8.2.5 Effect of Thermal Oxidation on the Diffusion Coefficient*

The thermal oxidation of silicon produces a higher-than-equilibrium number of silicon interstitials that will in turn increase the diffusivity of any atom that has an interstitialcy component. Conversely, if an atom diffuses solely by vacancies, extra interstitials will depress the number of vacancies present and decrease the diffusivity. Boron, phosphorus, and arsenic all show enhancements, while antimony shows a retardation. The enhancement effect is seen in both wet and dry oxidations and is more pronounced at lower oxidation/diffusion temperatures. Early work reported a retarding effect above 1150°C–1200°C, which corresponds approximately to the temperature where oxidation-induced stacking faults begin to shrink, but more recent data merely show the effect disappearing at high temperature.

The number of interstitials generated at the interface will depend on the oxidation rate, which decreases with time (thickness). Plus, some of the interstitials diffuse into the oxide and recombine with incoming oxygen, while the rest diffuse into the silicon and eventually recombine. If the additional interstitials affect only the interstitial component of D, then, from Eq. 8.10, $D = D^*(1 + f_I N_I / N_I^*)$, and the average diffusivity $<D>$ over a diffusion/oxidation time t is given by

$$<D> = \frac{1}{t}\int_0^t Dt = D^*\left(1 + \frac{f_I}{N_I^*}\right)\int_0^t (N_I - N_I^*)dt \qquad 8.22$$

N_I is the concentration of interstitials where diffusion is occurring, and N_I^* is their equilibrium concentration. Various complex expressions for the interstitial supersaturation have been derived (16–20). It is predicted that for low temperatures and dry oxygen, the supersaturation starts at zero and increases with time. At longer times, it is experimentally deduced to be given by

$$N_I - N_I^* = K\left(\frac{dx}{dt}\right)^p \qquad 8.23$$

where p is in the 0.2–0.4 range and K is a constant. Since dx/dt decreases as the oxide layer thickness increases, the supersaturation decreases with time in this region. The supersaturation also decreases with depth into the wafer. A simple analysis gives

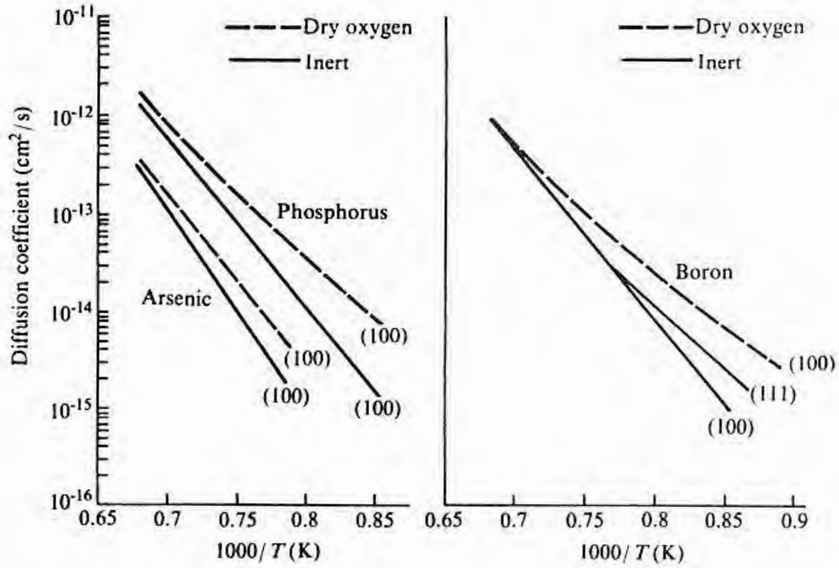
$$N_I - N_I^* \sim e^{x/L} \qquad 8.24$$

where L is a characteristic length, measured experimentally to be 30 μm at 1000°C and 25 μm at 1100°C (21). Fig. 8.6 shows the magnitude of the enhancement of D for boron, arsenic, and phosphorus.

*See references 11–20.

FIGURE 8.6

Effect of oxidation during diffusion on diffusion coefficient. (*Source:* D.A. Antoniadis et al., *Appl. Phys. Lett. 33*, p. 1030, 1978, and D.A. Antoniadis et al., *J. Electrochem. Soc. 125*, p. 813, 1978. Reprinted by permission of the publisher, The Electrochemical Society, Inc.)



8.2.6 Effect of Diffusing Direction (Orientation)

If the crystal in which diffusion is occurring is anisotropic, the value of D will depend on direction. In the general case, Eq. 8.4 becomes

$$J_i = D_{ij}\, \partial N \partial x_j \tag{8.25}$$

from which it is seen that the diffusion coefficient is a second-rank tensor. In cubic crystals, examples of which are Si and GaAs, properties that are second-rank tensors are independent of crystallographic orientation (22) (which is the reason that Eq. 8.5 is identical to Eq. 8.7). However, recalling that, depending on the diffusion mechanism, a deviation of the density of vacancies and/or interstitials may affect D, it can be surmised that a nonisotropic flow of either of them could cause a nonisotropic effective D. It is, in fact, experimentally observed that in silicon, orientation-dependent diffusions are often encountered (23–26) during simultaneous thermal oxidation and diffusion. For diffusion in an inert atmosphere, there is no evidence of anisotropic diffusion coefficients (14, 27). This orientation-dependent phenomenon can be explained by the different rates at which nonequilibrium numbers of interstitials are generated by oxidation on different crystallographic surfaces. The effect becomes greater as the temperature is lowered, and its magnitude at any temperature is in the order of (100) > (110) > (111). Impurities showing the effect will be the same ones showing oxidation-enhanced diffusion (OED), and the magnitude of the effect can be judged by the curves of Fig. 8.6.

