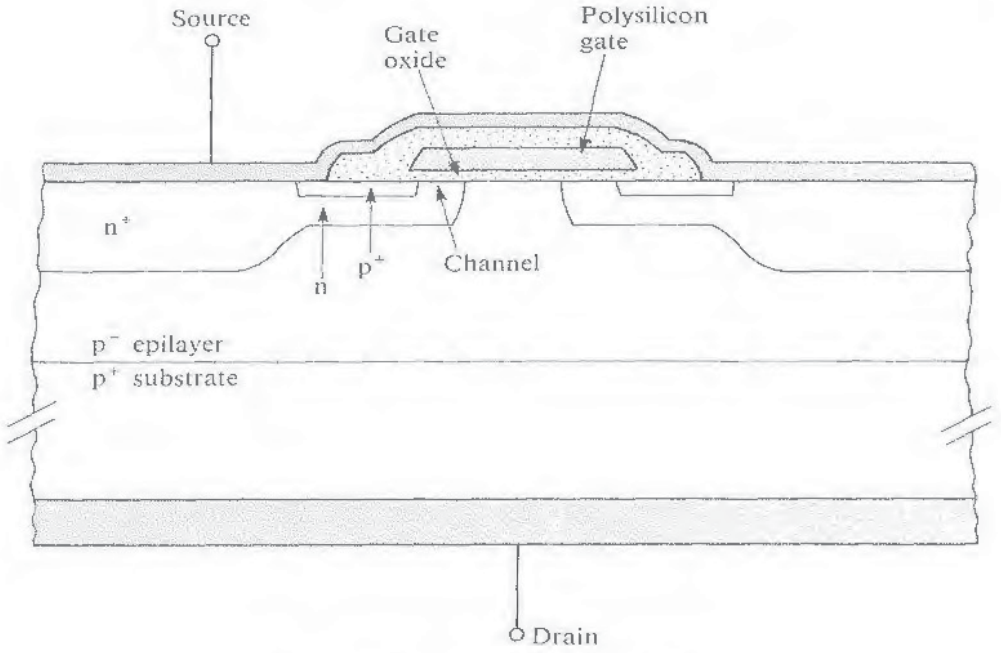


Figure 3.25. A p-channel device.

region doping levels, this contact normally forms a Schottky diode. Its reverse breakdown voltage is in series with the forward voltage of the body–drain diode when reverse current is conducted, so the drain–source voltage drop under these conditions ($V_{SD(on)}$) is high, typically 5 to 6 V.

Many variants on the basic VDMOS FET have been devised and are more or less generally available. These include the IGBT mentioned briefly in Section 1.3, and several other MOS/bipolar devices. They are discussed further in Chapter 20. Current-sensing MOSFETs and radiation-hardened devices are discussed in Chapter 19.

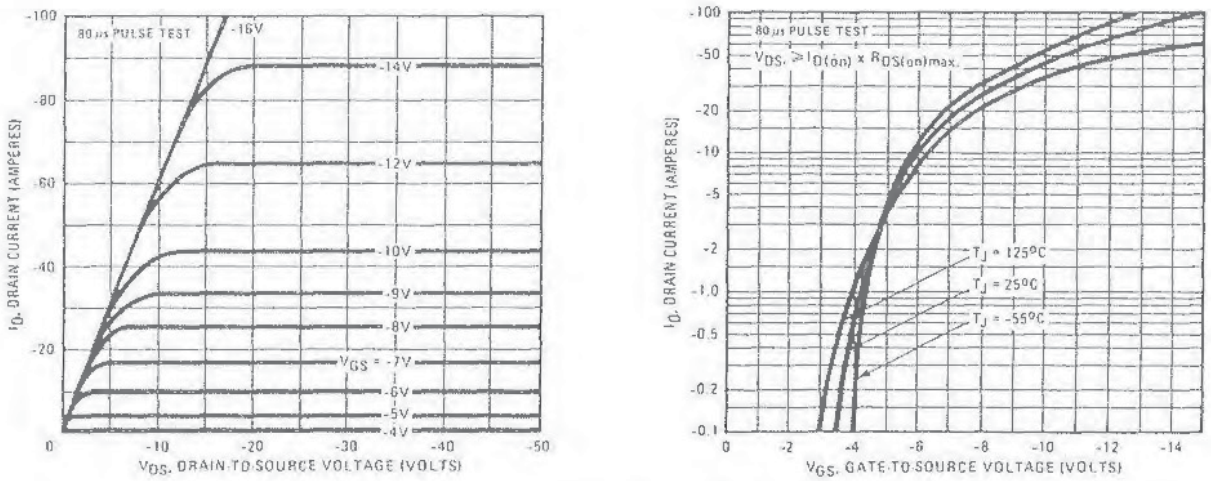


Figure 3.26. Characteristics of a p-channel device (an IRF 9540).

SUMMARY

The total drain–source voltage of the conducting MOSFET comprises the ohmic and the nonlinear voltages illustrated in Figure 3.2. For high-voltage devices in the ON state, the resistance of the epitaxial layer is dominant, and $R_{DS(on)}$ varies approximately as $V_B^{2.5}$. The theory governing these parameters is discussed in Appendix 6. Devices benefit from an increase in the doping level in the drain under the gate oxide. For low-voltage devices it is most important to minimize the contact and lead resistances, and a reduction in the device feature size is beneficial.

A simple charge-control model predicts the following dependence of I_D on V_{GS} and V_{Ch}:

For $V_{GS} > V_T$ and $V_{GS} - V_T > V_{Ch}$,

$$I_D = \mu_e \frac{w}{l} C_{ox}[(V_{GS} - V_T)V_{Ch} - \tfrac{1}{2}V_{Ch}^2] \tag{3.10}$$

For $V_{GS} > V_T$ and $V_{GS} - V_T < V_{Ch}$,

$$I_D = I_{D(Sat)} = \tfrac{1}{2}\mu_e \frac{w}{l} C_{ox}(V_{GS} - V_T)^2 \tag{3.11}$$

At high levels of drain current, electron velocity saturation effects lead to a predicted linear variation of drain current with V_{GS},

$$I_{D(Sat)} = \tfrac{1}{2}C_{ox} w(V_{GS} - V_T)\upsilon_s \tag{3.14}$$

Close to threshold, $V_{GS} \approx V_T$, there is an exponential variation of $I_{D(Sat)}$ with V_{GS},

$$I_{D(Sat)} \propto \exp(eV_{GS}/\alpha kT) \tag{3.15}$$

This simple theory can be fitted quite closely to the experimental characteristics of devices, but cannot easily be reconciled with their known physical characteristics. On closer examination it is seen to contain a number of self-contradictions. The more sophisticated theory discussed in Appendix 4 avoids these, but without the introduction of a field-dependent mobility, serious discrepancies with the experimental results remain.

Transconductance is maximized by obtaining the maximum gate width from a given area of silicon. This is inversely proportional to feature size, and is aided by the use of a cellular geometry. A discussion of the implications of some of the different cell patterns that have been used or proposed is given in Appendix 5.

A rise of temperature causes an increase in $R_{DS(on)}$. In the subthreshold region $I_{D(Sat)}$ increases with temperature, because of the reduction in the threshold voltage. At higher current levels the reduction in the electron mobility with temperature offsets this, and above a certain value of $I_{D(Sat)}$, where the temperature coefficient is zero, it decreases. These effects are illustrated in Figure 3.24.

Holes have a lower mobility and a lower saturation drift velocity than electrons, so p-channel devices have a lower transconductance and a higher $R_{DS(on)}$ than otherwise similar n-channel devices. Their voltage drop in reverse conduction is higher, but their reverse breakdown performance is better.

Several parasitic components influence the behaviour of power MOSFETs. As well as the normal parasitic resistance, inductance and capacitance associated with the semiconductor, the leads, and the packaging, several parasitic active components are also inherently present. These include the body–drain diode, the source–body–drain bipolar junction transistor, and a series junction field-effect transistor in the drain drift region. These are illustrated in Figures 3.15 and 3.16. These parasitics limit device performance and set constraints on device design. The body–drain diode is forward biased by any inverse voltage that develops across the transistor and so provides a reverse conduction path. Modern designs of transistors allow the body–drain diode to avalanche safely when reverse biased and so provide a means of dissipating energy stored in inductive switching circuits.

REFERENCES

1. H. C. Pao and C. T. Sah, "Effects of diffusion current on characteristics of metal–oxide (insulator)–semiconductor transistors." *Solid-State Electron.* **9,** 927–937 (1966).
2. R. F. Pierret and J. A. Shields, "Simplified long-channel MOSFET theory." *Solid-State Electron.* **26,** 143–147 (1983).
3. J. A. Cooper, Jr. and D. F. Nelson, "High-field drift velocity of electrons at the Si–SiO_2 interface as determined by a time-of-flight technique." *J. Appl. Phys.* **54,** 1445–1456 (1983).
4. G. Baccarani, M. R. Wordeman, and R. H. Dennard, "Generalised scaling theory and its application to a 1/4 micrometer MOSFET design." *IEEE Trans. Electron Devices* **ED-31,** 452–462 (1984).
5. C. Hu, M.-H. Chi, and V. M. Patel, "Optimum design of power MOSFET's." *IEEE Trans. Electron Devices* **ED-31,** 1693–1700 (1984).
6. H.-R. Chang and B. J. Baliga, "Impact of cell breakdown upon power DMOSFET on-resistance." *IEEE Trans. Electron Devices* **ED-34,** 2360 (1987).
7. V. A. K. Temple, "Ideal FET doping profile." *IEEE Trans. Electron Devices* **ED-30,** 619–626 (1983).
8. A. Lidow and T. Herman, "Process for manufacture of high power MOSFET with laterally distributed high carrier density beneath the gate." U.S. Pat. 4,593,302 (1986).
9. B. J. Baliga and J. P. Walden, "Improving the reverse recovery of power MOSFET integral diodes by electron irradiation." *Solid-State Electron.* **26,** 1133–1141 (1983).

CHAPTER 5

Fabrication and Reliability

5.1. FABRICATION

5.1.1. A Standard Process for n-Channel Devices

Most VDMOS FETs are fabricated in processes involving several deposition, implantation, and diffusion stages, and some five or six separate masking steps. The photolithography should permit a 2- to 3-μm minimum line width to be defined. This means that whole-wafer proximity printing may be used, together with chemical (wet) etching methods. A number of more modern device designs utilize a smaller cell size, and so demand a finer geometry. This is particularly true for devices intended for low-voltage applications. For these, the difficulty of maintaining the required definition and registration tolerances over the whole slice is overcome by the use of a step-and-repeat masking process. A mask covering a single device or a few devices only, is projected onto the photoresist, usually with a $\times 10$ reduction, and progressively stepped across the whole wafer. This technique has to be used when larger diameter slices (6 in. rather than 4 in.) are used to increase throughput. Likewise, plasma (dry) etching techniques are then needed in order to avoid undercutting and lateral spread during etching. For a comprehensive review of all aspects of fabrication technology, the reader is referred to Ref. 1.

As was made clear in Chapter 1, VDMOS FETs are fabricated on $\langle 100 \rangle$-oriented slices. The slice thickness has to be chosen with some care. Heat dissipated in the active parts of the device has to be removed through the full thickness of the substrate. The thicker the slice, the greater is the thermal resistance of the final device. However, too thin a slice will lead to breakages during manufacture and a lower yield, particularly with larger-diameter slices. Typically, slice thicknesses between 250 and 500 μm (10–20 mils) are used. The slice thickness also adds to $R_{DS(on)}$. However, heavy doping with phosphorus ensures that the resistivity is low. Typical values exceed 10^{24} atoms/m^3 (10^{18} cm^{-3}), giving 2×10^{-4} Ω m (0.02 Ω cm), or less.

After the normal cleaning and polishing routines, the first step is to grow the n-type epitaxial layer in which the device will be formed. Ideally this should be defect-free. Its thickness and doping level are determined by the intended voltage rating of the transistors. For 50-V devices, as little as 10 μm of material may be deposited. The resistivity would then typically be about 0.01 Ω m (1 Ω cm), corresponding to a phosphorus doping level of about 5×10^{21} m^{-3} (5×10^{15} cm^{-3}). The thickness and resistivity of the epitaxial layer increase as the voltage rating

increases. For 500-V devices, the thickness is more likely to be 50 μm and the resistivity 0.2 Ω m (20 Ω cm). The corresponding phosphorus concentration is 2×10^{20} m^{-3} (2×10^{14} cm^{-3}).

The surface is oxidized for the first masking process. This defines the p-type wells that in the final device connect the channel regions back to the source contacts. It also defines the p-type regions that will underlie the gate vias and the bonding pads. Similar p-type regions form the guard rings for the high-voltage structure around the periphery of the chip. Windows are etched through the oxide, and some 10^{18} boron atoms per square meter (10^{14} cm^{-2}) are implanted at low energy (<10 kV) and then diffused in to form a junction at a depth of about 5 μm. In an optimized process, some compensation has to be made for the doping level of the epilayer. The sheet resistance is typically about 200 Ω/square. During the high-temperature treatment, some outdiffusion of dopant from the substrate into the epilayer occurs.

The surface is left oxidized for the second masking sequence, which is the reverse of the first. All the oxide, *except* that covering the p-type wells, is removed. The slice is then ready for a high-energy (>100 kV) implant of phosphorus, everywhere except into the p-type wells. The situation is illustrated in Figure 5.1(a). Some 10^{16} to 10^{17} atoms per square meter are implanted (10^{12} to 10^{13} cm^{-2}). This gives enhanced conductivity to what will become the drain drift regions under the gate oxide.

Growth of the gate oxide is the next stage. The surface is cleaned, except for the thicker layer of oxide that covers the p-type wells. It is then subjected to a dry thermal oxidation, giving a uniform film of high-integrity silicon dioxide, some 50 to 100 nm thick. Onto this, some 0.5 μm of polycrystalline silicon, heavily doped with phosphorus, is deposited. This will form the gate web. Its sheet resistance is usually in the region of 30 Ω/square. Next, a layer of oxide is deposited over the whole surface by low-pressure chemical vapor deposition. This will form the mask for the subsequent, self-aligned diffusions. A cross-section through the device at this stage is shown in Figure 5.1*b*.

The third masking process defines the pattern of cells, hexagonal or square, that covers the active area of the device. The cells are centered on the previously formed p-type wells, which still carry their caps of oxide. The regions within the cell areas are etched down to the silicon surface, except for the thicker oxide covering the p-type wells. At the same time, fingers of the polycrystalline silicon layer are defined for the gate vias, as are annular rings around the periphery of the chip which will form part of the guard ring structure. This stage is shown in Figure 5.1*c*.

The slice is now ready for an implant of about 10^{18} boron atoms per square meter (10^{14} cm^{-2}). This is normally carried out at about 50 kV, and is followed by a drive-in diffusion for up to 1 or 2 hours at around 1200°C. This forms a pn junction with the n-type epitaxial layer at a depth of some 2 to 3 μm. The junction extends about 2 μm under the edge of the mask, because of lateral diffusion. It is this region immediately under the gate oxide that forms the channel region of the FET. Note that the high-temperature treatment drives the earlier phosphorus implant deeper into the epitaxial layer, and causes further outdiffusion of the substrate dopant.

Any thin coating of oxide formed on the exposed silicon surface is cleaned off,

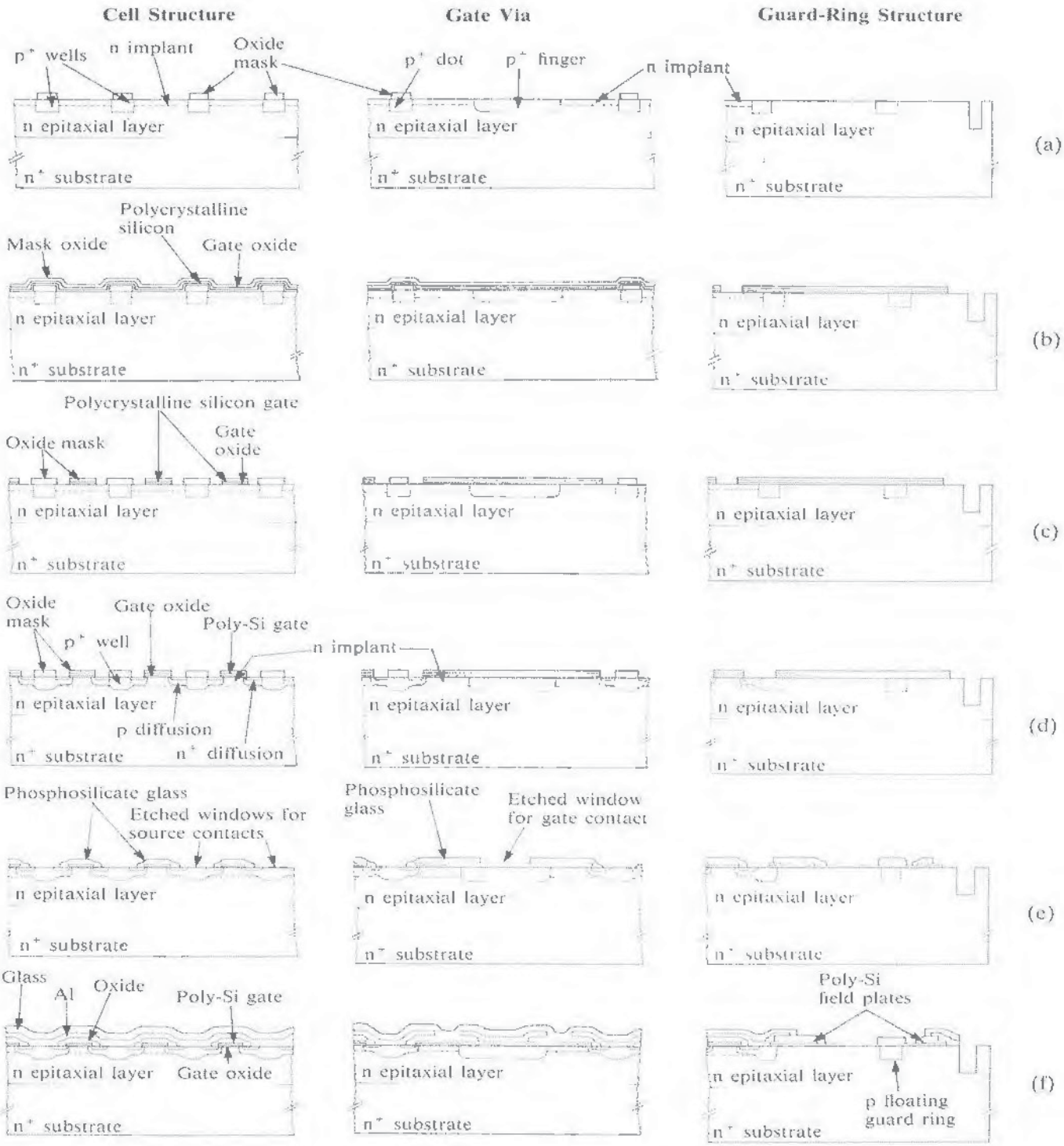


Figure 5.1. A typical processing sequence for a power MOSFET. (*a*) Section through the device after epitaxial growth and p and n implants; (*b*) following the growth of the gate oxide and the deposition of the poly-Si gate; (*c*) following the etching of the cells and the definition of the gate web; (*d*) after two self-aligned diffusions have defined the channel region; (*e*) following the etching of the contact windows; (*f*) the finished device.

and the slice is exposed to a short phosphorus diffusion, typically for about 30 minutes at about 950°C, through the same windows. This forms what will become the source regions. The surface concentration is high, and a relatively abrupt junction is formed at a depth of about 1μm. The junction extends rather less than 1 μm under the edge of the oxide mask.