8.2.7 Effect of Heavy Doping

As was shown in the discussion on vacancies, the density of charged point defects can change with doping concentration when the level approaches n_i. Most silicon dopants have a diffusion component depending on charged vacancies and will thus show a doping-level dependency. In addition, the field enhancement discussed in the previous section (and which becomes noticeable only when $N > n_i$) must also be included where appropriate. In the case of boron, and perhaps the other acceptors as well, diffusion is apparently by a neutral complex—for example, $B^- V^+$. Arsenic and antimony diffuse by V^0 and V^-; phosphorus, by V^0 and $V^=$; and boron and gallium, by V^0 and V^+. Thus, the expressions for D^* are as follows (28):

$$\text{As, Sb} \qquad D = g\left(D^{*0} + D^{*-}\frac{n}{n_i}\right)$$

$$\text{P} \qquad D = g\left[D^{*0} + D^{*=}\left(\frac{n}{n_i}\right)^2\right]$$

$$\text{B, Ga, Al} \qquad D = \left(D^{*0} + D^{*+}\frac{p}{n_i}\right)$$

When $N > n_i$, the neutral vacancy contribution to the total D generally becomes small. If, at the same time, the diffusing profile is such that an appreciable electric field occurs and g (from Eq. 8.21) approaches 2, then, if D^* is the diffusivity in intrinsic material, for donors,

$$\text{As, Sb} \qquad D \rightarrow 2D^*\frac{n}{n_i}$$

$$\text{P} \qquad D \rightarrow 2D^*\left(\frac{n}{n_i}\right)^2$$

For acceptors, field-aided diffusion appears to be negligible, so

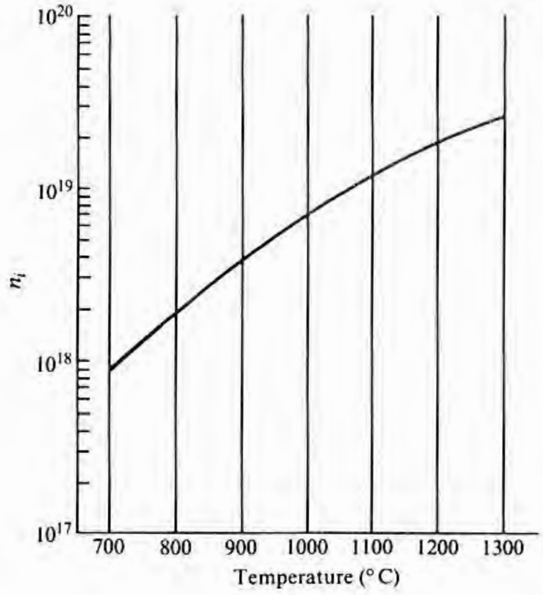
$$\text{B, Ga, Al} \qquad D \rightarrow D^*\frac{p}{n_i}$$

In intermediate doping ranges and at some temperatures, contributions from charge states other than those just listed may also become important. Fig. 8.7 is a curve of n_i versus temperature for silicon and can be used to tell when the doping concentrations are in the region where these effects may occur.

When the concentration of dopants in silicon approaches the 10^{21} atoms/cc range, precipitates, complexes, and atomic misfit strain begin to play a role in diffusion. For arsenic, complexes begin to form at slightly above 10^{20} atoms/cc, have a reduced diffusivity,

FIGURE 8.7

n_i versus temperature for silicon. (*Source:* From data in F.J. Morin and J.P. Maita, *Phys. Rev.* 96, p. 28, 1955.)



and are not fully ionized. The result is that D_{As} versus concentration has a maximum as shown in Fig. 8.8 (29). This figure also shows the increase in diffusivity with concentration up to the point where complexes form. In the case of phosphorus, the strain effect (see next section) becomes important for concentrations above about 4 × 10^{20} atoms/cc.

8.2.8 Effect of Temperature

From Eq. 8.6, the simple expression for D is of the form $D = D_0e^{-E/kT}$. When the doping level is below n_i over the whole temperature range being covered and when there are no oxidation enhancement effects—that is, in the intrinsic diffusivity range—Eq. 8.6 is appropriate. E will be different for each element, and in cases where differing diffusing conditions can favor a particular charge state for the point defect involved, different E values for each state are required. Diffusion data are sometimes presented in graphic form, in which case D versus $1/T$ gives a straight line if Eq. 8.6 is appropriate. When the data give a straight line, an alternative is to present D_0 and E in tabular form, as is done in Table 8.2 for silicon. Unfortunately, most gallium arsenide diffusants are not well enough behaved to be depicted in this manner. However, the table still lists various n- and p-dopants.

8.2.9 Effect of Stress/Strain

Lattice strain will change the bandgap (30, 31) and thus can affect the distribution of vacancies and the diffusivity (28, 31). It has also been suggested that, in the case of GaAs, it affects the jump fre-

FIGURE 8.8

Effect of concentration on arsenic diffusivity. (The peak position at intermediate temperatures follows the line.)
(*Source:* Adapted from R.B. Fair, *Semiconductor Silicon/81*, p. 963, Electrochemical Society, Pennington, N.J., 1981.)