We have now formed the basic transistor. The channel is defined by the two pn junctions, each following the edge of the mask. The fact that both ends of the channel are defined using the gate structure as a mask ensures precise alignment of gate and channel and accurate control of the channel length. This is the essence of the self-aligned process. The situation at this stage is shown in Figure 5.1*d*.

All the oxide formed during the previous processing (except the gate oxide, which is masked by the poly-Si) is removed using hydrofluoric acid, and a new, thin layer of oxide is grown over the whole surface. Onto this a thicker layer of phosphorus-doped silicon dioxide is deposited. This can easily be made to reflow and so form smooth glassy contours over the edges of the polycrystalline silicon. This aids the continuity of the subsequent metallization stages.

The fourth mask defines windows for the source contacts. These windows are etched down to the silicon surface, as shown in Figure 5.1*e*, and aluminum is evaporated over the entire top surface of the wafer. At the same time windows are etched down to the polycrystalline silicon layer over the gate vias, so the aluminum contacts there as well. The fifth mask enables the aluminum to be etched, so as to separate the main source contact from the gate contact and also from the field rings at the chip periphery. After removal of the photoresist, a final layer of phosphosilicate glass is deposited over the whole wafer for protective purposes. A section through the device at this stage is shown in Figure 5.1*f*. A sixth mask is used to define windows which are etched through the glassy layer down to the aluminum of the source and gate bonding pads.

The individual MOSFETs are now fully formed. A drain contact is made, for example by evaporating chromium, nickel, and silver layers onto the rear of the wafer. The individual chips are then each probe-tested. The wafer is scribed and broken, and the chips separated, cleaned, die-bonded, and connected.

5.1.2. Process Variations

In the previous section, we have described what may be thought of as a standard process for the fabrication of VDMOS FETs. Each manufacturer has proprietary variations. Some processes are more complex than the one described, making use of the full range of techniques developed for VLSI, in particular, plasma etching and step-and-repeat photolithography. They may employ as many as eight masks. Others aim at greater simplicity and lower cost. One manufacturer has introduced [2] a four-mask process with an interdigitated (noncellular), aluminum-on-polysilicon gate. This has been optimized for high-voltage (500 V), high-frequency applications. Another manufacturer has described [3] a very different interdigitated design, intended to maximize gate width and minimize ON resistance in a low-voltage device. This process applies the technology of VLSI to a development of the VVMOS device. We give a brief account, next, of each of these very different designs.

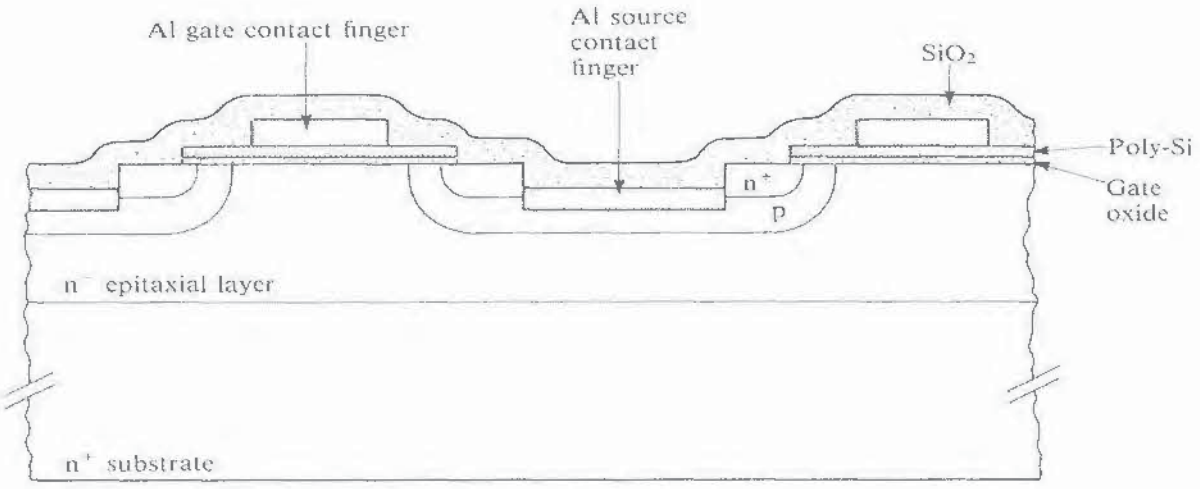


Figure 5.2. Cross-section through the active region of a metal gate device.

The first process aims at achieving a high yield from a simple process with low masking tolerances and a reduced sensitivity to defects. It achieves the required ratings by employing a larger chip area. While this is more costly in raw material, it brings the immediate benefits of lower ON resistance, lower thermal resistance, and lower current density. A cross-section through the active region is shown in Figure 5.2. The two main novel features can be clearly seen. The source–body contact fingers lie at the bottom of trenches etched through the source diffusion and into the body region. This removes the need for the p^+ diffusion, while minimizing the effect of the parasitic bipolar transistor. Because the source metal does not overlie the gate, C_{GS}, and therefore C_{iss}, are reduced. The polycrystalline silicon gate fingers are overlaid with aluminum metallization. This reduces the gate turn-on propagation delay, and ensures that at high frequency all parts of the transistor are fully activated. Only one of the masking steps is critically aligned, compared with the conventional process where the first four masks are each critical in their alignment. The electrical characteristics of the device are: $BV_{DSS} = 500$ V, $I_D = 16$ A, $R_{DS(on)} = 0{\cdot}3\ \Omega$, $C_{iss} = 2$ to 2.5 nF, $Q_g = 55$ to 75 nC.

The second design is shown in plan and cross-section in Figure 5.3, and an outline of the process sequence is illustrated by Figure 5.4. The devices are formed in an 8-μm-thick, phosphorus-doped epilayer of resistivity 0.006 Ω m (0.6 Ω cm). This is grown on an antimony-doped substrate of resistivity less than $6 \times 10^{-5}\ \Omega$ m (6 mΩ cm). A boron implant and diffusion forms the body region, with the junction at a depth of about 4 μm. This is followed by an implant and diffusion of arsenic to form the source with a junction depth of 0.4 μm. Thin films of silicon dioxide and silicon nitride are deposited, then a thicker layer of oxide, as shown in Figure 5.4*a*. Rectangular grooves are then cut to a depth of 2.5 μm, by reactive-ion etching, and after cleaning the surface, the 50-nm-thick gate oxide is grown, as in Figure 5.4b. A polycrystalline silicon layer 0.5 μm thick and heavily doped with phosphorus is deposited next, and its surface is oxidized. This is followed by a deposition of a second layer of poly-Si, this time undoped, which fills the grooves and leaves a planar surface, as seen in Figure 5.4*c*. This is etched

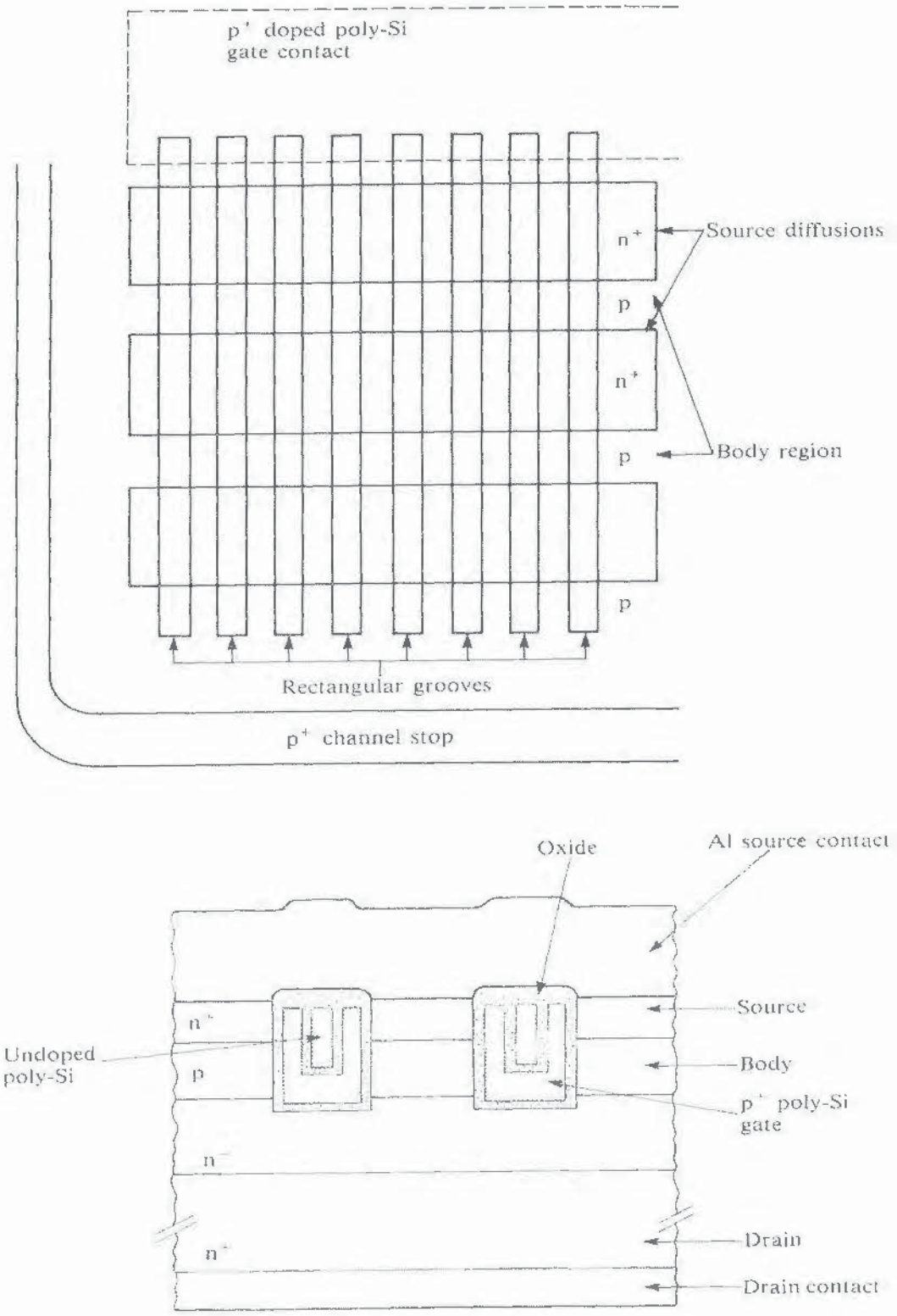


Figure 5.3. A very high-packing-density device. (*a*) Plan view; (*b*) cross-section.

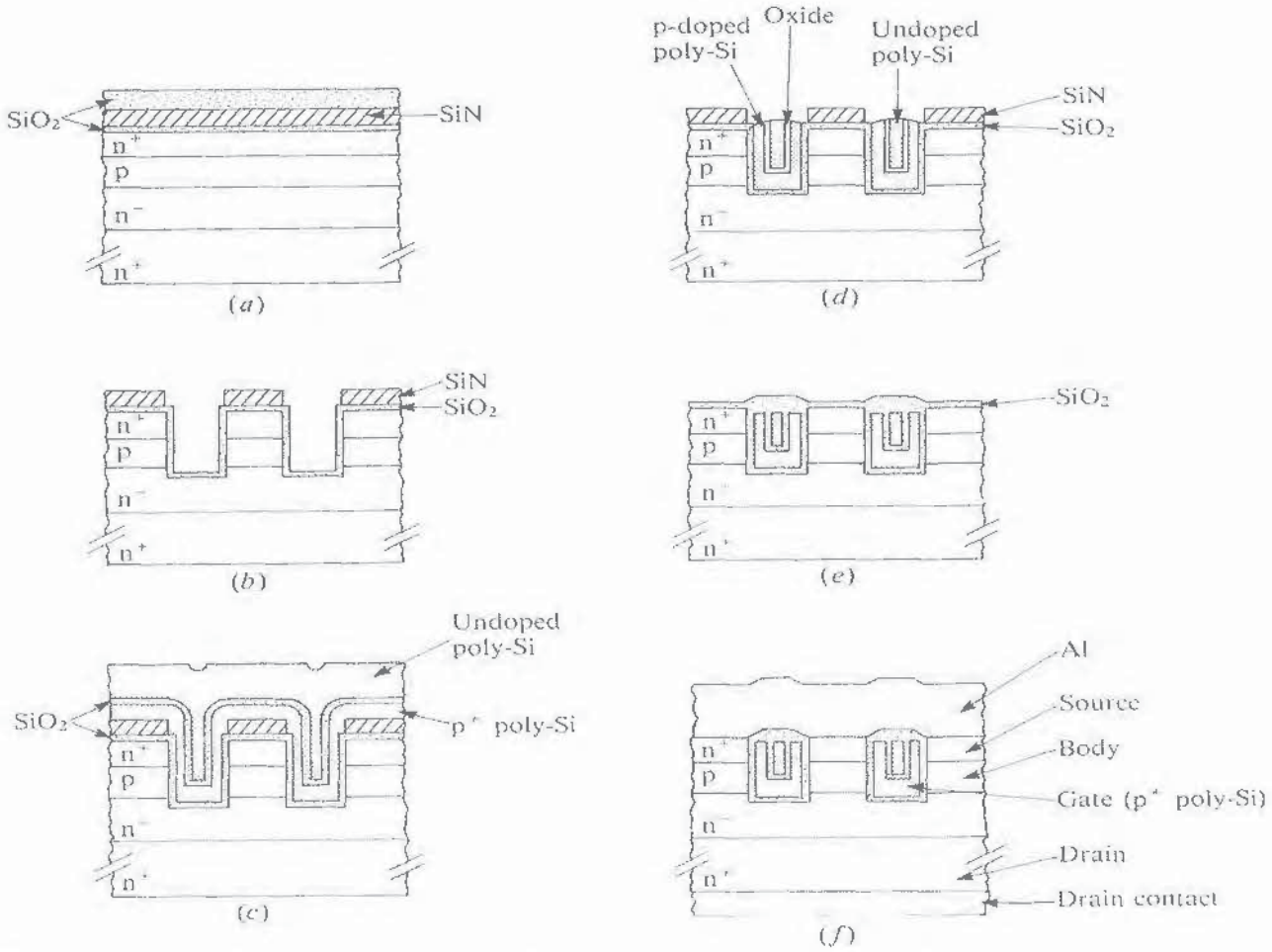


Figure 5.4. Processing sequence for the device shown in Figure 5.3. (*a*) After epitaxial growth, p and n diffusions, and the deposition of insulating layers; (*b*) etched rectangular grooves; (*c*) poly-Si infilling; (*d*) poly-Si etch; (*e*) local oxidation; (*f*) the final device.

back to the nitride layers, again using a reactive-ion etch. These then act as a mask for a local oxidation of the poly-Si of the gate fingers, as in Figure 5.4*d* and *e*. The nitride and gate oxide are removed from the surface of the source, and an aluminum contact layer is evaporated everywhere, as shown in Figure 5.4*f*. The entire gate and channel structure is self-aligned with a single mask. As a result it is possible to fabricate the gate grooves on a pitch of 4 μm. This gives a gate width per unit area of 0.5 m/mm^2 and a value of 0.14 Ω mm^2 for $R_{DS(on)}A$. The electrical characteristics of a device of total area 3.8 × 4.0 mm are $BV_{DSS} = 32$ V, $R_{DS(on)} = 9$ mΩ, $g_{fs} = 62$ S, $C_{iss} = 6.4$ nF, $t_r = 50$ ns, $t_f = 200$ ns.

5.2. QUALITY AND RELIABILITY

5.2.1. Introduction

The VDMOS FET has been designed from the outset as a highly reliable component with a high manufacturing yield. We distinguish the terms *quality* and *reliability* as

follows:

Quality is inversely related to the proportion of defective or out-of-specification parts at the time of shipment. Statistical sampling methods are used to enable the supplier to monitor the outgoing quality level (OQL). Power MOSFETs can achieve an assured electrical outgoing quality level with fewer than 40 defective parts per million (AOQL < 40 ppm). This is obtained by rigorous monitoring and control of the incoming materials and the process and assembly at every stage, and by the careful screening of individual components. At this level of AOQL, individual device inspection by the purchaser is likely to be counterproductive, in that the extra handling involved may well introduce more defective components than it finds.

Reliability measures the capability of the device to perform as specified over an extended operational lifetime. In common with suppliers of other semiconductor components, power-MOSFET manufacturers have attempted to quantify the reliability of their products under the conditions likely to be experienced in normal operation. For this to be of use to the design engineer, reliable data on failure rates have to be presented in such a way that they can readily be related to a wide range of possible operating conditions. In order to achieve this, the most likely modes of failure have to be identified, and a precise definition of what constitutes a failure has to be made. This is usually quite straightforward, and involves a particular parameter such as leakage current or $R_{DS(on)}$ going outside specification. To obtain these data on a reasonable time scale and at reasonable cost, some means has to be found to accelerate the degradation and failure processes. Otherwise, the information would take so long to acquire that, before it was available, the device would likely to be obsolete. For this information to be reliably extrapolated to normal operating conditions, the physical causes of each mode of failure have to be understood. This enables one or more significant stress parameters, such as temperature or electric field, to be isolated. Next, a physical law relating the failure rate to a particular stress parameter has to be established. Finally, the validity of the law is tested, and the data required by the user obtained in accelerated life tests carried out under overstressed conditions. For the results to be statistically meaningful, tests have to be made on a large number of devices. Then confidence levels can be assigned to each predicted lifetime.

In the next section we give a brief review of the terms used in assessing reliability. Later we go on to examine the principal causes of VDMOS failure. Practical aspects of the reliability of power MOSFETs are discussed further in Chapter 17.

5.2.2. Basic Reliability theory

A plot of the failure rate against time, for many components, follows the classic *bathtub* curve. This is shown in Figure 5.5. The instantaneous failure rate is the sum of three components. At early times the failure rate is high. This is caused by devices that are slightly defective in some way, but which still meet the device

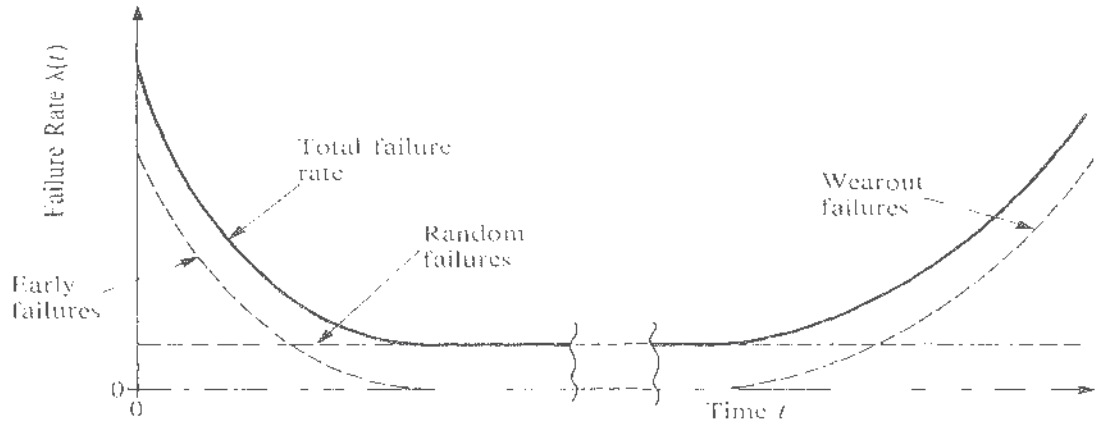


Figure 5.5. The "bathtub" curve.

specification initially. They are thus sent out as good devices, but the defect causes a rapid change in some parameters, so that they are soon out of specification. They are known as *infant mortalities*. They can often be isolated by subjecting all components to a period of *burn-in*, that is, a short test at high temperature and sometimes with other imposed stresses such as high voltage or high current. It is, of course, important that the burn-in should not itself induce new defects and significantly accelerate the ultimate failure of good devices. One of the great benefits of having readily available and reliable life-test data on components is that the equipment manufacturer is relieved of the trouble and cost of running a burn-in schedule, with all its attendant risks.

Throughout the life of the component, failures may occur randomly, for no correlated or clearly ascribable reasons. It is often assumed that during the main period of the life of many components this is the dominant failure mode. The failure rate is taken to be constant and to depend on the level of stress imposed. Under conditions of low stress it may be very low indeed, and then, this period may last for many millions of hours. Really, a random failure phase is most likely to be observed in a complex system in which failures result from many independent and individually unlikely causes. Most causes of failure in a single component, after infant mortalities have been screened out, are associated with a wearout mechanism. Any observed period of apparently constant failure rate is likely to be at best an approximation, and coincidental.

Following any period of random failure comes the final period of *wearout* failure. The failure rate rises as physical and chemical degradation processes accelerate, until finally no good components remain.

It is necessary to define some terms which can be applied to a population of components subjected to life test. The *reliability* $R(t)$ represents the probability that a component survives to the time t. The probability that the component fails sometime during the period t to $t + \mathrm{d}t$ is $f(t)\,\mathrm{d}t$, where $f(t)$ is known as the *failure probability density function*. The probability that the component fails sometime during the period t to $t + \mathrm{d}t$, *given that it survives to the time* t, is $\lambda(t)\,\mathrm{d}t$, where $\lambda(t)$ is known as the *failure rate*. It is $\lambda(t)$ that is plotted in Figure 5.5. Clearly,

$$f(t) = \lambda(t)R(t) \tag{5.1}$$

Other relationships between these parameters may easily be derived. The probability of failure during the whole of the period from the start of life to time t is given by

$$\int_0^t f(t)\,\mathrm{d}t = 1 - R(t) = F(t) \tag{5.2}$$

where $F(t)$ is known as the *cumulative failure probability function*. Note that $F(0) = 0$, $R(0) = 1$, $F(\infty) = 1$, and $R(\infty) = 0$. Differentiation of Equation (5.2) shows that

$$f(t) = \frac{-\mathrm{d}R}{\mathrm{d}t} \tag{5.3}$$

And substituting Equation (5.3) into Equation (5.1) gives

$$\lambda(t) = \frac{-\mathrm{d}R/\mathrm{d}t}{R(t)} \tag{5.4}$$

Integration of Equation (5.4) gives:

$$R(t) = \exp\left(-\int_0^t \lambda(t)\,\mathrm{d}t\right) \tag{5.5}$$

The relationship between these parameters for a component that follows the "bathtub" curve is illustrated in Figure 5.6.

During any random failure period, when the failure rate λ is constant, $R(t) = e^{-\lambda t}$, $F(t) = 1 - e^{-\lambda t}$, and $f(t) = \lambda e^{-\lambda t}$. It is only during a period when λ is constant that it is helpful to the user for the failure rate of a component to be specified. When this is possible, the failure rate is usually measured in units representing one failure per 10^9 device hours. These are known as failure units, or FITS. Sometimes percentage failures per 1000 hours are quoted. Note that a failure rate of 1% per 1000 hours corresponds to 10^4 FITS.

The *lifetime* $\langle t \rangle$, or the *mean time to failure* MTTF, is defined as

$$\langle t \rangle = \int t f(t)\,\mathrm{d}t \tag{5.6}$$

and the standard deviation σ about the mean lifetime, is given by

$$\sigma^2 = \int (t - \langle t \rangle)^2 f(t)\,\mathrm{d}t = \int t^2 f(t)\,\mathrm{d}t - \langle t \rangle^2 \tag{5.7}$$

It is easily shown that when λ is constant, $\langle t \rangle = \sigma = 1/\lambda$. The term *mean time between failures* (MTBF) is often used for $\langle t \rangle$, although really this is best applied to a system containing many subsystems or components that are repaired or replaced on failure.

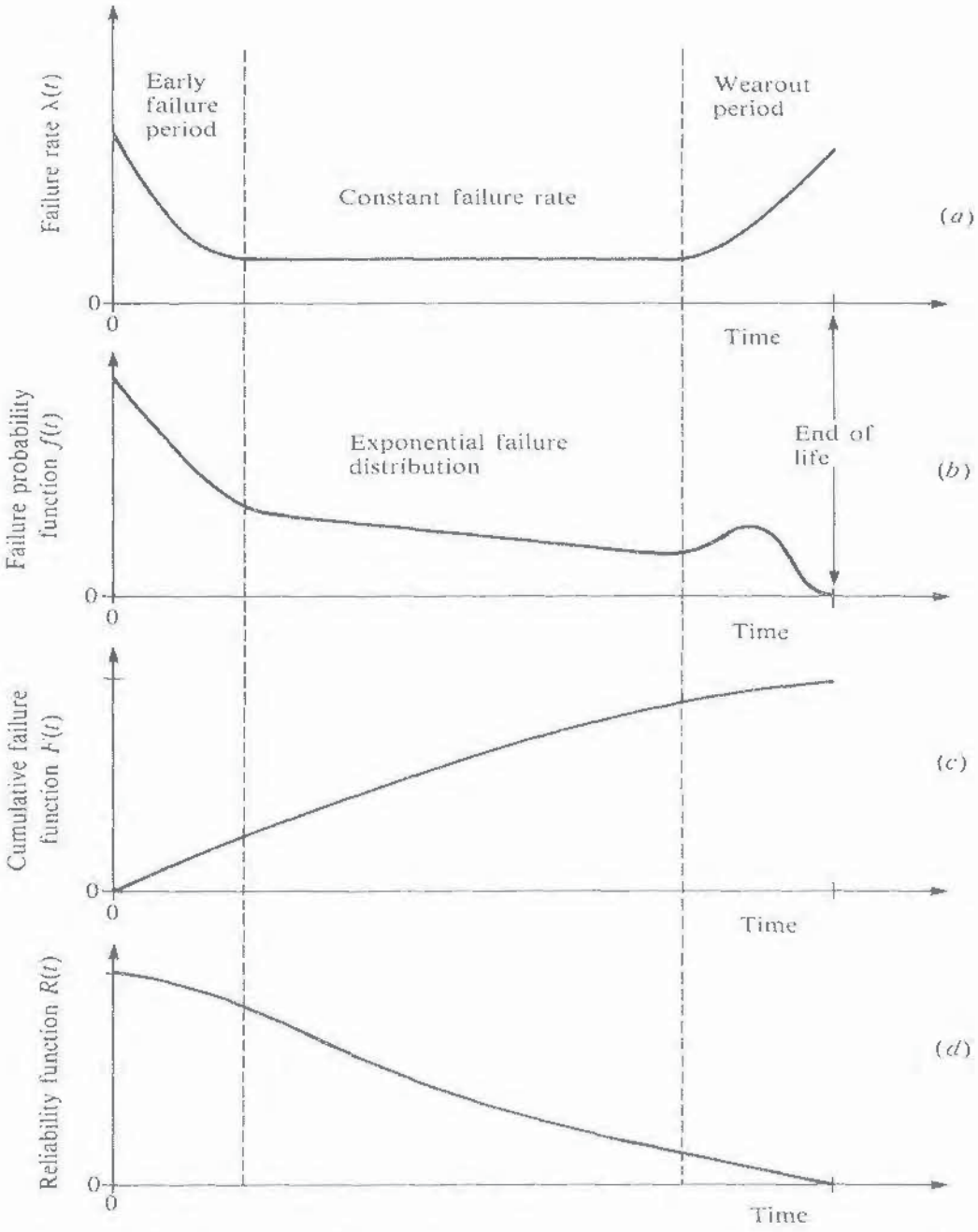


Figure 5.6. Variation of the failure parameters during life when failure follows the "bathtub" curve. (*a*) Failure rate $\lambda(t)$; (*b*) failure probability density function $f(t)$; (*c*) cumulative failure probability function $F(t)$; (*d*) reliability function $R(t)$.

In many instances, the cumulative failure function for semiconductor devices follows a *lognormal* distribution. The implication of this is that failure is the random consequence of the *product* of a number of independent factors. It is perhaps better to think of it as occurring with a probability proportional to the exponential function of the *sum* of many independent events. As these events result from the operation of the component, this is essentially a wearout

mechanism. The failure probability distribution function takes the form

$$f(t) = (2\pi)^{-1/2}(ts)^{-1}\exp\left(\frac{-[\ln(t/t_m)]^2}{2s^2}\right) \tag{5.8}$$

In Equation (5.8), t_m is the geometrical mean of the failure times and is also the *median* lifetime, that is, the time at which 50% of the components will have failed. The factor s is sometimes called the dispersion. It represents the range of lifetimes about their mean value $\langle t \rangle$, which is given by:

$$\langle t \rangle = t_m \exp(s^2/2) \tag{5.9}$$

In Figure 5.7, we have plotted the various failure functions for a typical lognormal failure distribution. In Figure 5.7*c*, it can be seen that a graph of $f(t)$ versus $\ln t$ gives a normal or Gaussian distribution. The integral of this distribution, shown in Figure 5.7*d*, can be transformed into a straight line by the use of the so-called probability scale for the ordinate, as shown in Figure 5.7*e*. The validity of the lognormal distribution function can be established by recording the times of successive failures in this way. The fraction of the total population of devices under test that have failed is recorded on a normal probability scale, against time plotted on a logarithmic scale. Once the straight line is established, the median lifetime and the dispersion can easily be determined.

When the life test is carried out on a small sample from the total population of components, the supposition is that the distribution of the samples is representative of that of the whole population. However, let us say that we test a very small sample of ten components. When the first one fails, we might at first sight presume that this time is representative of the time to failure of 10% of the population. This is unduly pessimistic, because it is most unlikely that *all* 10% of the population would have failed by this time, had they all been put on test. There are methods available to correct for this. One that is quite accurate and easy to apply is known as Bernard's method. The ith failure out of a sample of n is regarded as representing the fraction $(i - 0.3)/(n + 0.4)$, rather than i/n, of the total population. This is what should be plotted on the ordinate scale of Figure 5.7*e*.

In general, knowledge of the median lifetime of a component is of little value to an equipment designer. Much more important is to know the likely time to the first failure among the many components used in a given system. When a lognormal failure distribution is established for a sample of these components, an estimate can be made of the likely time to failure of, say, the first 0.1% of components. Furthermore, depending on the sample size used, confidence limits, expressing the probability that such an estimate is representative of the whole population, can be calculated.

Many of the degradation processes in semiconductors involve the formation of dislocations and other crystal defects, or the diffusion and aggregation of impurities. All of these are thermally activated processes, having a characteristic

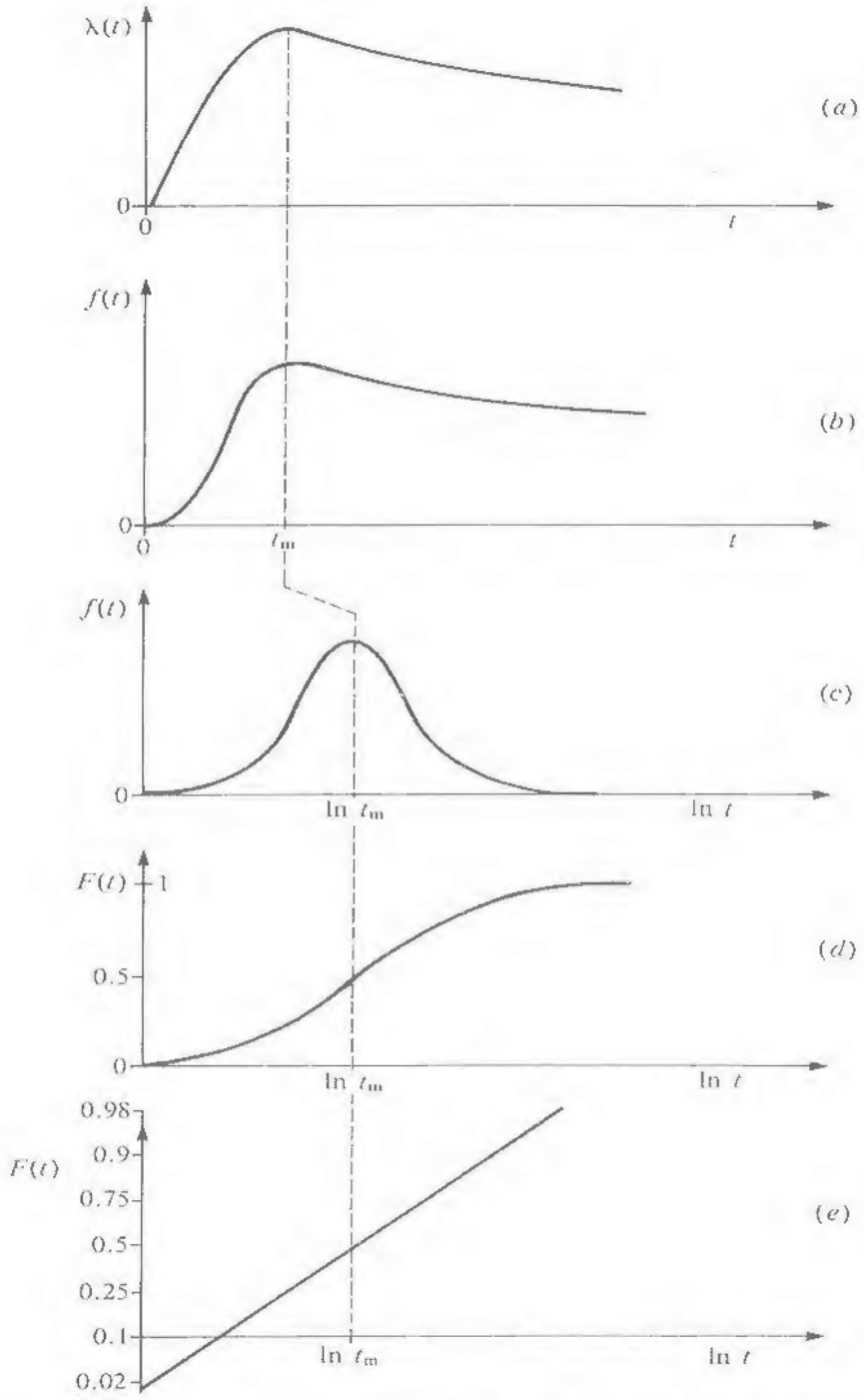


Figure 5.7. Variation of the failure parameters during life when failure follows the lognormal distribution. (*a*) Failure rate $\lambda(t)$; (*b*) failure probability density function $f(t)$; (*c*) $f(t)$ versus time plotted on a logarithmic scale; (*d*) cumulative failure probability function $F(t)$ versus time plotted on a logarithmic scale; (*e*) $F(t)$ plotted on a normal probability scale, versus time plotted on a logarithmic scale.

activation energy, E_a. They are enhanced by an increase in temperature, according to the law of Arrhenius [4]:

$$\text{rate of activation} \propto \exp(-E_a/kT) \tag{5.10}$$

The time required for a sufficient number of such events to produce a condition of failure is thus proportional to $\exp(E_a/kT)$, and it follows the lognormal distribution function. This can be established by plotting the median lifetime (or some other characteristic time) on a logarithmic scale versus the reciprocal of the absolute