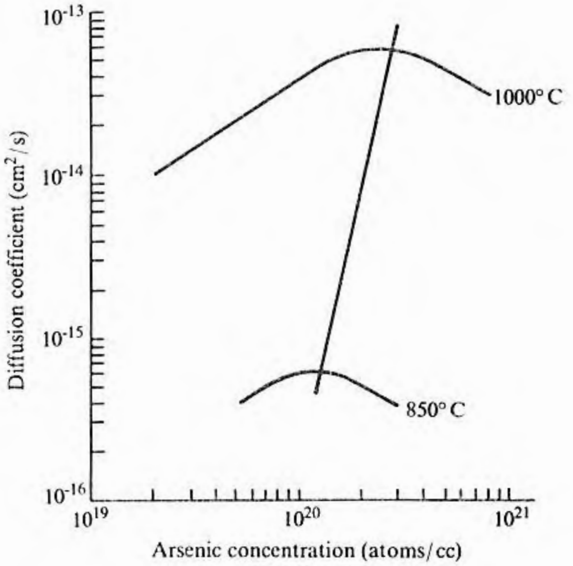


TABLE 8.2

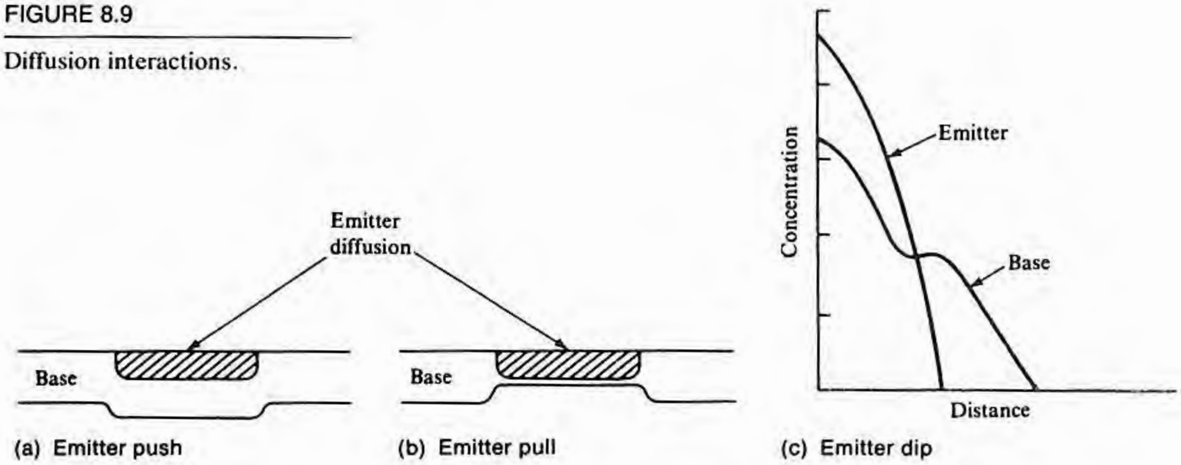
Diffusion Coefficients

	Silicon			*Gallium Arsenide*	
	D_0 (cm^2/s)	*E (eV)*		*D_0 (cm^2/s)*	*E (eV)*
Donors			*Donors*		
Phosphorus			Sulfur	0.0185	2.6
V^0	3.85	3.66	Selenium	3000	4.16
V^-	4.44	4.0	Tellurium		
$V^=$	44.2	4.37			
Arsenic					
V^0	0.066	3.44			
V^-	12	4.05			
Antimony					
V^0	0.214	3.65			
V^-	15	4.08			
Acceptors			*Acceptors*		
Boron			Beryllium	7.3×10^{-6}	1.2
V^0	0.037	3.46			
V^+	0.41	3.46			
Aluminum			Magnesium	0.026	
V^0	1.385	3.41			
V^+	2480	4.20	Zinc		
Gallium					
V^0	0.374	3.39	Cadmium		
V^+	28.5	3.92	Mercury		
Self-Interstitial					
			Gallium	0.1	3.2
			Arsenic	0.7	5.6

Source: From references 28 and 29.

FIGURE 8.9

Diffusion interactions.



(a) Emitter push (b) Emitter pull (c) Emitter dip

quency directly (32). Strain can arise from atomic misfit of diffused or implanted atoms (see section 8.7) or from layers such as SiO_2 or silicon nitride deposited on the surface. In particular, the stress around windows cut in oxide and nitride produces noticeable changes in D values (32, 33).

8.2.10 Emitter Push, Pull, and Dip

Simple theory predicts that each impurity will diffuse quite independently of others that may be present. There are, however, at least three phenomena—emitter push, pull, and dip—where this is not true (10, 34). These interactions are shown in Fig. 8.9. The earliest one reported (in 1959) was emitter push, which caused a gallium base to move deeper under a diffused phosphorus emitter (35). The effect was also present with boron base diffusions and indeed, in some cases, made it difficult to obtain the desired narrow bipolar base widths. Sometime later, when gallium bases and arsenic emitters were used, the reverse situation, emitter pull, was reported. Emitter push and pull can be explained in terms of the concentration of vacancies associated with the emitter diffusion (3, 31). The other interactive effect shown in Fig. 8.9 is a dip in the concentration of base impurity that is sometimes seen at or near the intersection of the emitter profile with the base profile. In this case, internal field retardation appears to cause the dip (34, 36).

8.3 SOLUTIONS TO THE DIFFUSION EQUATIONS

Impurity diffusion mathematics parallels that for carriers, except that the diffusion coefficient has more causes of variability in the region of interest. Combining Fick's first law with the continuity equation and assuming that D is independent of concentration give Fick's second law:

$$\frac{\partial N}{\partial t} = D\frac{\partial^2 N}{\partial x^2} \tag{8.26}$$

where t is the diffusion time. In three dimensions,

$$\frac{\partial N}{\partial t} = D\nabla^2 N \tag{8.27}$$

When D is a function of the concentration N, Eq. 8.26 becomes

$$\frac{\partial N}{\partial t} = \frac{\partial(D\partial N/\partial x)}{\partial x} = \frac{\partial D}{\partial x}\frac{\partial N}{\partial x} + D\frac{\partial^2 N}{\partial x^2} \tag{8.28}$$

In regions where the diffusion coefficients are concentration dependent, closed-form analytical solutions for the boundary conditions of interest are usually not available. In those cases, numerical integration and computer modeling are used to obtain solutions. Large programs are available that are designed to run on mainframe computers—for example, SUPREM, the Stanford University process engineering model for diffusion and oxidation. Nevertheless, it is still very instructive to first look at solutions of Eq. 8.26 for various boundary conditions applicable to semiconductor processing. To a first approximation, such solutions are valid. However, of more importance, the use of the simple closed-form solutions in studying a process does not screen the engineer from the physical processes as the massive computer modeling programs do.

Eq. 8.26 may be solved by separation of the variables (37)—that is, by considering that

$$N(x,t) = X(x)\Upsilon(t) \tag{8.29}$$

where X is a function only of x and Υ is a function only of the time t. The general solution to Eq. 8.29 is

$$N(x,t) = \int_0^\infty [A(\lambda)\cos\lambda x + B(\lambda)\sin\lambda x]e^{-\lambda^2 Dt}d\lambda \tag{8.30}$$

Solutions with boundary conditions appropriate for semiconductor processing are given in the following sections. Where multiple impurities are present, it is assumed that each impurity species diffuses independently of all others. Thus, separate, independent solutions to the diffusion equation can be obtained for each impurity. The total impurity concentration $N'(x,t)$ can be determined by summing the individual distributions. The net impurity concentration is given by $|\Sigma N_A - \Sigma N_D|$.