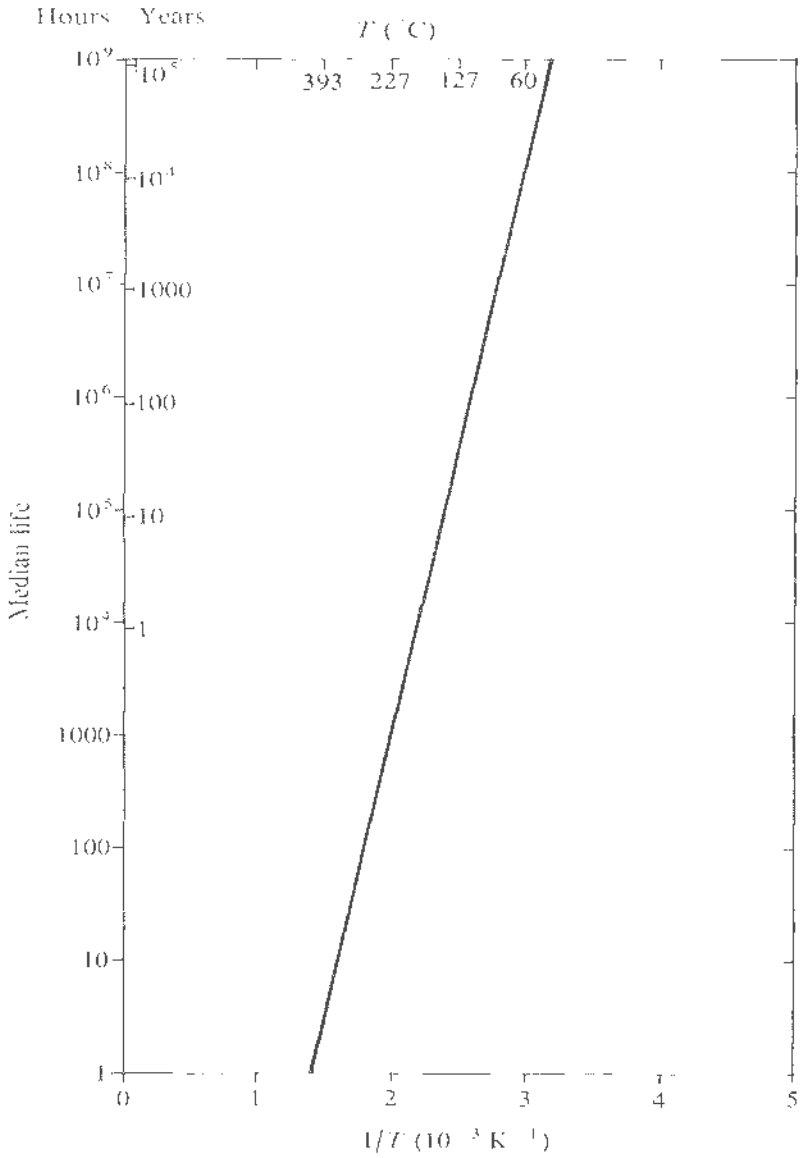


Figure 5.8. An Arrhenius plot: a graph of the median time to failure, plotted on a logarithmic scale, against the reciprocal of the absolute temperature. The graph shown corresponds to the example given in the text, namely, the activation energy is 1 eV and the median life at 227°C is 1000 hours.

temperature, as shown in Figure 5.8. The slope of the line is governed by the activation energy, which in some instances can be determined independently.

Establishing the value of E_a with a degree of confidence is the basis of the use of high temperatures to accelerate life testing. To give an example, if E_a is 1 eV, the lifetime is proportional to $\exp(e/kT) = \exp[(11{,}602\ \mathrm{K})/T]$. The expected life of a device run at 500 K (227°C) is shorter than that of one run at 400 K (127°C) by a factor of $\exp[11{,}602(1/400 - 1/500)]$, that is, by a factor of $\exp(5.8) = 330$. Provided that the same failure mechanism is dominant at both temperatures, a median lifetime of 1000 hours measured at the higher temperature leads to a predicted median lifetime of 330,000 hours ($37\frac{2}{3}$ years) for a device run at the lower temperature.

5.2.3. The Main Causes of VDMOS Failure

Device failure, as with most semiconductor devices, is usually the result of faulty circuit design causing one or more of the device operating limits to be exceeded. Particularly vulnerable periods occur during the switching of inductive circuits, when overvoltage or overrapid reapplication of voltage (excess dV/dt during diode recovery) may occur. The most common causes of failure resulting from such misuse are:

1. Gate short, usually caused by excessive voltage applied to the gate, either in circuit or during handling (electrostatic discharge).
2. Safe-operating-area failure, caused by one or more excessive thermal transients, or by poor heat-sinking.
3. Avalanche failure, when excessive avalanche current is drawn during unclamped inductive switching.
4. Fused leads, following excessive transient overload current.

Failures that may develop during long periods of normal use at or near to the specified limits of the device operating conditions usually take one of the following forms:

1. Gate failure. The gate leakage current increases, and eventually the transistor becomes stuck in the ON state, leading to a catastrophic circuit failure.
2. Drain–source leakage increase. The transistor becomes overheated and thermal runaway ensues.
3. Connection failure. Thermomechanical fatigue or corrosion at a wire bond, or at the die bond, leads to a rise in the thermal impedance R_{Th} or in $R_{\mathrm{DS(on)}}$, and hence in dissipation.

In each case the most likely symptom of approaching failure is a gradual increase in the device temperature, with the attendant risk of thermal runaway during onerous operating conditions.

The underlying physical causes of these progressive failures mechanisms are conveniently divided into those associated with the die and those associated with

the packaging. Two die defects are the most common. The first is the electric field distortion brought about by the accumulation of polar molecules and ionic contaminants in the passivation of the high-field region, in particular the field rings around the periphery of the chip. The second is the growth of defects in the gate oxide. These usually take the form of voids or inclusions. Several mechanisms that cause failure in other types of semiconductor devices do *not* arise in practice in properly constructed VDMOS FETs. These include the development of slow trapping levels in the active region, microcracking of the aluminum metallization, and electromigration of the metallization (which can occur at high current densities). Packaging failures include thermal fatigue of the die bond brought about by differential expansion between the silicon and the header, wire-bond fatigue, and corrosion, usually caused by adsorbtion of water (OH^-) onto the chip surface. We deal with each of these failure mechanisms in turn, examining how they may be enhanced for the purposes of accelerated life-testing.

Field Distortion. This manifests itself as a gradual increase in the drain–source leakage current I_{DSS}. Allowed to progress far enough, it can lead to thermal runaway. Accelerated life tests are carried out at 150 and 175°C, with 80% of the rated drain–source breakdown voltage (BV_{DSS}) applied, and $V_{GS} = 0$ V. These are known as high-temperature reverse-bias (HTRB) tests. Tests carried out at different temperatures and different bias levels have established that temperature acceleration follows the Arrhenius law with an activation energy of 1 eV, and that bias acceleration increases the failure rate by the factor $e^{V/(31\text{ V})}$. Published data have usually been treated as though a constant failure rate were to be expected, and the results are quoted with an upper confidence level of 60% in equivalent FITS at a temperature of 90°C. See for example Figure 19 of Appendix 7. Note that as this failure mechanism is essentially one of wearout, burn-in is not capable of reducing the failure rate.

Oxide Defects. The accumulation of oxide defects causes a gradual increase in the gate–source leakage current I_{GSS}, leading eventually to a possible rupture of the gate oxide and a gate–source short circuit. Accelerated gate stress life tests are carried out at 150 and 175°C with $V_{DS} = 0$ V and $V_{GS} = 20$ V, the maximum rated value. Again, even though this is essentially a wearout mechanism, it is usually assumed that a constant failure rate occurs. Thermal acceleration follows the Arrhenius law, with $E_a = 0.4$ eV. The electric field in the gate oxide, E_{ox}, which is proportional to V_{GS}, enhances the failure rate by the factor $\exp(-E_{ox}/E_c)$, where $E_c = 10.8\text{ V}/\mu\text{m}$. Again, burn-in is not effective. The estimated time to 1% failure probability is shown in Figure 18 of Appendix 7.

Die-Bond and Wire-Bond Fatigue. Differential thermal expansion between the silicon die and the header, and thermomechanical and current pulsing stresses, cause fatigue and degradation of the die bond. The thermal impedance is very sensitive to this effect and soon increases, leading to an increase in the junction temperature. The wire bonds may degrade in a similar manner, causing $R_{DS(on)}$ to rise and become unstable. The vulnerability of devices to these effects is tested in two ways. In the power cycling test the device is switched on for some two or three minutes, during which time heat generated by the dissipation in the die

raises the case temperature from an ambient of 30°C to 100 or 130°C. The transistor is then switched off and forced-air-cooled back to 30°C. These cycles are continued to failure. In the temperature cycling test the devices are taken from −55 to +150°C and back over a period of about 30 minutes. No bias is applied. Lifetests usually extend for at least 10,000 cycles. In this case degradation results from differential thermal expansion. Plastic encapsulated (TO-220) devices, which have to use hard-soldered bonds, are likely to be more vulnerable than those that are hermetically sealed (TO-3), which can use soft-soldered connections. Transient temperature gradients created in the power cycling test give rise to further stresses. In each case failure times have been found to vary according to the Arrhenius law, with an activation energy of about 0.3 eV.

Surface Corrosion. The ingress of water molecules into non-hermetically-sealed packages leads to device deterioration in two ways. First, ingress into the field termination region distorts the electric field distribution and degrades the voltage breakdown characteristics, causing leakage currents, as described previously. Secondly, hydroxyl ions penetrating to the regions of the bonding pads stimulate cathodic corrosion of the aluminum contact. This leads to an increase in the values of $R_{DS(on)}$ and V_{SD}, both of which become unstable. Drain bias accelerates this effect. However, the increased working temperature of an operating device helps to exclude the water vapor. Samples of non-hermetically-sealed devices are subjected to a *temperature–humidity -bias* test of at least 1000 hours at 85°C and 85% relative humidity with $V_{GS} = 0$ V and $V_{DS} = 10$ V or at the full rating (BV_{DSS}). A variant on the test that is sometimes used is the *highly accelerated temperature and humidity stress test* (HAST). This is a high-pressure (0–40 psig), high-humidity (50–80% relative humidity) test normally carried out for several tens or hundreds of hours with $V_{DS} = 8$ V. Similar failure mechanisms occur.

Avalanche Failure. A test has been introduced recently to detect any degradation brought about by subjecting devices to repetitive avalanche breakdown. The test circuit is similar to that shown in Figure 14 of Appendix 7.

SUMMARY

Power MOSFETs are fabricated by different manufacturers using a range of processes of differing sophistication and complexity. The processes use all the techniques normally associated with MOS integrated-circuit fabrication.

A particularly important part of the design is that of the field rings around the periphery. These are formed at the same time as the main device. They are critical for the achievement of the voltage rating, especially when the body–drain diode is required to sustain controlled avalanche breakdown.

Power MOSFETs have been designed from the outset as high-quality and highly reliable components. Device failures that are not the result of faulty circuit design usually follow a lognormal distribution function. This indicates wearout, rather than random failures, and means that burn-in is of limited benefit. The

degradation processes are accelerated by increased temperature. The rate of degradation often follows the law of Arrhenius, indicating an activation-energy-controlled process. Humidity and applied electric field also act as accelerating stress factors for many of the degradation mechanisms.

REFERENCES

1. S. M. Sze (Ed.), *VLSI Technology*. McGraw-Hill, New York, 1983.
2. T. Daly, "New technology makes power MOSFETs faster, more efficient." *Power Convers. Intell. Motion* (*PCIM*) pp. 14–18 (1988).
3. D. Ueda, H. Takagi, and G. Kano "An ultra-low on-resistance power MOSFET fabricated by using a fully self-aligned process." *IEEE Trans.* on *Electron Devices* **ED-34,** 926–930 (1987).
4. S. Arrhenius, "Über die Dissociationswärme und den einfluss der temperatur auf den Dissociationsgrad der Elektrolyte" ("On the heat of dissociation and the influence of temperature on the degree of dissociation of an electrolyte"). *Z. Phys. Chem.* **4,** 96–116 (1889).

CHAPTER 19

Special-Purpose MOSFETs

19.1. CURRENT-SENSING MOSFETs

19.1.1. Current Sensing

Many power-MOSFET applications incorporate device current sensing to implement control or protection features. Traditional current-sensing techniques suffer from a number of problems. Current-sensing resistors are inefficient, increasing heat removal problems. Current-sense transformers have bandwidth and saturation limitations, and Hall-effect devices need to be sophisticated to overcome accuracy and temperature-dependence problems. Wire-wound resistors and transformers introduce inductance into the circuit, which limits high-frequency operation. Furthermore, all these methods involve extra components that must be sourced and stocked, put some obstacle in the way of automated assembly, use relatively bulky components, and increase equipment cost. Current-sensing MOSFETs have the benefits that special current-sensing components are not essential, automated assembly capability is not affected, and switching performance is unimpaired.

Current-sensing MOSFETs have been developed to meet the need of current monitoring in the power device. In current-sensing MOSFETs a known fraction of the source current is diverted into a second, independent source terminal. This had led to the device being sometimes called a current mirror [1]. The sense current is typically of the order of $\frac{1}{1500}$ of the drain current, and can be converted into a scaled-voltage indication of the drain current in a way that is efficient in terms of power, cost, and board space.

19.1.2. Principle of Operation

A vertical diffused power MOSFET consists of many parallel cells formed on the upper surface of the silicon die. An enhanced channel is formed around the edge of each cell to allow vertical drain–source conduction. When fully inverted, the channel regions act as many similar resistors in parallel, all with equal, positive temperature coefficients. This situation results in current being evenly distributed between cells. It is this even distribution of current that allows current sensing by the detection of the current flowing through a few cells, thereby giving a scaled indication of the total drain–source current.

19.1.3. Structure

The source metallization covering a small number of ordinary cells is isolated from the main source metallization. A separate bonding pad is provided to allow

Figure 19.1. Current-sensing MOSFET die. The sense cells are visible just above the central current-sense bonding pad. Just above that is the Kelvin source bonding pad. The main source pad is at the top of the picture; the gate pad is at the bottom. The dots on the source pads are shorts to the underlying p' region, used to give the device improved avalanche and dv/dt characteristics. (Courtesy of International Rectifier Corp.)

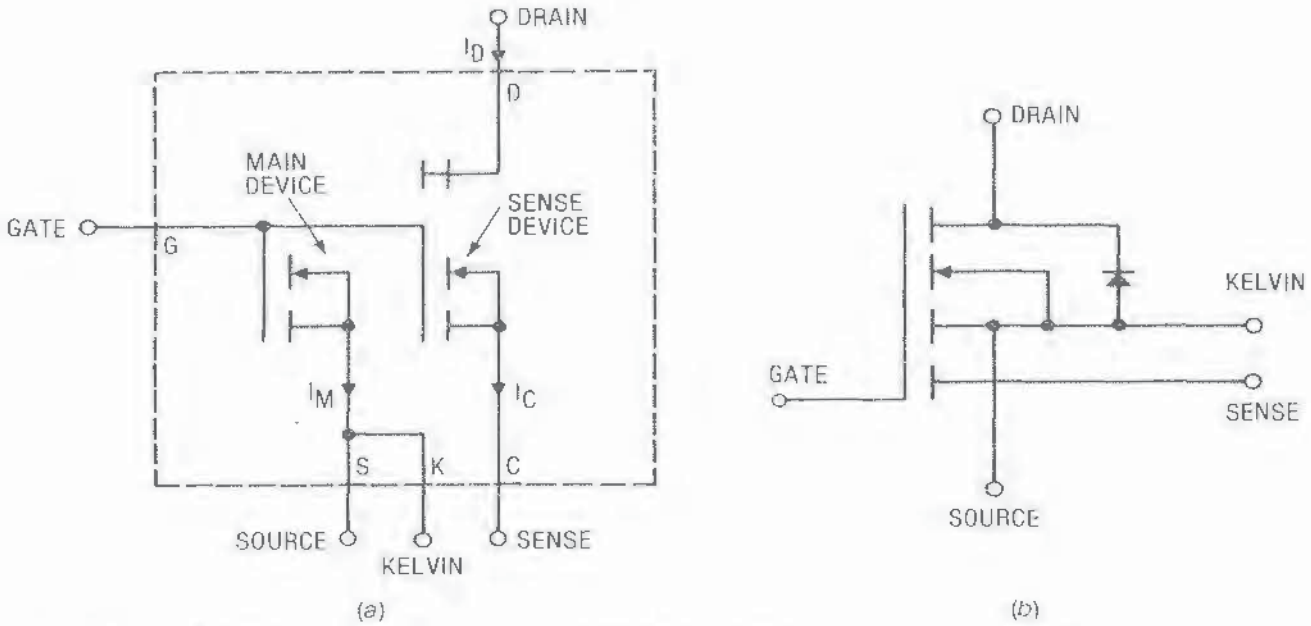


Figure 19.2. Current-sensing MOSFET. (*a*) Schematic representation; (*b*) circuit symbol.

connection of the sources of the isolated cells to the current-sense terminal. Two connections are made to the main source metallization via separate bonding pads. One connection is used as the main source terminal and carries the main source current. The other, known as the Kelvin contact, is connected to the main source metallization close to the sense cells. The Kelvin terminal is used to provide a return path for the diverted sense current. The Kelvin contact is located close to the source cells in order to minimize the voltage drop in the main source metallization. Figure 19.1 shows a photograph of a current-sensing MOSFET die. The equivalent circuit of the device is shown in Figure 19.2, along with the circuit symbol for the device. As Figure 19.2 shows, the current-sense MOSFET can be represented by two parallel devices—a main device and a sense device. The ratio of main-device current to sense-device current is termed the sense ratio, and is denoted by r. The intrinsic sense ratio r_0 is that ratio obtained when the effect on current distribution of the measuring system and any parasitic elements is zero.

19.1.4. Static Characteristics

An equivalent circuit for a fully enhanced device is shown in Figure 19.3, with typical element values for a 100-V device and a 500-V device.