8.3.1 Diffusion from Infinite Source on Surface

This condition is one of the more common conditions used and gives the profiles that were shown in Figs. 8.2 and 8.3. The impurity concentration at the surface is set by forming a layer of a doping source on the surface. The layer either is thick enough initially or else is continually replenished so that the concentration N_0 is maintained at the solid solubility limit of the impurity in the semiconductor during the entire diffusion time. Assuming that the semiconductor is infinitely thick, the solution is as follows (37):

$$N(x,t) = N_0[1 - \left(\frac{2}{\sqrt{\pi}}\right) \int_0^z e^{-z^2} dz \qquad 8.31$$

where $z = x/2\sqrt{Dt}$. The integral is a converging infinite series referred to as the error function, or erf(z), and occurs in the solution of many diffusion problems. Toward the end of this chapter, some properties of the error function and a short table of values are given (see Table 8.12). Using the erf abbreviation, Eq. 8.31 can be written as

$$N(x,t) = N_0\left[1 - \text{erf}\left(\frac{x}{2\sqrt{Dt}}\right)\right] \qquad 8.32$$

The expression $1 - \text{erf}(z)$ is often referred to as the complementary error function erfc(z). If there is a background concentration N_1 of the same species at the beginning of diffusion, Eq. 8.32 becomes

$$N(x,t) = N_1 + (N_0 - N_1)\left[1 - \text{erf}\left(\frac{x}{2\sqrt{Dt}}\right)\right] \qquad 8.33$$

If there is a background concentration N_1 of a different species, then they act independently, and

$$N(x,t) = N_1 + N_0\left[1 - \text{erf}\left(\frac{x}{2\sqrt{Dt}}\right)\right] \qquad 8.34$$

If the species are of opposite type, a junction will occur when

$$N_0\left[1 - \text{erf}\left(\frac{x}{2\sqrt{Dt}}\right)\right] = N_1 \qquad 8.35$$

The value of x where this occurs is given by

$$x_j = 2\sqrt{Dt}\,\text{erf}^{-1}\left(1 - \frac{N_1}{N_0}\right) \qquad 8.36$$

which can be rewritten as

$$x_j = A\sqrt{t} \qquad 8.37$$

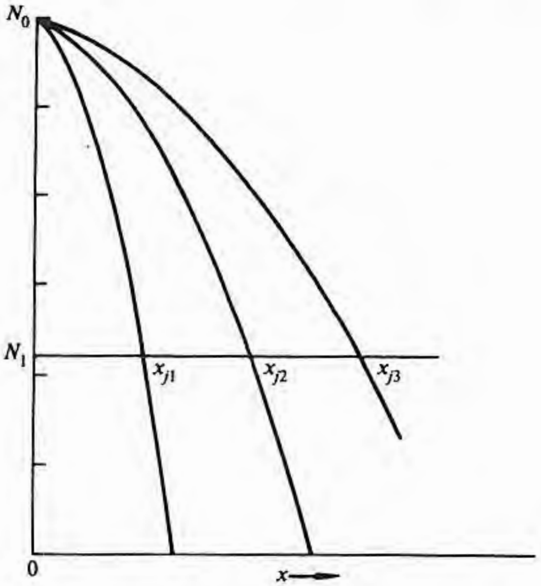


FIGURE 8.10

Position of junction after diffusion times in ratio of 1:4:9. (x_j is linear with $\sqrt{t}$ regardless of N_0/N_1 ratio.)

where A is a constant given by $A = 2[\text{erf}^{-1}(1 - N_1/N_0)]\sqrt{D}$. Thus, the junction depth x_j increases as the square root of the diffusion time, as illustrated in Fig. 8.10 where the diffusion times are in the ratio of 1:4:9. A linear plot of x_j versus $\sqrt{t}$ provides assurance that D is remaining constant over the range of concentrations covered. Then, a best-fit curve of several values of x_j and t can be used to determine a particular x_j and t that can be substituted into Eq. 8.36 to determine D.

EXAMPLE □ If N_0 is 5×10^{19} atoms/cc and a diffusion is made into an opposite-type background of 10^{16} atoms/cc, how long will it take to diffuse to the point where the junction is 1 μm deep if the temperature is such that $D = 10^{-14}$ cm²/s?

Substituting into Eq. 8.36 (cm and s are a consistent set of units) gives

$$10^{-4}\text{ cm} = (2 \times 10^{-7}\text{ cm}/\sqrt{\text{s}}) \times \sqrt{t}\,\text{erf}^{-1}(1 - 10^{16}/5 \times 10^{19})5 \times 10^{2}\sqrt{\text{s}} = \sqrt{t}\,\text{erf}^{-1}(0.998)$$

From Table 8.12, erf(0.998) ≅ 2.18. Thus, t = 52,900 s, or 14.7 hours. □

Because N_0 is set by the solubility limit of the dopant being used, at a given temperature, only one N_0 is available per dopant per semiconductor. The use of a diffusion with the concentration set by solid solubility gives good control, but the surface concentration N_0 may be much higher than desired. To provide greater flexibility in the choice of N_0, a two-step diffusion process comprised of a short diffusion from an infinite source followed by its removal and a continued diffusion from the thin layer of impurities already diffused in is often used. The total number of impurities S available is set by the first diffusion and is given by

$$\begin{aligned} S &= \int_0^\infty N(x)dx \\ &= \int_0^\infty N_0\left[1 - \operatorname{erf}\left(\frac{x}{2\sqrt{Dt}}\right)\right]dx = \frac{2N_0}{\sqrt{\pi}}\sqrt{Dt} \end{aligned} \qquad 8.38$$

The behavior of the second diffusion is described in the next section.

8.3.2 Diffusion from Limited Source on Surface

At $t = 0$, a fixed number S/cm^2 of impurities is on the surface. N is given by (37)

$$N(x,t) = \frac{S}{\sqrt{\pi Dt}}\, e^{-x^2/4Dt} \qquad 8.39$$

In this case, the distribution is Gaussian[10] and not erf. The surface concentration $N(0,t)$ continually diminishes with time as shown in Fig. 8.11 and is given by

$$N(0,t) = \frac{S}{\sqrt{\pi Dt}} \qquad 8.40$$

Thus, in the limited-source case, the surface concentration decreases linearly with $\sqrt{t}$. The junction depth, unlike that of the previous case, varies in a more complex manner:

$$e^{-x_j^2/4Dt} = \frac{N_B}{S}\sqrt{\pi Dt} \qquad 8.41a$$

or

$$x_j = \left[4Dt \log\left(\frac{S}{N_B\sqrt{\pi Dt}}\right)\right]^{1/2} \qquad 8.41b$$

Eq. 8.39 was derived based on S being located at $x = 0$, but if S actually is in a layer from a previous short diffusion of the type described by Eq. 8.32, and if the Dt product of the second diffusion

[10] Functions of the form $y = e^{-x^2/a^2}$ are referred to as Gaussian.

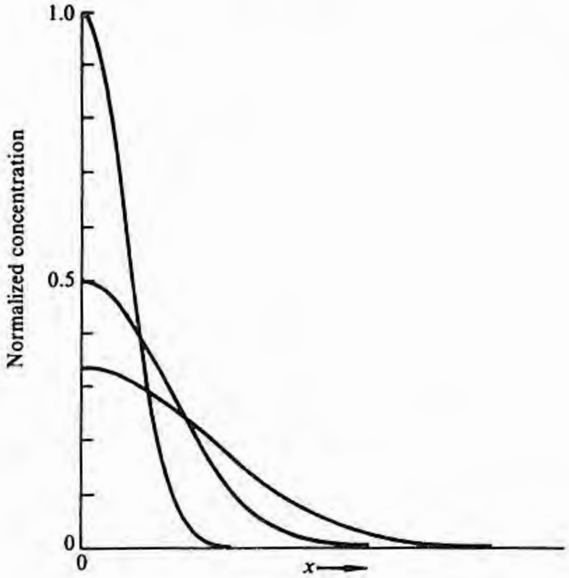


FIGURE 8.11

Diffusion from limited source plotted on linear scale for diffusion times in ratio of 1:4:9. (Note that S decreases as square root of time.)

is as much as four or five times that of the first one, the impurities from the first can still be considered as all lying at $x = 0$. S for the second diffusion will be the amount of impurities introduced during the first diffusion. Substituting the value for S from Eq. 8.38 into Eq. 8.39 gives

$$N(x',t,t') = \left(\frac{2N_0}{\pi}\right)\left(\frac{Dt}{D't'}\right)^{1/2} e^{-x'^2/4D't'} \qquad 8.42$$

and the surface concentration is

$$N_0(t,t') = \frac{2N_0}{\pi}\sqrt{\frac{Dt}{D't'}} \qquad 8.43$$

where the prime values indicate values for the second diffusion. When the second Dt product is not large compared to the first, as is likely during short, low-temperature diffusions, the exact solution instead of the approximations of Eqs. 8.42 and 8.43 can be used (38, 39):

$$N(x,t,t') = N_0\frac{2}{\sqrt{\pi}}\int_{\sqrt{\beta}}^{\infty} e^{-m^2}\,\mathrm{erf}(\alpha m)dm \qquad 8.44$$

where

$$\alpha = \sqrt{\frac{Dt}{D't'}}$$

$$\beta = \frac{x^2}{4(Dt + D't')}$$

(Representative values for the integral are tabulated in Table 8.13.) The surface concentration is given by

$$N_{\text{sur}} = N_0 \frac{2}{\pi} \tan^{-1} \sqrt{\frac{Dt}{D't'}} \qquad 8.45$$

Since $\tan^{-1} x = x - x^3/3 + x^5/5 - \ldots$, the difference between the exact solution for N_0' and that given by Eq. 8.43 is only a few percent when $D't' > 5Dt$.

Ion implantation is an alternative to the first diffusion. The ions implanted can be counted very accurately, and they are quite close to the surface (see Chapter 9). They will not have the atoms all located on one plane, however, and the boundary conditions are much more like those discussed in the next section.

8.3.3 Diffusion from Interior Limited Source

This condition is like the case just described, except that the sheet of impurity atoms is located at $x = X_0$ where X_0 is in the interior of the semiconductor so that the source is depleted by diffusion in both directions. In this case (37), assuming that X_0 is far removed from a free surface,

$$N(x,t) = \frac{S}{2\sqrt{\pi Dt}} e^{-(x - X_0)^2/4Dt} \qquad 8.46$$

Fig. 8.12 shows two profiles with their respective diffusion times differing by a factor of 4. This exponential profile has the same shape as the Gaussian or normal distribution often encountered in probability theory. For that application, it is written as

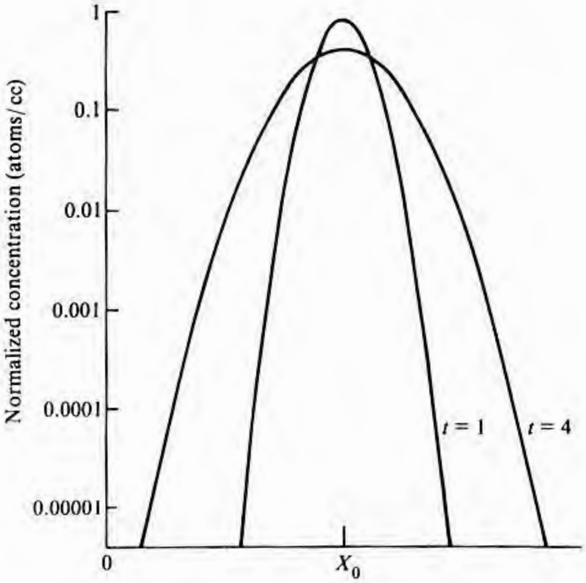
$$\frac{N_i}{\sigma\sqrt{2\pi}} e^{-(x - X_0)^2/2\sigma^2} \qquad 8.47$$

where N_i is the number of observations and σ is the standard deviation.[11] A comparison of Eqs. 8.46 and 8.47 shows that N_i

[11] 68.3% of the area under the curve is contained in the portion of the curve between $x - X_0 = 1\sigma$ and $x - X_0 = -1\sigma$.

FIGURE 8.12

Profile for diffusion from source S of width = 0 located at $x = X_0$.



corresponds to S and $\sigma = \sqrt{2Dt}$. This correlation is useful when considering subsequent diffusions after an ion implant since the initial ion implant profile is often described by a total flux ϕ (S), a range R_p (X_0), and a standard deviation ΔR_p (σ or $\sqrt{2Dt}$).

EXAMPLE □ Suppose that an ion implant is characterized by a total dose ϕ of 10^{13} atoms/cm^2, a range of 0.2 μm, and a standard deviation of 0.1 μm. What will the profile look like after a 20 minute heat cycle at a temperature where $D = 10^{-14}$ cm^2/s? Neglect the fact that because of the shallowness of the implant, there might be a surface boundary effect.