With the source and the sense terminals connected together, all cells have essentially equal gate–source bias. The gate voltage must exceed the minimum necessary gate voltage by two or three volts so that the effects of parasitic voltage drops within the device and its package are minimized. Under these conditions

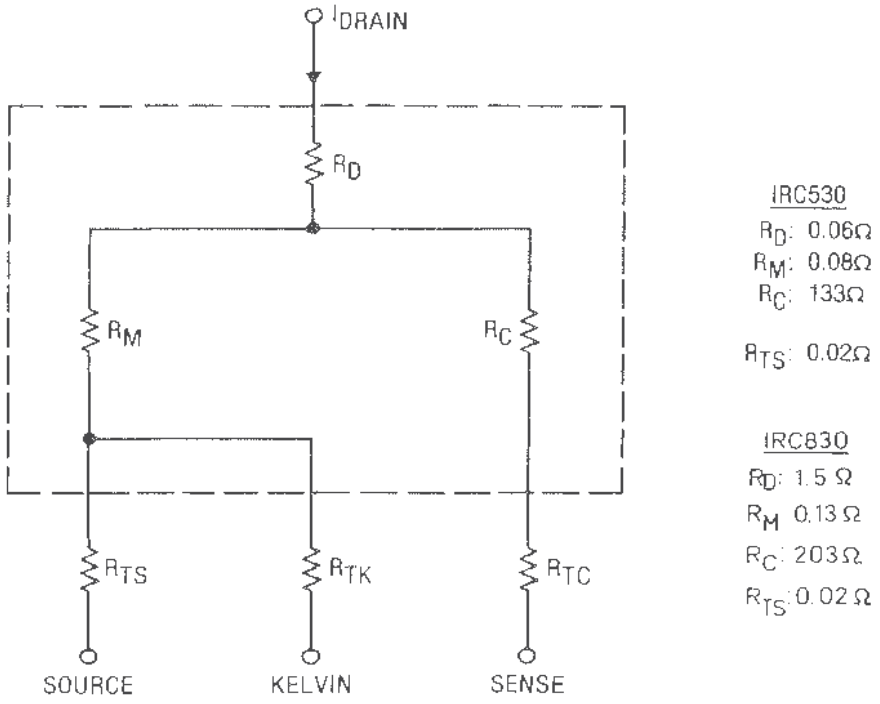


Figure 19.3. Equivalent circuit of a fully enhanced current-sensing MOSFET for static conditions for a 100 V device (IRC 530) and a 500 V device (IRC 830).

the channel resistance can be assumed to be the same for all cells. In Figure 19.3 the equivalent resistance of the parallel cells of the sense device is shown as R_C, and that of the main device cells is shown as R_M. Thus

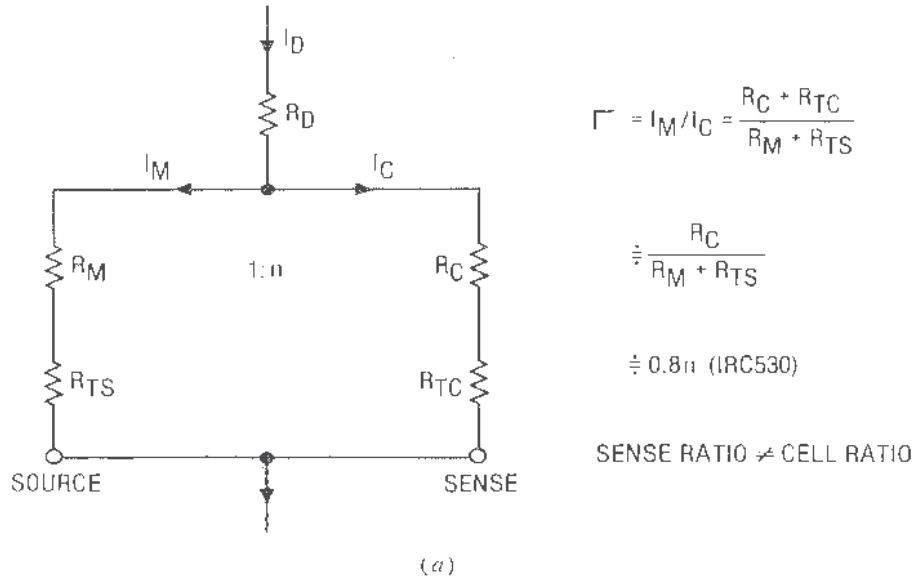
$$R_C = nR_M \tag{19.1}$$

where n is the cell ratio. R_D is the bulk drain resistance of the device, and R_T represents terminal resistances, including metallization, bonding-wire, and post resistances.

With the sense and source terminals connected, and neglecting parasitic resistances, the ratio of main-cell to sense-cell current is given by

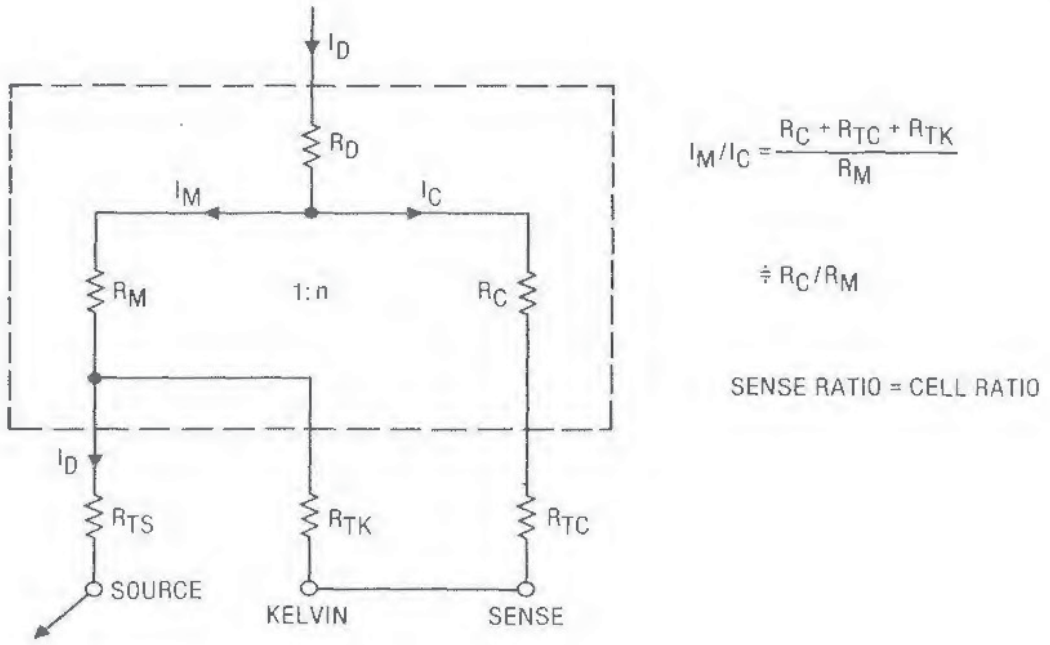
$$\frac{I_M}{I_C} = \frac{R_C}{R_M} = n \tag{19.2}$$

Thus, in the absence of any external resistance, the sense ratio is theoretically equal to the cell ratio. Since R_C and R_M are altered in the same proportion by changes in temperature, drain-current magnitude, and manufacturing spread in $R_{DS(on)}$, the sense ratio remains constant with variation of these conditions. If, however, parasitic resistances external to the silicon die are introduced into the two current paths, the sense ratio will be disturbed from its intrinsic value, as Figure 19.4*a* shows. Furthermore, the external resistances will not behave in the same way as R_C and R_M, and therefore the sense ratio becomes dependent on temperature, current magnitude, and spread in $R_{DS(on)}$. Parasitic impedances can also affect the current ratio. The Kelvin connection is designed to reduce the amount of parasitic resistance and inductance in series with the main cells. The Kelvin connection provides a path for returning the sensing current to the source



(*a*)

Figure 19.4. Preserving the sense ratio by using the Kelvin source. (*a*) Sense current returned to the source terminal; (*b*) sense current returned to the Kelvin source.

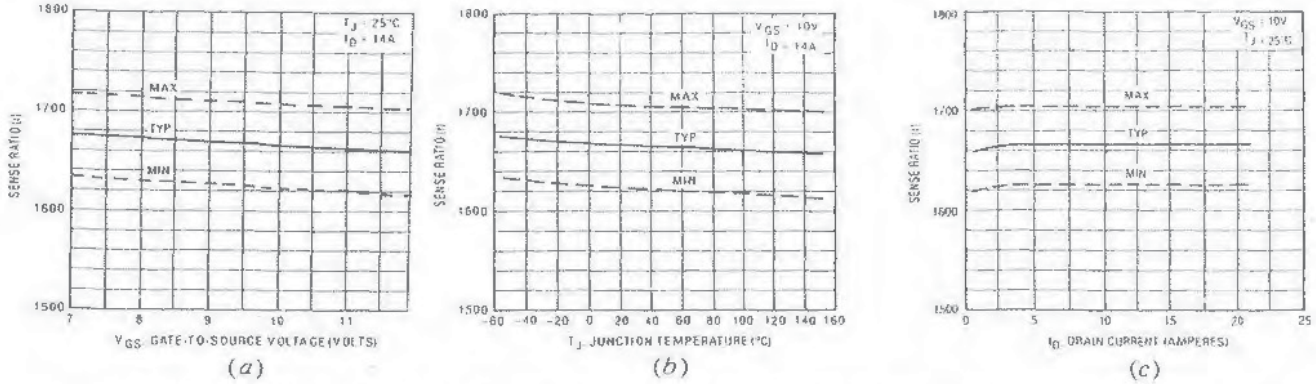


(*b*)

Figure 19.4. (*Continued*)

without including the source wire bond, the source pin, and much of the source metallization. As Figure 19.4*b* illustrates, the sense ratio is then less dependent on parasitic resistances.

The typical data-sheet characteristics shown in Figure 19.5 are based on the virtual-earth current-sensing method (see Section 19.1.7) and therefore reflect variations of the sensing ratio of the packaged device with no external resistance introduced into the sensing circuit. The nominal intrinsic sense ratio of this device is specified as 1665, with $I_D = 14$ A, $V_{GS} = 10$ V, and $T_J = 25°C$. The graphs show a possible manufacturing spread of 2.5% in the sense ratio and a shift in ratio of approximately 0.3% for a change in temperature of 100°C.



(*a*) (*b*) (*c*)

Figure 19.5. Variation of the sense ratio with (*a*) gate voltage, (*b*) junction temperature, and (*c*) drain current.

19.1.5. Reverse Conduction

With the MOSFET off and current flowing through the integral diode, the current again divides between the main cells and the sense cells. However, this division of the current is not necessarily in the same ratio as for forward conduction, since although the ratio of the channel resistances may be the same, there may be differences in the voltage drops of the diodes. The sense current may therefore only be used to give an approximate indication of the device current. A more accurate current measurement may be obtained if the channel is turned on during diode conduction [2]. Provided the forward drop is below approximately 0.6 V, the current will tend to divide in the ratio of the resistances of the main and sense cells.

During reverse recovery of the MOSFET intrinsic diode, large sense-current spikes will be produced when the drain–source voltage is rising due to the recovery current of the sense-cell diodes.

19.1.6. Dynamic Characteristics

Parasitic capacitances associated with the sense and main devices are not in the same ratio as the intrinsic current sense ratio due to the capacitance of the current sense bondpad. This suggests that at turnoff there will be spurious spikes in the sense signal due to the charging of the output capacitance of the MOSFET as the drain voltage rises. However, in practice, other factors dominate [3].

Current spikes observed at turn-on and turn-off often originate in the measuring equipment. The current sense ratio of a current-sensing MOSFET is typically in the region of 1500:1. Signals of the order of a few millamps or a few tens of millivolts are therefore being measured in intimate proximity to circuits carrying tens of amps, with rates of change of current of tens or hundreds of amps per microsecond. Voltages that can swamp the signal are easily generated in the sensing circuit and the ground lead of an oscilloscope probe. Induced errors usually take the form of spikes of opposite polarity at turn-on and turn-off. Even wrapping the ground lead around the probe tip can be insufficient protection. A clean signal may be obtained by removing the probe tip and soldering the sense resistor across the inner tip and the adjacent earth sheath. The resistor is then located directly across the sense and Kelvin leads of the MOSFET package.

Once these precautions have been taken, the sense signal will still contain false peaks at turn-on and turn-off. These peaks are due to a difference in current ratio between the linear operating region and the fully enhanced operating region. In the linear, or constant-current, region the current divides in the ratio of the total channel widths of the main and sense cells. In the fully enhanced region the current division between the main and sense cells is determined mainly by the ratio of the path resistances. Generally, the ratio of the resistance of the drift region of the sense cells to the resistance of the drift region of the main cells is close but not identical to the sense ratio in the linear region. When an external sense resistor is included in the sense path, the ratio of path resistances becomes significantly different.

There is another effect at work during the switching interval that tends to

produce spurious signals—transformer action. Even if all magnetic coupling between the main and sense circuits external to the device package is eliminated, there remains the possibility of coupling within the package itself. The drain and source leads form the primary of a transformer, while the sense and Kelvin leads form a secondary. In the usual lead configuration the two loops overlap. Therefore the package constitutes a poorly coupled 1:1 transformer. Since the current ratio in the two circuits is of the order 1500:1, even a low degree of coupling can induce a significant error, and large spikes should be expected in the sense signal during periods of high di/dt. These spikes do not in fact appear during switching since for most of the switching interval the MOSFET is in the linear region and the current in the sense circuit is a function of the transconductance of the sense cells and the gate voltage. However, transformer action will induce spikes in the current sense signal if the current sense MOSFET is maintained on, and its drain current is switched by some external means. The transformer effect clearly has implications for the bandwidth of the current-sensing MOSFET [3].

19.1.7. Virtual-Earth Sensing Method

Two methods are commonly employed for obtaining a scaled voltage indication of drain current. These are the virtual-earth method and the series-resistor method. The virtual-earth method is less sensitive to temperature variations, while the series resistor method is simpler and cheaper.

The virtual-earth method is illustrated in Figure 19.6. The virtual-earth action of the operational amplifier ensures that the sense and Kelvin pins are effectively at the same voltage, so that no voltage which might upset the sense ratio is introduced into the sensing circuit. Operational-amplifier input bias currents do not significantly affect the current ratio. For example, using the device whose characteristics are given in Figure 19.5, an input bias current of 500 nA would result in an error of 0.1% for a drain current of 0.8 A. Amplifiers with bias currents less than 25 nA are readily available. If necessary, bias-current offset

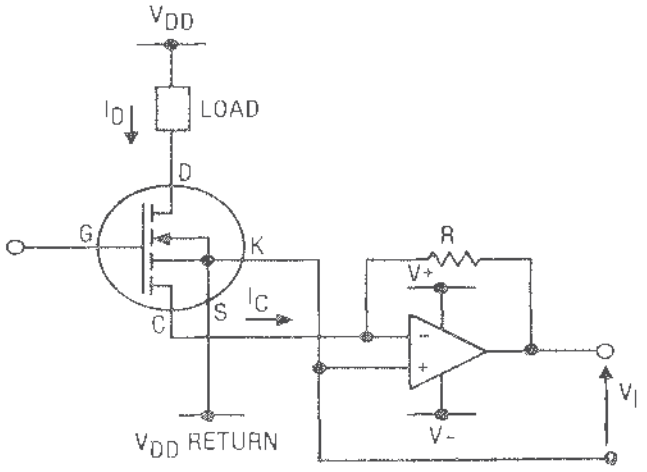


Figure 19.6. Virtual-earth sensing circuit.

may be reduced by locating a resistor between the noninverting terminal and ground. Input offset voltage adds directly to the output voltage. The error it introduces may be significant if a low gain has been chosen. For example, peak-detection circuitry may involve comparing the output with a 200-mV reference. A maximum offset of 2 mV or so would then be required from the operational amplifier.

The operational amplifier must also have a slew rate adequate to allow it to track rapid changes in current. For example, with 0.5-V/A output, and the MOSFET switching 10 A in 100 ns, an amplifier slew rate of 50 V/μs would be required. Many power-supply and motor-control integrated circuits use a current limit threshold voltage of the order of 200 to 300 mV, in which case a slew rate of 2 to 3 V/μs would be adequate.

Capacitance charging currents give rise to sense-current spikes. Current spikes may also be present as part of normal circuit operation, for example diode reverse-recovery spikes. If it is required that these should not trigger overcurrent detection circuitry, filtering or blanking of the spikes may be necessary [4]. An *RC* filter may be incorporated in the operational-amplifier circuit. The operational-amplifier output should not be allowed to be driven into saturation, as the overload recovery time is likely to be long. A zener diode, located across the feedback resistor, may be used to limit the output-voltage swing. Also, the operational amplifier will leave the linear mode if the output becomes slew-rate-limited. Delays in changing slewing directions may cause very short submicrosecond spikes to appear as pulses 1 to 2 μs wide. These may be difficult to filter and may trigger overcurrent detection circuitry. Therefore slew-rate-limited operation should be avoided.

The disadvantages associated with the virtual-earth circuit are that (a) dual power supplies are required, (b) the current-sense voltage indication is negative, (c) an active device is required to accomplish the current to voltage transformation. These disadvantages are overcome, at the expense of reduced accuracy, with the series-resistor method.

19.1.8. Series-Resistor Sensing Method

The voltage output is obtained by placing a resistor in the path between sense and Kelvin, as shown in Figure 19.7*a*. The voltage across the resistor is a scaled voltage indication of the device current. In some applications the control circuits may of necessity be referenced to the source connection. In this case the source and Kelvin pins have to be connected as shown in Figure 19.7*b*, and the consequent loss in sensing accuracy due to the inclusion of the parasitic source resistance has to be accepted.

The effect of including the series sensing resistor in the sense circuit is to increase the sense ratio. Figure 19.8 shows the equivalent circuit with the sensing resistor included. The sense ratio is then given by

$$r = \frac{R_S}{R_M} + r_0 \tag{19.3}$$

Therefore it is possible to calculate the sense ratio for a given value of R_S, once

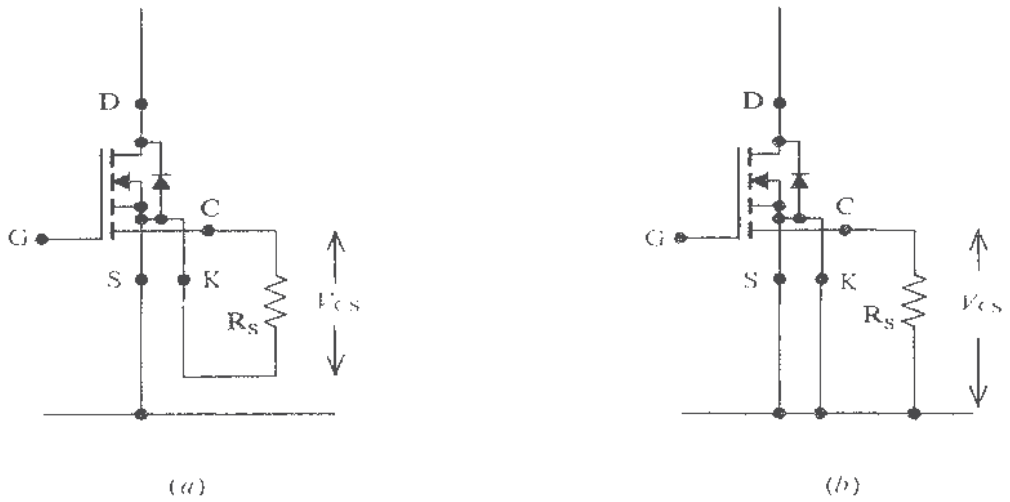
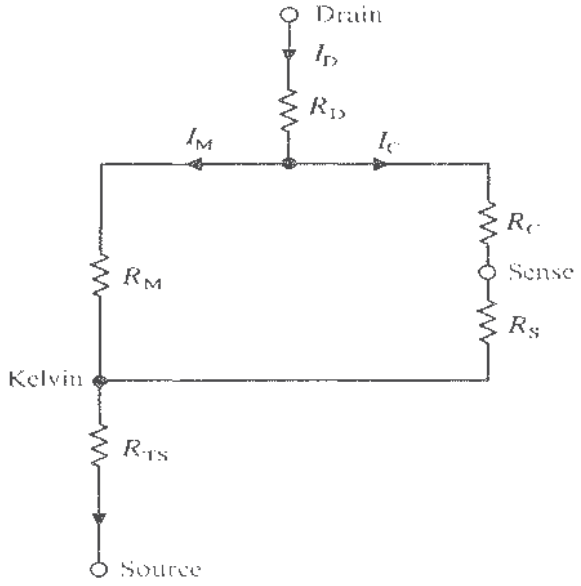


Figure 19.7. Sense-resistor connections. (*a*) Via the Kelvin source; (*b*) via the main source.

R_M and r_0 are known. However, R_M is the main device cell resistance; it is subject to manufacturing spread and is strongly dependent on temperature. Therefore r will vary with these parameters. The variation increases as R_S increases, until the limiting case of $R_S \gg R_C$ is reached when the voltage across R_C is essentially equal to the voltage across the main cell resistance. Then the main device cell resistance is being used to monitor device current, and the current sense output is equal to the drain voltage.

A second disadvantage associated with the resistor sensing method is that the



$$I_M/I_C = \frac{R_C + R_S}{R_M}$$
$$= \frac{R_C}{R_M} + \frac{R_S}{R_M}$$
$$= \left(\frac{1}{R_M}\right)R_S + r_0$$

Figure 19.8. Alteration of the current ratio due to the addition of a sensing resistor.

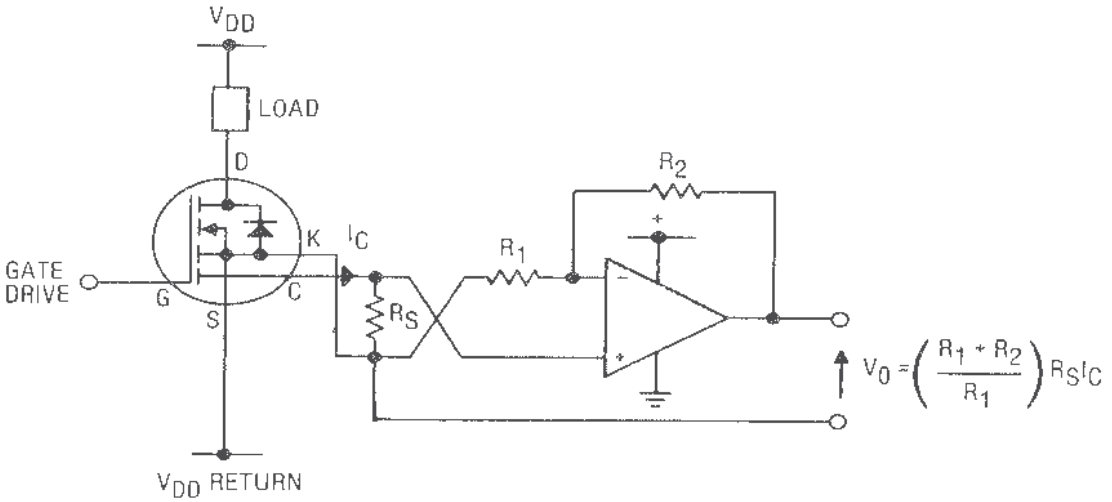


Figure 19.9. Resistor sensing circuit with amplification.

sensitivity range is limited. The circuit sensitivity is given by:

$$\text{sensitivity} = \frac{V_{\text{sense}}}{I_D} = \frac{R_S}{(r+1)} \tag{19.4}$$

As R_S increases, r increases and the maximum available sensitivity approaches a limit. To obtain a higher output voltage, or to allow the use of a low-value sensing resistor in order to reduce the effect of temperature on the sense ratio, an operational amplifier may be used to amplify and buffer the sense voltage as shown in Figure 19.9. The operational amplifier is connected so that a positive output is obtained and only a single positive supply is required. However, the input common-mode voltage range of the amplifier must include the lower power rail.

The static accuracy is dependent on operational-amplifier offsets. The input offset voltage of the amplifier adds directly to the sense voltage and therefore must be kept small relative to V_{sense} at the lowest current value that is of interest. Input bias currents may cause significant error if the noninverting effective source resistance is large.

19.1.9. Applications

Current-sensing MOSFETs find application in all circuits where current is a controlled parameter, for example in motor drives, where current and torque are directly related. Current sensing allows fault protection in power supplies, motor controllers, and other equipment in which overload protection is required. Start-up current limiting in lamp and motor loads can extend system life and reduce component cost. Current-waveform shaping is also possible for such applications as print-hammer drivers, solenoids, and stepper-motor windings.

The Kelvin contact of the current-sensing MOSFET may act as a gate return path in circuits in which very fast switching is required. By referencing the gate drive

circuit to the Kelvin pin, the common-mode source inductance associated with the main source lead is eliminated and faster turn-on is possible.

Power supplies often incorporate a current-sensing function which can be implemented with the current-sense MOSFET. In resonant power converters the current-sense signal may be used to implement zero-crossing turn-off. Also, current-mode control of power supplies, which requires sensing of the current in the switching device, is becoming widely used. Examples of this application are given in Chapter 11.

19.2. RADIATION-HARDENED POWER MOSFETs

19.2.1. Introduction

This section deals with the effects of certain kinds of radiation, both naturally occurring and man-made, on the operation of power electronic circuits containing power MOSFETs. Both ionizing and non-ionizing radiation are considered, along with the origins and damage effects of primary radiation interactions and secondary induced interactions. The effects of radiation can be limited by circuit design strategies and the use of radiation-resistant (rad-hard) MOSFETs [5–8].

19.2.2. Types of Radiation

The principal kinds of radiation in the operating environment of an electronic system are gamma rays, X-rays, neutrons, trapped protons and electrons, and cosmic rays. The rate of incidence of gamma rays is of considerable interest when one considers the effects of nuclear weapon bursts on electronics. This rate quantity is usually known as "gamma dot", or the time derivative of the gamma flux, expressed in rad (Si)/s. Another type of radiation, known as electromagnetic pulse (EMP), is generated by nuclear-weapon bursts. While of great interest to circuit and system designers, this does not affect devices and will not be discussed here.

Types of radiation may be broadly classified into two categories: ionizing and non-ionizing. Photon radiation in the gamma and X-ray part of the spectrum is a primary constituent of ionizing radiation. Gamma rays and X-rays are both photons, differing only in their wavelengths. Gamma rays have wavelengths of a fraction of an angstrom, while X-rays have wavelengths of several angstroms. X-rays interact with atoms primarily through the Compton effect, in which a loosely bound electron is scattered along with the incident photon, resulting in ionization of the atom. In a higher-energy interaction, a high-energy photon (gamma ray), upon approaching a nucleus, may be annihilated, resulting in the production of an electron–positron pair. At low photon energies, the matter interaction is via the photoelectric effect, in which an electron is scattered and the incident photon is absorbed. In all cases electrons are scattered. These electrons possess both charge and mass. They therefore readily interact with matter, causing electrical (and chemical) changes to occur. Secondary ionizing radiation is generated when these free electrons collide with the atoms, resulting in bremsstrahlung X-rays.

Trapped protons and electrons are contained by the earth's magnetic field in regions known as the Van Allen belts. These belts form a torus about the earth with the fields converging in the regions of the north and south poles. Additionally, a magnetic anomaly known as the South Atlantic Anomaly (SAA) causes a convergence of the field lines in that area. Protons, being both massive and charged, interact readily with matter, causing ionization damage. Due to their mass, they also cause displacement damage similar to that produced by neutrons. Electrons cause direct ionization damage and produce further secondary interactions via bremsstrahlung radiation.

In addition to protons and electrons, the near-earth radiation environment contains many heavy ions which stream in from the sun and from deep space. These ions (cosmic rays) mainly range from mass numbers around that of iron to simple protons. These particles are responsible for a form of radiation damage known as single-event upset (SEU), in which a single ion can cause a logic circuit to change state or fail. They also have the potential to cause failure in a power MOSFET. Various shielding strategies have been developed to deal with SEU, so that the particles expend their energy in some intervening medium.

The final constituent of the radiation environment is the neutron. The neutron, having no charge, does not cause direct ionization damage. However, it is relatively massive and causes displacement damage within crystal structures. It can also induce reactions which result in the emission of ionizing electrons and photons. Secondary gamma radiation can cause ionization damage even though the incident neutron does not. A similar situation exists when an excited nucleus returns to its ground state by emitting an electron or a photon.

19.2.3. Radiation Effects in MOSFETs

In crystal structures of the sort found in semiconductors, ionizing radiation forms electron–hole charge separations, or "pairs". These electron–hole pairs have two primary effects in a power MOSFET structure, depending upon where they are formed. In the relatively high mobility material which constitutes the MOSFET body and epitaxial diffusions, electron–hole pairs have a very brief existence. They either recombine or are swept out very rapidly. If the circuit configuration permits, a large pulse of drain–source current can result (Figure 19.10). In a

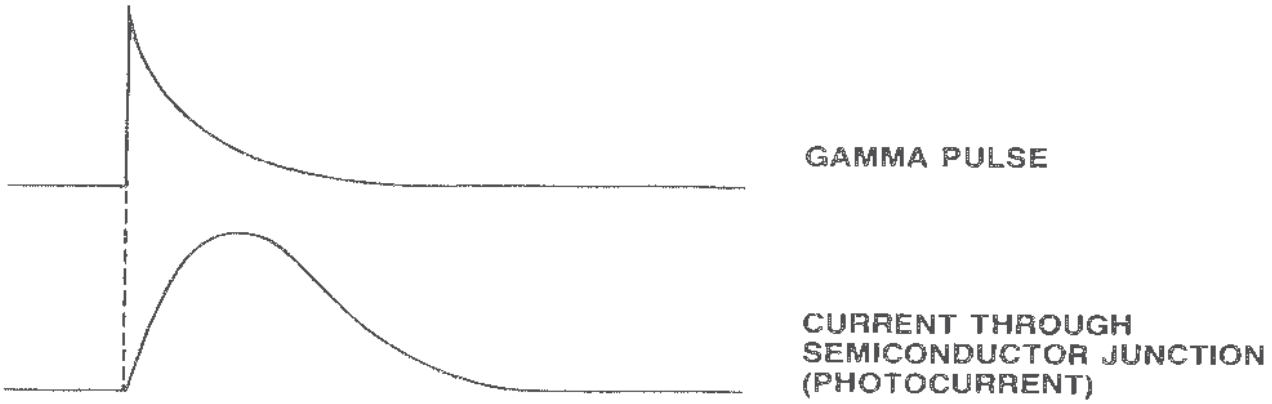


Figure 19.10. Photocurrent waveform produced by a pulse of gamma radiation.

bipolar transistor this may well result in second-breakdown failure. Power MOSFETs in which the parasitic bipolar transistor is effectively muted should be better able to resist damage due to radiation-induced turn-on, provided SOA failure of the parasitic bipolar transistor can be avoided [9–11].

If, however, the electron–hole pairs are generated in the very low-mobility areas of the gate oxide or interface states, recombination is very slow, resulting in persistent charge separations. The negative components (electrons) of these pairs are removed upon reaching the gate structure by virtue of the gate bias voltage applied to the MOSFET. This same gate bias voltage causes the positive components (holes) to collect at the oxide–channel interface, where they act as a permanent bias source. The proximity of these charges to the channel interface enhances this effect. The presence of this charge results in the equivalent V_{GS} of the MOSFET moving in the negative direction and eventually, after sufficient exposure to ionizing radiation, becoming negative. At this point, the MOSFET, which began as an enhancement-mode device, becomes a depletion-mode device (n-channel MOSFET).

The above discussion assumes low to moderate dose rates. At high dose rate [10^8 rad(Si)/s], recombination rates are insufficient to rid the structure of charge separations, resulting in reduced blocking voltage capability and very high drain currents. These photon induced currents (photocurrents) may be many times the steady current rating of the device. The only terrestrial source of such high gamma rates is a nuclear-weapon burst. Various circuit strategies have been developed to deal with this event and are described in Section 19.2.8.

Neutrons are primarily non-ionizing due to their lack of electric charge, and interact with semiconductor crystal structures by causing displacements of the constituent atoms. These defects (called Frankel defects) act to deplete the majority carriers (electrons) in the channel structure, resulting in increased resistivity. This effect is most pronounced in material where dopant concentrations are initially small, such as high-voltage MOSFETs. One may view this process as the canceling of donor atoms in the doped material, resulting in fewer free electrons available as carriers.

The secondary ionizing effects of neutron interactions should not be overlooked, as these effects became pronounced at high neutron fluence levels (10^{10} n/cm^2 s). Gamma rays, bremsstrahlung X-rays, and Compton electrons are generated when neutrons interact with both the semiconductor material and the material used to package the device. The ionizing effects due to the secondary radiation emissions following neutron bombardment are identical to those of incident gamma radiation.

19.2.4. Effects of Ionizing Radiation on the Characteristics of Standard Power MOSFETs

The main effect of ionizing radiation is to introduce charges into the gate oxide which produce a shift in the threshold voltage [12]. The threshold voltage of n-channel MOSFETs decreases with increasing total dose (Figure 19.11) while for p-channel MOSFETs the threshold voltage increases (Figure 19.12). This change of threshold is essentially independent of dose rate and depends only on the total

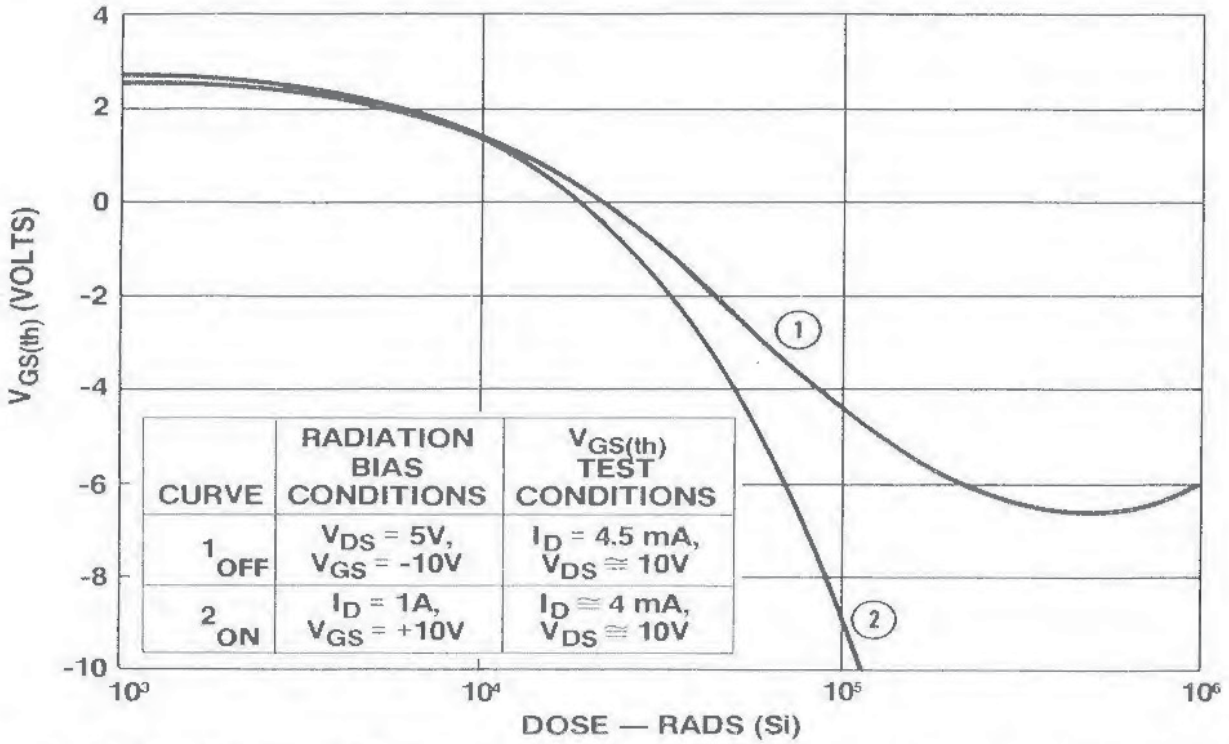


Figure 19.11. Gate threshold voltage versus total dose for a conventional n-channel power MOSFET (IRF 130).