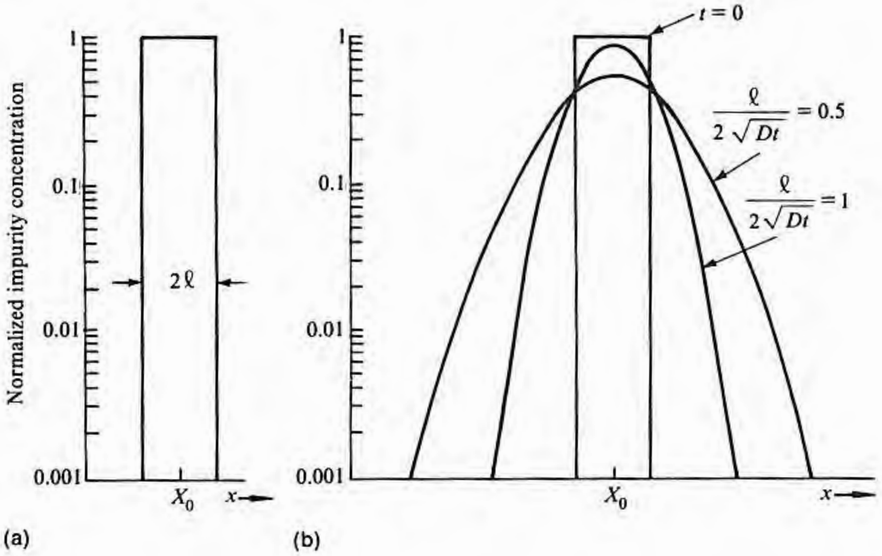
ϕ is the total number of impurities and hence corresponds to S. The scatter during implant is equivalent to a first diffusion from an initial sheet source. $\Delta R_p = \sigma = 0.1\ \mu\text{m} = \sqrt{2Dt} = 10^{-5}$ cm. Solving for Dt gives 5×10^{-11} cm^2. The subsequent diffusion has a Dt product of 10^{-14} cm^2/s × 1800 s = 1.8×10^{-11}cm^2, which is much less than the equivalent Dt of the implant. Hence, it would be expected to change the profile very little. However, as will be discussed in section 8.3.13, in this case the final distribution can be calculated by substituting a $(Dt)_{\text{eff}} = \Sigma D_1t_1 + D_2t_2 = 5 \times 10^{-11} + 1.8 \times 10^{-11} = 6.8 \times 10^{-11}$ cm^2 for the Dt of Eq. 8.46. □

8.3.4 Diffusion from Layer of Finite Thickness

This condition is a case similar to the one just described, except that initially a rectangular rather than a sheet source distribution is located internally to the body of the semiconductor. As is shown in Fig. 8.13a, if the initial thickness 2ℓ is much greater than $\sqrt{Dt}$, it is

FIGURE 8.13

Profile for diffusion from interior layer of width = 2ℓ located at $x = X_0$.



best treated as the two separate step distributions to be described in the next section. Of more interest is the case where $\sqrt{Dt}$ is greater than ℓ, where the peak distribution N_0 decreases with increasing diffusion time.

Application for this set of boundary conditions can occur if the dopant redistribution of multiple layers of molecular beam epitaxy after heat treatments is being considered. It is less applicable to conventional epitaxial processing since autodoping will usually overshadow the diffusion.

To simplify the algebra, consider that a new origin is chosen with its 0 at the X_0 of Fig. 8.13a. If the width of the doped layer is 2ℓ, $N(x,t)$ is given by (37)

$$N(x,t) = \frac{N_0}{2}\left[\operatorname{erf}\left(\frac{\ell + x}{2\sqrt{Dt}}\right) + \operatorname{erf}\left(\frac{\ell - x}{2\sqrt{Dt}}\right)\right] \qquad 8.48$$

where 2ℓ is the width of the layer. As diffusion progresses, impurities will diffuse out in both directions as shown in Fig. 8.13b. The peak concentration N_0 will decrease as $\operatorname{erf}(\ell/2\sqrt{Dt})$ and thus will change imperceptibly until $\ell/2\sqrt{Dt}$ becomes less than about 2.

8.3.5 Diffusion from Concentration Step

The initial boundary conditions for this case, illustrated in Fig. 8.14 as the abrupt steps of $t = 0$, approximate the conditions just after a conventional epitaxial deposition and very closely match those after molecular beam epitaxy (MBE). If there are a series of thin MBE layers, the previous case may be more appropriate. This case also provides an approximation of a grown junction crystal distribution. However, grown junction technology has now been obsolete for 20 years. Diffusion from each side of the step can be considered to proceed independently as described by Eqs. 8.49 and 8.50 (37):

$$N(x,t) = \frac{N_1}{2}\left[1 - \operatorname{erf}\left(\frac{x}{2\sqrt{D_1 t}}\right)\right] \tag{8.49}$$

$$N(x,t) = \frac{N_2}{2}\left[1 + \operatorname{erf}\left(\frac{x}{2\sqrt{D_2 t}}\right)\right] \tag{8.50}$$

N_1 is the doping level for $-x$ and $t = 0$; N_2, for $+x$ and $t = 0$. D_1 is the diffusion coefficient for the N_1 species; D_2, for the N_2 species. Eqs. 8.49 and 8.50 may be added together to give a single expression covering diffusion from both sides of the step:

$$N(x,t) = \frac{N_1}{2}\left[1 - \operatorname{erf}\left(\frac{x}{2\sqrt{Dt}}\right)\right] + \frac{N_2}{2}\left[1 + \operatorname{erf}\left(\frac{x}{2\sqrt{Dt}}\right)\right] \tag{8.51}$$

FIGURE 8.14

Profile for diffusion occurring at a concentration step. (In this example, N_1 and N_2 have the same Dt product.)