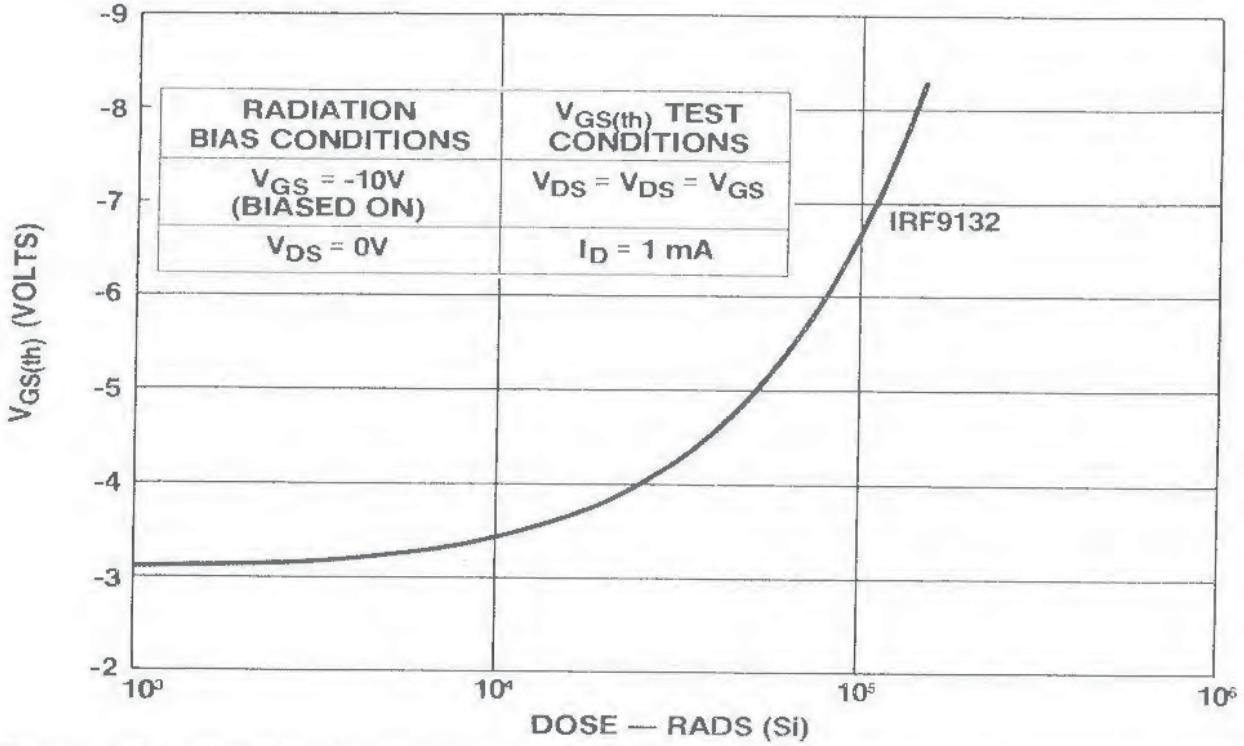


Figure 19.12. Gate threshold voltage versus total dose for a conventional p-channel power MOSFET (IRF 9132).

dose, except when the temperature excursions of the die are sufficient to accelerate recombination of this charge to a point where significant decay occurs.

As can be seen from Figure 19.11, the variation in threshold voltage is roughly independent of bias conditions for doses of up to about 30 krad. Thereafter the threshold voltage decreases much more rapidly for the biased-on condition. The threshold shift appears to be essentially independent of frequency and drain–source voltage [13]. Gate drive circuitry can be designed to nullify the threshold shift by overriding it with an appropriate level of biasing. For very heavy dose levels this may not be possible if the bias voltage required exceeds the maximum allowable gate–source voltage. Application of power to the main circuit may have to be delayed at switch-on in order to allow the gate bias-generating circuitry to become operative.

19.2.5. Effects of Ionizing Radiation on Leakage Current

Figure 19.13 shows the effect on drain–source leakage current of total radiation dose for a p-channel power MOSFET. Since for a p-channel MOSFET the threshold voltage becomes more negative with increasing total dose, the increased leakage current cannot be due to the threshold voltage approaching the gate voltage. Leakage current increases with radiation dose due to the trapping of charge in the oxide covering the field termination structure. This charge causes a local increase in electric field and leads to increased leakage current or, conversely, breakdown at a lower voltage. However, experiments on p-channel MOSFETs suggest that this explanation is simplistic and in fact the mechanisms at work are more complex [14]. Generally, the decrease in breakdown voltage increases with the applied drain–source voltage during irradiation [15]. The fall in blocking voltage will

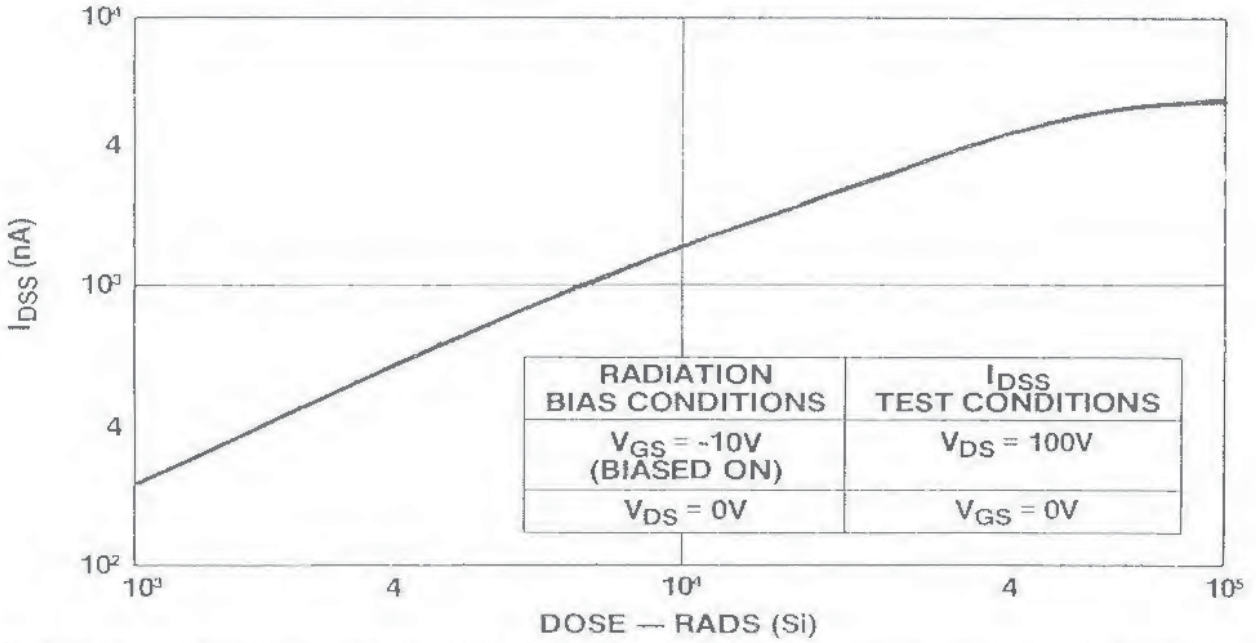


Figure 19.13. Drain–source leakage current versus total dose for a conventional p-channel power MOSFET (IRF 9132).

require the selection of a device with a voltage rating high enough to accommodate the fall in blocking voltage produced by the radiation. This results in a higher $R_{DS(on)}$ and a device that is more susceptible to a shift in $R_{DS(on)}$ due to neutron radiation.

19.2.6. Effects of Non-ionizing Radiation on the Characteristics of Standard Power MOSFETs

The effect of neutron radiation is to increase the $R_{DS(on)}$ of the power MOSFET. An empirically determined expression [16] gives the relationship between resistivity and neutron fluence:

$$\rho_n = \rho_0 \exp\left(\frac{\Phi}{444 n_0^{0.77}}\right) \tag{19.5}$$

where ρ_n is the resistivity after exposure, ρ_0 is the resistivity before radiation, Φ is the neutron fluence, and n_0 is the starting dopant density. The increase of $R_{DS(on)}$ depends on the resistivity of the silicon, and therefore on the MOSFET voltage rating [17]. For 100-V MOSFETs the increase is small, but at 400-V it is significant (Figure 19.14). The increase of ON resistance will result in increased conduction losses in switching applications but will have little effect in linear applications. In switching applications allowance may be made for the increased $R_{DS(on)}$ by selecting a device with an appropriately low initial value of $R_{DS(on)}$. Alternatively, the increase in $R_{DS(on)}$ can be accepted and allowance made for the increased power dissipation by use of an appropriate-sized heatsink. Neutron

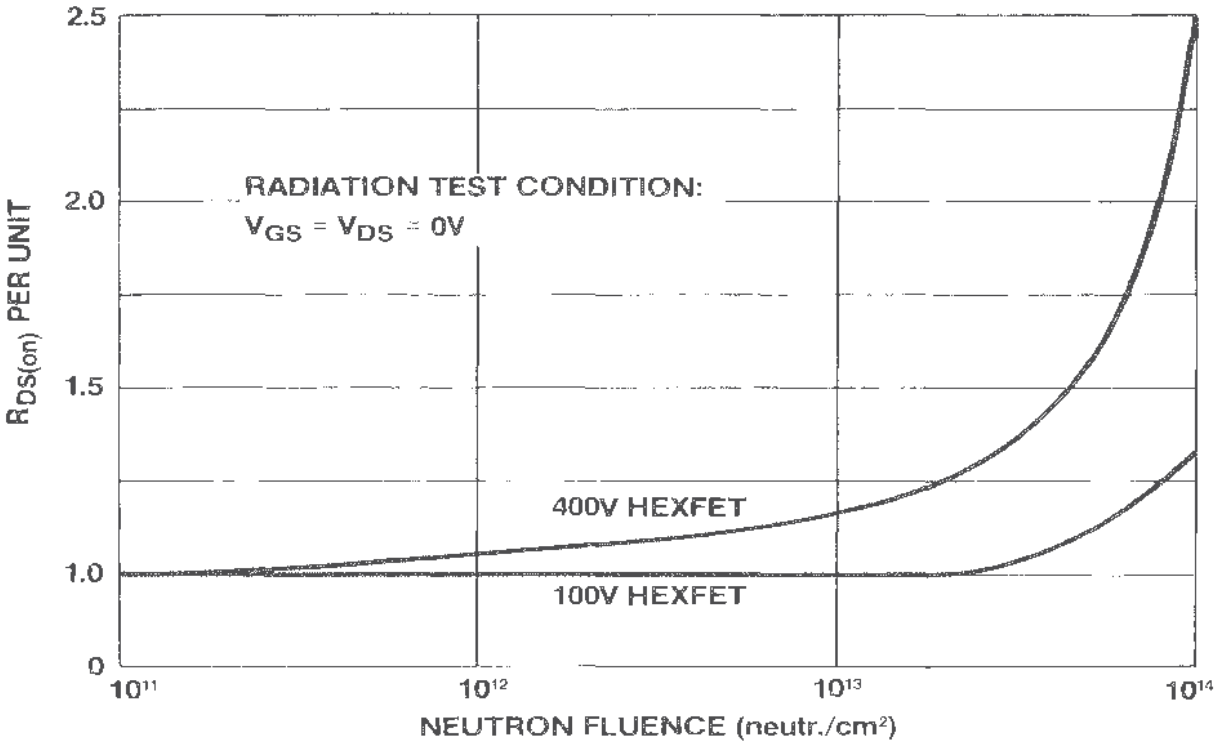


Figure 19.14. ON resistance versus neutron fluence level for conventional MOSFETs.

radiation also has a major effect on the forward voltage drop of the IGBT (see Section 20.2). This is due more to the effect of the radiation on minority-carrier lifetime than to an increase in the resistivity of the silicon [18].

19.2.7. Radiation Countermeasures

In order to construct electronic circuits capable of operating reliably in hostile radiation environments, strategies must be adopted to protect the semiconductor devices. These strategies fall into three broad categories:

1. System shielding.
2. Radiation-tolerant circuit topologies.
3. Radiation-hardened semiconductors.

In system shielding a suitable material is interposed between the radiation source and the electronics. This material is chosen to deplete the energy of the incident radiation by collisions with the material atoms or absorption in nuclear reactions [19]. The degree to which this is used depends on the application. In stationary land-based applications, such as in nuclear power plants, lead shielding is practical but in space applications weight restrictions severely limit the amount of shielding that is possible. Similarly, more shielding is possible in land vehicles and ships than in aircraft.

19.2.8. Circuit Designs to Protect MOSFETs against the Effects of Radiation

Radiation-resistant circuit topologies are ones in which there is reduced sensitivity of the circuit to abnormal behavior of the semiconductor components, the objective being to prevent permanent damage to any circuit component. For example, circuits may be designed so that passive components limit the current which flows between the power rails if all semiconductor components lose their blocking capability [20]. Conventional buck converter and boost converter circuits are unsuitable in this respect, as Figure 19.15 shows. Push-pull and resonant

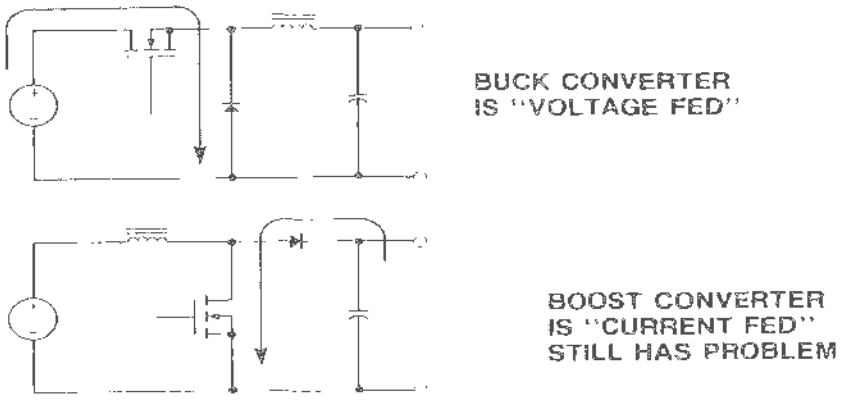


Figure 19.15. Conventional converter circuits experience short circuits between the power rails when semiconductors lose their blocking capability due to gamma radiation.

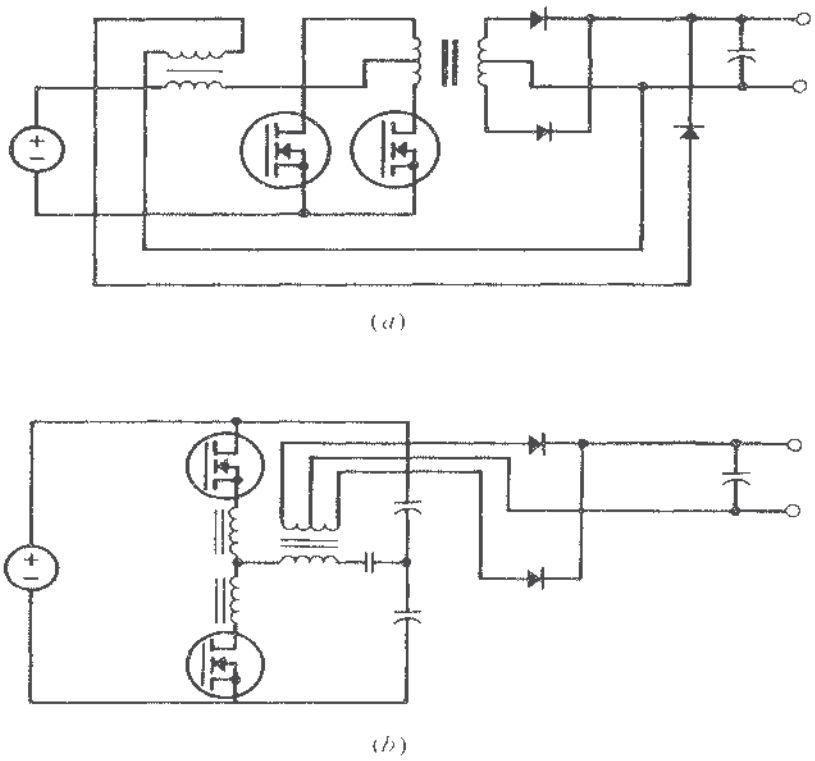


Figure 19.16. Power converter circuits capable of surviving prompt gamma pulses. (*a*) Weinberg converter; (*b*) resonant converter.

circuits, such as those shown in Figure 19.16, rely on inductance to absorb power-rail voltage for the brief period during a gamma-ray pulse when all semiconductor devices are likely to be conducting. Alternatively a radiation detector, such as a p–i–n diode, can initiate shutdown of the equipment with crowbarring of supply rails, as illustrated in Figure 19.17. As well as crowbarring, all FET channels can be turned on to divert current away from the parasitic bipolar transistor, thereby reducing the chance of second breakdown.

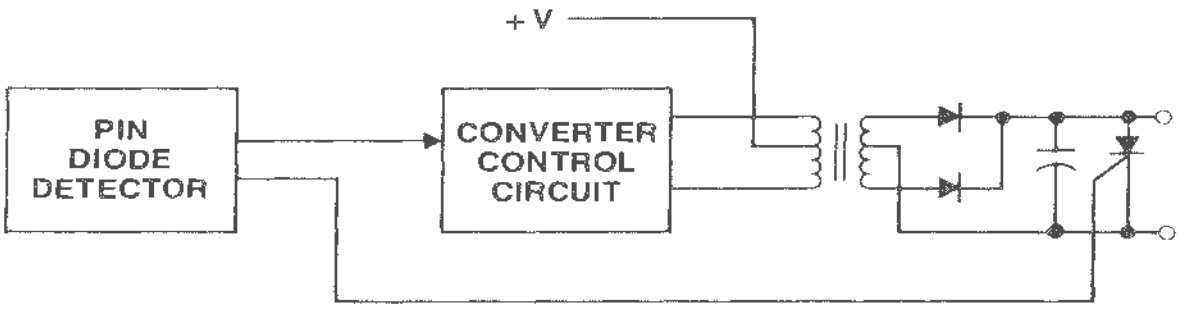


Figure 19.17. Circuit protection by radiation detection and shutdown. Circuits can be protected against damage if the gamma radiation is detected by a p-i-n diode which shuts down the circuit and triggers the crowbar.

19.2.9. Radiation-Resistant Power MOSFETs

Radiation hardened ("rad-hard") power MOSFETs have been successfully produced and continue to be a focus of intense effort. Manufacturers are as yet reluctant to discuss the methods used to obtain resistance to radiation, in part because of the obvious military applications for these devices, although some work has been reported [21, 22]. It is clear that the gate-oxide structure traditionally employed in power MOSFETs requires modification to reduce the threshold shift resulting from ionizing radiation [23]. There can be no doubt that presently available rad-hard MOSFETs are capable of withstanding large doses of radiation without suffering disabling loss of tranconductance or a major shift in threshold voltage.

Assessing the quality of radiation-resistant devices poses some difficulties, since the device is altered by radiation and radiation sources are not always readily available. Nondestructive testing of some characteristics is possible [24]. Where verification of product quality involves irreversible changes in the device tested, the product has to be sampled. The number of samples required to obtain statistically valid results may be reduced by testing the samples beyond normal limits [25]. Bias conditions applied to the device in the minutes after exposure to radiation influence the degree of permanent threshold shift and therefore may affect the accuracy of the measurements [26].

19.3. LOW-THRESHOLD-VOLTAGE MOSFETs

In an attempt to make power MOSFETs more compatible with 5-V logic, some manufacturers have introduced MOSFETs with a threshold voltage that is typically half that of standard ones. They can therefore be driven directly from TTL or other 5-V logic families without the need for level shifting. The reduction in threshold voltage is usually achieved by a reduction in the thickness of the gate oxide, typically from 1000 to 500 Å [27]. Since the gate oxide is thinner, less voltage is required to produce the same electric field in the silicon under the gate and the same inversion depth. Halving the thickness of the gate oxide halves the threshold voltage and doubles the transconduetance of the MOSFET. The threshold voltage may also be reduced by a reduction of the doping density of the body region [28], but considerations such as the need to avoid punchthrough restrict the adjustment to the threshold voltage that can be made by this means.

Since low threshold voltage and high transconductance are generally desirable features, it is worth considering why the oxide thickness in ordinary power MOSFETs is 1000 Å and not 500 Å. The primary reason is that it becomes more difficult to grow an oxide layer of the required integrity as the oxide thickness decreases. Defects and variation in thickness become a more serious problem as the oxide thickness approaches the dimension of the defect. Furthermore, hot-electron effects, ion migration, and a lower activation energy for gate-failure phenomena all make it more difficult to maintain an acceptable lifetime for the gate. However, despite these difficulties, power MOSFETs with a threshold voltage in the range of 1 to 2 V can be made and are widely available.

Halving the gate oxide thickness increases gate-source and drain–source capacitances. An increase in these capacitances will tend to increase switching

times, while the lowering of the threshold voltage has the opposite effect. For most of the turn-on period when the drain voltage is falling, the drain region is separated from the gate by the parasitic JFET between cells (see Section 3.6.2). However, the transconductance of the device is higher, giving an enhanced Miller effect. There are thus a number of opposing influences on the switching time. A direct comparison of the switching speed of low-threshold MOSFETs with that of ordinary MOSFETs is difficult to make, since the gate-drive conditions are necessarily different for the two types of device. If the gate drive is assumed to be a voltage generator in series with a resistor, the value of the voltage generator should clearly be halved for the low-threshold device (for a device with half the threshold voltage), while the series resistor may either be kept the same, be halved to give the same initial value of gate current, or be reduced by a factor of four to give the same initial gate power. If one maintains the same initial gate power, the switching time of the low-threshold MOSFET is almost half that of an equivalent conventional device [29].

Applications for low-threshold power MOSFETs include interfacing with 5-V logic circuits and use in battery-operated equipment where insufficient voltage may be available to fully turn on normal-threshold voltage devices. However, in many interface applications, where the load is an LED, a lamp, a relay, or a solenoid, the fast switching capability of the power MOSFET is not of particular benefit, and the choice between the power MOSFET and the bipolar Darlington transistor in this application tends to hinge on cost. It should not be assumed that full enhancement of the logic-level MOSFET is achievable when driven from any 5-V logic. For example, the guaranteed minimum high-level output voltage of low-power Schottky logic (2.4 V) is inadequate for full enhancement of most logic-level power MOSFETs [30].

REFERENCES

1. N. Zommer and J. Biran, "Power current mirror devices and their applications." *Proc. Power Convers. Int. Conf.* June., pp. 275–283 (1986).
2. C. E. Cordonnier and D. L. Zaremba, "Application considerations for SenseFET power MOSFET devices." *Proc. Power Convers. Int. Conf.* Munich, May, pp. 47–65 (1987).
3. D. A. Grant and R. Pearce, "The dynamic performance of current sensing power MOSFETs." *Electron. Lett.* **24**(18), 1129–1131 (1988).
4. D. L. Zaremba, Jr., "Current sense power MOSFETs." *Power Electron. Conf.* pp. 1–7 (1988).
5. G. C. Messenger and M. S. Ash, *The Effects of Radiation on Electronic Systems*. Van Nostrand-Reinhold, New York, 1986.
6. J. F. Kircher and R. E. Brown, *Effects of Radiation on Materials and Components*. Reinhold, New York, 1964.
7. U. S. Defense Dept., *The Effects of Nuclear Weapons*. Dept. of Energy, U.S. Gov. Printing Office, Washington, DC.

8. N. Rudie, *Principles and Techniques of Radiation Hardening*. Western Periodical Co., Hollywood, CA, 1976.
9. H. Volmerange and A. Witteles, "Radiation effects on MOSPOWER transistors." *IEEE Trans. Nucl. Sci.* **NS-29**(6), 1565–1568 (1982).
10. W. R. Dawes, Jr., T. A. Fischer, C. C.-C. Huang, W. J. Meyer, C. S. Smith, R. A. Blanchard, and T. J. Fortier, "Transient Hardened Power MOSFETS." *IEEE Trans. Nucl. Sci.* **NS-33**(6), 1425–1427 (1986).
11. A. E. Waskiewicz, J. W. Groninger, V. H. Straham, and D. M. Long, "Burnout of power MOS transistors with heavy ions of californium-252." *IEEE Trans. Nucl. Sci.* **NS-33**(6), 1710–1713 (1986).
12. D. L. Blackburn, D. W. Berning, J. M. Benedetto, and K. F. Galloway, "Ionizing radiation effects on power MOSFETS during high-speed switching." *IEEE Trans. Nucl. Sci.* **NS-29**(6), 1555–1558 (1982).
13. S. S. Seehra and W. J. Slusark, "The effect of operating conditions on the radiation resistance of VMOS power FETS." *IEEE Trans. Nucl. Sci.* **NS-29**(6), 1559–1564 (1982).
14. R. D. Pugh, A. H. Johnston, and K. F. Galloway, "Characteristics of the breakdown voltage of power MOSFETS after total dose irradiation." *IEEE Trans. Nucl. Sci.* **NS-33**(6), 1460–1464 (1986).
15. D. L. Blackburn, J. M. Benedetto, and K. F. Galloway, "The effect of ionizing radiation on the breakdown voltage of power MOSFETS." *IEEE Trans. Nucl. Sci.* **NS-30**(6), 4116–4121 (1983).
16. M. G. Buehler, "Design curves for predicting fast-neutron-induced resistivity changes in silicon." *Proc. IEEE* **56**(10), 1741–1743 (1968).
17. D. Blackburn, T. C. Robbins, and K. F. Galloway, "VDMOS power transistor drain-source radiation dependence." *IEEE Trans. Nucl. Sci.* **28**(6), 4354 (1981).
18. A. R. Hefner, D. L. Blackburn, and K. F. Galloway, "The effect of neutrons on the characteristics of the insulated gate bipolar transistor (IGBT)." *IEEE Trans. Nucl. Sci.* **NS-33**(6), 1428–1434 (1986).
19. A. Holmes-Siedle and R. F. A. Freeman, "Improving radiation tolerance in space-borne electronics." *IEEE Trans. Nucl. Sci.* **NS-24**(6), 2259–2265 (1977).
20. W. E. Abare and W. K. Martindale, "Dose rate tolerant HEXFET power supply." *IEEE Trans. Nucl. Sci.* **NS-28**(6), 4380–4383 (1981).
21. G. B. Roper and R. Lowis, "Development of a radiation hard N-channel power MOSFET." *IEEE Trans. Nucl. Sci.* **NS-30**(6), 4110–4115 (1983).
22. J. M. McGarrity, "Considerations for hardening MOS devices and circuits for low radiation doses." *IEEE Trans. Nucl. Sci.* **NS-27**(6), 1739–1744 (1980).
23. G. Singh, K. F. Galloway, and T. J. Russell, "Radiation-induced interface traps in power MOSFETS." *IEEE Trans. Nucl. Sci.* **NS-33**(6), 1454–1459 (1986).
24. H. E. Boesch, Jr. and J. M. McGarrity, "An electrical technique to measure the radiation susceptibility of MOS gate insulators." *IEEE Trans. Nucl. Sci.* **NS-26**(6), 4814–4818 (1979).
25. A. I. Namenson, "Hardness assurance and overtesting." *IEEE Trans. Nucl. Sci.* **NS-29**(6), 1821–1826 (1982).
26. P. V. Dressendorfer, J. M. Soden, J. J. Harrington, and T. V. Nordstrom, "The effects of test conditions on MOS radiation hardness results." *IEEE Trans. Nucl. Sci.* **NS-28**(6), 4281–4287 (1981).
27. C. F. Wheatley and H. R. Ronan, "Switching waveforms of the LEFET: A 5-volt gate-drive power MOSFET." *IEEE Power Electronics Spec. Conf.* pp. 238–246 (1984).

28. M. Sciammas and D. Kapur, "Low threshold power FETs: A sensible solution," *Powertechnics* June, pp. 28–31 (1986).

29. L. Marechal, "The logic level FET at switching," *Proc. Power Convers. Inter. Conf.* pp. 197–208 (1984).

30. J. E. Wojslawowicz., "L²FETs challenge bipolar power devices," *Power Convers. Intell. Motion (PCIM)* January, pp. 27–30 (1988).

APPENDIX 5

Cell Geometry

In Section 3.4 the device transconductance g_m, and hence the current-controlling capability of the MOSFET, are shown to be proportional to the gate width w. As a result, an important design objective must be to obtain as wide a gate as possible on a given area of silicon.

In principle, VDMOS FET gates may be laid out as linear arrays, interdigitated with the source, as shown in Figure 3.9*a*. Then, the gate width per unit area is simply $w/A = 1/a$, where a is the pitch of the array. Photolithographic line definition and mask registration set a lower limit to the pitch size.

Aside from these considerations, it is also important that spacing between the channel diffusions, b, should always exceed a certain minimum value. The reason for this is that built into the VDMOS structure is a parasitic JFET in series with the drain. This is described in Section 3.6. It is most likely to have a serious effect in high-voltage devices, in which the epitaxial layer is more lightly doped. What happens is that the conduction path below the gate is pinched off by the depletion layers that form around the junction with the p-type regions.

The alternative to the linear gate geometry is a cellular array. There are several reasons why, in practice, the latter is normally preferred. By comparison, the linear structures have a lower packing density (w/A), they give rise to a higher device capacitance, and they usually require rather more complicated photolithographic masks. The long gate fingers give rise to significant resistance when a polycrystalline silicon gate is used, and this causes turn-on delays. However, one manufacturer advocates the use of a linear metal gate structure, in conjunction with a simple four-mask fabrication process, for producing inexpensive high-voltage devices. Another, quite sophisticated interdigitated design has been made for high-frequency applications [1].

Several different cellular structures have been proposed, and of these, a few remain in widespread production by different manufacturers. Designs have used cells of hexagonal, square, and even triangular shape, and they have been based on square or hexagonal lattices. Some of the options are illustrated in Figure 3.9. Technological advances have led to a reduction of the cell pitch, and hence to larger values of gate width for a given chip size. More than three million cells per square inch have been obtained, giving gate widths exceeding 1 m on a 10-mm^2 chip (100 × 150 mil). Reducing cell dimensions also gives rise to a more uniform current distribution in lower-voltage devices, for reasons discussed in Section 3.5, and this brings the further benefit that it lowers $R_{DS(on)}$.

Let us assume that the need to avoid the effects of the parasitic JFET, together with the level of technology used in fabrication, sets a lower limit on the spacing

between adjacent channel diffusions under the gate oxide. Note that this distance, b, is about 5 μm less than the breadth of the polysilicon arms forming the gate web. (It must be clearly distinguished from the gate *width*, w, which depends on the total length of these arms.) Then, for each of the possible cell structures there is an optimum cell pitch that maximizes the value of w/A, the gate width per unit area. Consider the case shown in Figure 3.9*b*, where hexagonal cells of side s lie on a hexagonal lattice of pitch a. Each cell occupies an area $\sqrt{3}\,a^2/2$ and provides a gate width of $6s$. The geometry of the configuration means that $a = b + \sqrt{3}\,s$. Thus,

$$\frac{w}{A} = \frac{12s}{\sqrt{3}\,(b + \sqrt{3}\,s)^2} \tag{A5.1}$$

Differentiating with respect to s and setting the derivative equal to zero shows that w/A is a maximum when $s = b/\sqrt{3}$, and hence $a = 2b$. Thus,

$$\left(\frac{w}{A}\right)_{\max} = \frac{1}{b} = \frac{1}{\sqrt{3}\,s} = \frac{2}{a} \tag{A5.2}$$

Results for other cell patterns may be derived similarly and are set out in Table A5.1. On this argument the hexagon-on-hexagon, square-on-square, and triangle-on-hexagon arrangements are equally good, and have a slight advantage over the others.

Clearly, the smaller b can be made, and with it a, the larger is $(w/A)_{\max}$. However, the limit on b is far from being the only constraint. In particular, the need to make a low-resistance ohmic contact to the source means that the cell size

TABLE A5.1. Optimum Cell Geometries

Figure	Configuration	Constraint	Condition for $(w/A)_{\max}$	$(w/A)_{\max}$
3.9(*a*)	Linear	$a = b + s$	$s \to 0$	$1/a = 1/b$
3.9(*b*)	Hexagons on hexagonal lattice	$a = b + \sqrt{3}\,s$	$s = b/\sqrt{3}$	$1/b = 1/\sqrt{3}\,s = 2/a$
3.9(*c*)	Hexagons on square lattice	$a = b + 2s$	$s = b/2$	$3/4b = 3/8s = 3/2a$
3.9(*d*)	Squares on square lattice	$a = b + s$	$s = b$	$1/b = 1/s = 2/a$
3.9(*e*)	Squares on an offset square lattice	$a = b + \sqrt{2}\,s$	$s = b/\sqrt{2}$	$1/\sqrt{2}\,b = 2/s = \sqrt{2}/a$
3.9(*f*)	Squares on a hexagonal lattice	$a = (2/\sqrt{3})(b + s)$	$s = b$	$\sqrt{3}/2b = \sqrt{3}/2s = 2/a$
3.9(*g*)	Triangles on hexagonal lattice	$a = \sqrt{3}(\sqrt{3}\,b + s)$	$s = \sqrt{3}\,b$	$1/b = \sqrt{3}/s = 6/a$

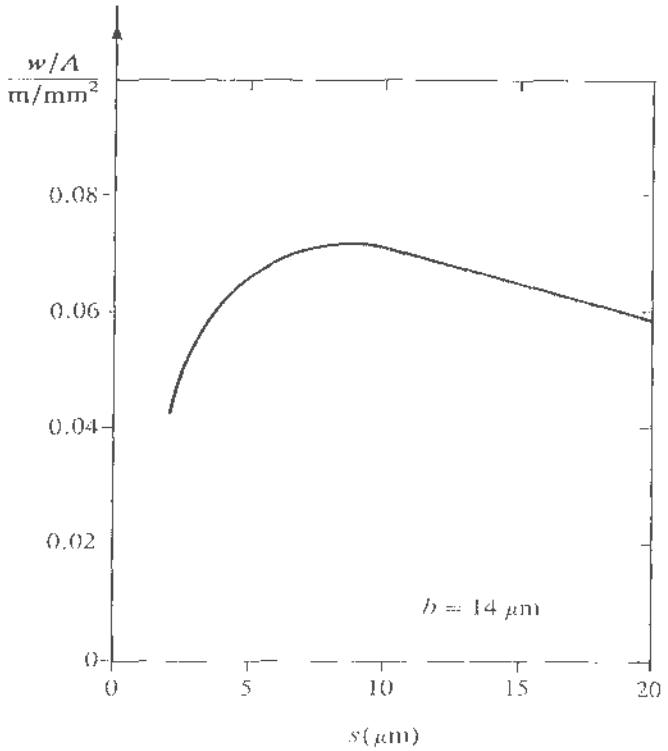


Figure A5.1. Variation of gate width per unit area with cell size. Hexagons on a hexagonal lattice with $b = 14\ \mu m$.

s may not be made too small either. As a result, it is normal practice for the cell pitch a to be made rather larger (for a given value of b) than the optimum cell pitch predicted by the simple theory and set out in Table A5.1. The important parameter $(w/A)_{max}$ is not very sensitive to such variations, as can be seen in Figure A5.1. Thus, the HEXFET shown in Figure 3.10*a* has $b = 14\ \mu m$ but $a = 38\ \mu m$ ($s = 14\ \mu m$). Even so, the gate width per unit area is 0.067 m/mm^2, compared with the optimum value of 0.071 m/mm^2 when $a = 28\ \mu m$ ($s = 8\ \mu m$).

There are two other important considerations. The first is concerned with the added resistance that results from the higher current density under the gate oxide between the p diffusions. This is less when the shortest distance between adjacent cells lies between their vertices rather than their sides. This goes some way to redress the slight disadvantage of the arrangements of Figures 3.9*c*, *e*, and *f*. The second consideration is the need for an even distribution of current around the whole width of the channel. This is obtained more easily with the more obtuse vertices of the hexagonal cell than it is with square or triangular cells.

Overall it can be said that the theoretical advantages of one cell arrangement over another are at most marginal. What matters is the exploitation of their respective merits by the development of sophisticated, specialized manufacturing techniques.

REFERENCE

1. D. Fuoss, "Vertical DMOS power field-effect transistors optimized for high-speed operation." *Proc. IEEE Int. Electron Devices Meet.* **82**, 250–253 (1982).

ISBN 0-471-82867-X