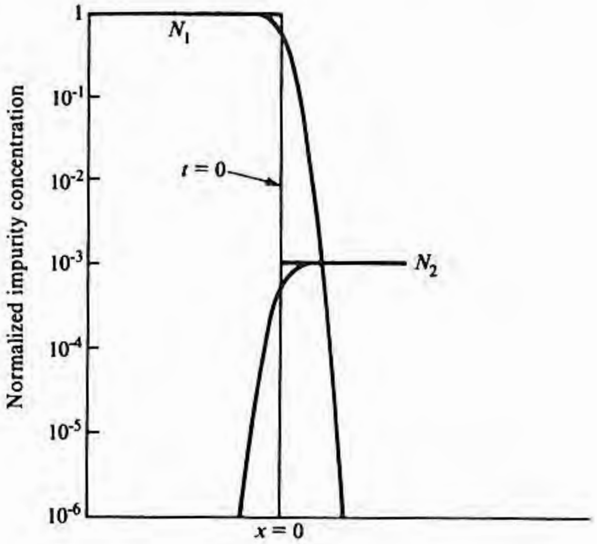


FIGURE 8.15

Dip in net concentration occurring near original high/low boundary because of a more rapid diffusant in the high-resistivity layer.

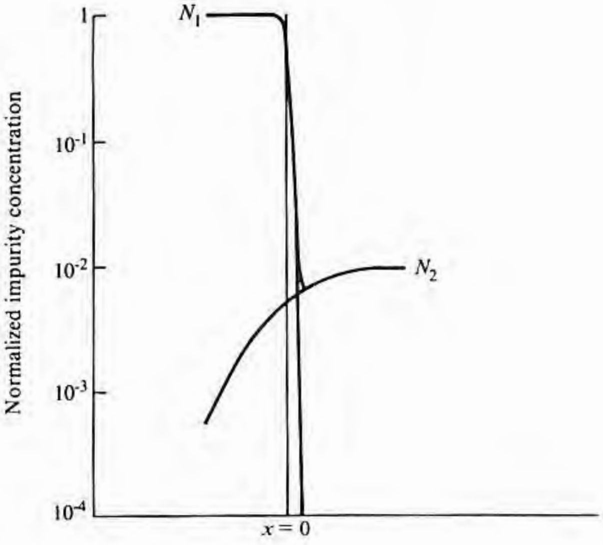


Fig. 8.14, along with the initial conditions, also shows a profile for each separate diffusion (Eqs. 8.49 and 8.50) and their sum (Eq. 8.51) for the case of the same impurity on each side of the step ($D_1t = D_2t$). When the two Dt's are not equal, substantially different distributions can sometimes occur. For example, if $N_1 >> N_2$ and $D_2 > D_1$, it is possible to get a dip in the concentration curve as shown in Fig. 8.15 (40). This dip might occur, for example, if an epi substrate were doped with antimony and overgrown with an arsenic layer. The maximum possible dip is a factor of 2 and occurs before the N_1 diffusion has perceptibly moved into the N_2 region. If a heavily doped n-substrate partially compensated by having some spurious p-dopant present is overgrown with a high-resistivity n-layer, then if $N_2 < N_{1p}$ and if $D_{1n} << D_{2p}$, a p-layer could occur in the epitaxial layer as shown in Fig. 8.16 (41). This happens if low-resistivity antimony substrates become contaminated with boron. The resulting p-layer is often referred to as a "phantom" p-layer, although it is indeed real. If N_1 and N_2 are of opposite type, the junction will move into the more lightly doped region as diffusion proceeds, regardless of the relative values of D_1 and D_2.

8.3.6 Diffusion from Concentration Step into Moving Layer

This case, as shown by Fig. 8.17, describes diffusion from a heavily doped substrate into an epitaxial layer being grown at a velocity v in cm/s. However, this case is ordinarily of little importance both because autodoping (see Chapter 7) will usually overshadow diffusion and because growth is generally much more rapid than diffu-

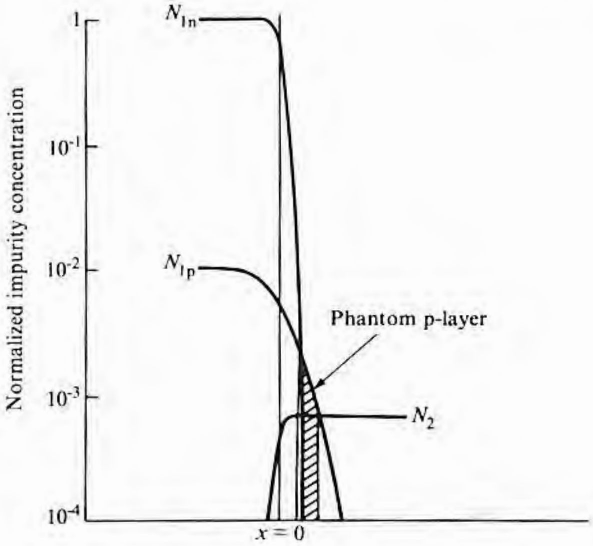


FIGURE 8.16

Origin of phantom p-layer that occurs in n-epi layers because of n-substrate contamination by a faster diffusing p-dopant.

sion. In the event that the epitaxial growth temperature is reduced to the point where autodoping is negligible, diffusion will be reduced to the point that the solution for diffusion into an infinite thickness (Eq. 8.50) can be used. Even at higher temperatures, the use of Eq. 8.50 causes little error in most cases.

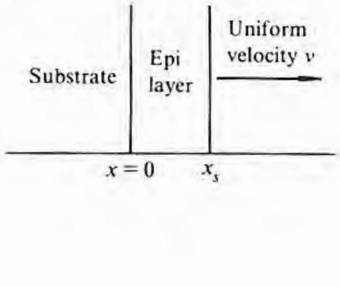
To account for the moving boundary, instead of Eq. 8.26, the equation to be solved is as follows (42, 43):

$$D\frac{\partial^2 N}{\partial x^2} = \frac{\partial N}{\partial t} + v\frac{\partial N}{\partial x} \qquad 8.52$$

where v is the epitaxial growth velocity. In the solutions to follow, v is assumed to remain constant. One solution represents diffusion from an infinitely thick substrate into the growing layer and is generally the one of interest. In the event that the "substrate" is a thin epitaxial layer just grown or a thin high-concentration layer diffused into a lightly doped wafer, then it is not infinite in extent, and its concentration will decrease with growth (diffusion) time. The other represents diffusion from the growing layer into the substrate. For the case of out-diffusion from an infinite substrate, the boundary conditions are as follows:

FIGURE 8.17

Geometry for diffusion during epitaxial growth.



$$N(-x,0) = N_1$$
$$N(-\infty,t) = N_1$$
$$J = -D\frac{\partial N}{\partial x} = (K + v)N \quad \text{at } x_s$$

where K is a rate constant describing the loss of dopant at the epi–ambient interface and x_s is the location of the interface. The solution is given by

$$\begin{aligned}\frac{N(x,t)}{N_1} &= \frac{1}{2}\operatorname{erfc}\left(\frac{x}{2\sqrt{Dt}}\right)\\ &\quad - \left(\frac{K+v}{2K}\right)e^{(v/D)(vt-x)} \times \operatorname{erfc}\left(\frac{2vt-x}{2\sqrt{Dt}}\right)\\ &\quad + \left(\frac{2K+v}{2K}\right)e^{[(K+v)/D][(K+v)t-x]}\\ &\quad \times \operatorname{erfc}\left(\frac{2(K+v)t-x}{2\sqrt{Dt}}\right)\end{aligned} \qquad 8.53$$

When the rate constant K approaches infinity (no impediment at the surface), Eq. 8.53 will predict a smaller N for a given x than will Eq. 8.50. When K goes to zero (no loss at the surface), Eq. 8.53 will predict a higher value for N at a given x than will Eq. 8.50. For v^2t/D greater than about 10, Eqs. 8.53 and 8.50 will very closely match, regardless of the K value. The reason for the insensitivity to K in this case is that the layer appears infinitely thick and no impurity gets to the boundary. Little data exist on K values, but data in reference 43 suggest values in the range of 2×10^{-7}cm/s. In the event that the substrate is not infinitely thick—for example, for one epitaxial layer deposited on top of another or on top of a diffused layer—then the solution is more related to diffusion from the finite thickness sources discussed earlier. (See reference 42 for further details.)

EXAMPLE □ For the case of a 1200°C silicon epitaxial deposition, calculate the v^2t/D values for a layer 0.05 μm thick if the deposition rate is 0.5 μm/minute and the substrate is doped with antimony.

The D value for antimony, found from Fig. 8.26, which appears in a later section, is about 10^{-13} cm^2/s. The time t to grow 0.05 μm at 0.5 μm/minute is 0.1 minute, or 6 s. The rate of 0.5 μm per minute equals 8.3×10^{-7} cm/s. Thus, $v^2t/D \cong 42$. As the film gets thicker, the expression gets larger. It also gets larger as v increases and as D decreases. D, as will be seen later, decreases rapidly as the temperature decreases. Typical silicon epitaxial deposition conditions are 1100°C at 1 μm/minute where v^2t/D will be much larger than 42. Thus, in most epitaxial depositions, $v^2t/D >> 10$, and Eq. 8.50 is applicable for thicknesses greater than a very small fraction of a micron. □

The other half of the diffusion, that from the moving layer into the initial material, is considerably simpler since there is no out-diffusion problem. When the initial material is infinitely thick, the boundary conditions are as follows:

$$N(-x,0) = 0$$
$$N(x_s,t) = N_2$$
$$N(-\infty,t) = 0$$

and when applied to Eq. 8.52 give

$$N(x,t) = \frac{N_2}{2}\left[1 + \operatorname{erf}\left(\frac{x}{2\sqrt{Dt}}\right) + e^{(v/D)(vt-x)}\operatorname{erfc}\left(\frac{2vt - x}{2\sqrt{Dt}}\right)\right] \qquad 8.54$$

When v^2t/D is large, this equation reduces to the nonmoving boundary solution of Eq. 8.50.

8.3.7 Out-Diffusion with Rate Limiting at Surface

This set of conditions would describe the motion of impurities in an initially uniformly doped semiconductor heated to a high temperature. The impurity flux J_s across the semiconductor–ambient interface at $x = 0$ is given by

$$J_s = K(N_e - N_s)$$

where K is the same rate constant discussed in the preceding case, N_s is the concentration of impurity in the semiconductor at the surface, and N_e is the equilibrium concentration of impurity in the ambient adjacent to the semiconductor. Note that if $N_e > N_s$, there will be in-diffusion rather than a loss of impurities by out-diffusion. If, as is common, N_e is assumed to be much less than N_s, then $J_s = -KN_s$, and, by setting the growth velocity v equal to zero, Eq. 8.53 can be used to calculate the impurity profile following out-diffusion (43). Thus,

$$\frac{N(x,t)}{N_1} = \operatorname{erf}\left(\frac{-x}{2\sqrt{Dt}}\right) + e^{(K/D)(Kt-x)}\operatorname{erfc}\left(\frac{2Kt - x}{2\sqrt{Dt}}\right) \qquad 8.55$$

For the limit of $K = 0$, nothing will escape; for the other limit of K very large so that there is no surface barrier, Eq. 8.55 reduces to $N(x,t) = N_1\operatorname{erf}(-x/2\sqrt{Dt})$. The solution for the case of $N_e \neq 0$ is available (44) but is more complex and not included. Graphical aids to its use are found in reference 45.

8.3.8 Diffusion from a Fixed Concentration into Moving Layer

These conditions could arise from a simultaneous diffusion and surface etching or evaporation. During some silicon diffusions, the initial diffusion source is a mixed oxide layer formed on the surface.

When the diffusion is done in an oxidizing atmosphere, as is normal, oxide will grow at the glass–silicon interface. However, the mixed oxide layer dissolves it as it forms so that the source of fixed concentration N_0 stays in contact with the moving wafer surface.

For the conditions of the surface moving in the $+x$ direction with a velocity v,

$$\begin{aligned} N(vt,t) &= N_0 && \text{for all } t\text{'s} \\ N(x,0) &= 0 && \text{for } x > 0 \end{aligned}$$

By changing to a new variable $x' = x - vt$ (x' is measured from the interface and not from the original position), Eq. 8.26 becomes

$$D\frac{\partial^2 N}{\partial x'^2} + v\frac{\partial N}{\partial x} = \frac{\partial N}{\partial t} \qquad 8.56$$

The solution is as follows (46):

$$\begin{aligned} N(x - vt,t) &\equiv N(x',t) \\ &= \frac{N_0}{2}\left[\left(\operatorname{erfc}\frac{x' + vt}{2\sqrt{Dt}}\right) + e^{-vx'/D}\operatorname{erfc}\frac{x' - vt}{2\sqrt{Dt}}\right] \end{aligned} \qquad 8.57$$

where x is measured from the original surface and x' from the actual boundary at a time t.

8.3.9 Diffusion through Thin Layer

One way in which these conditions could occur would be if diffusion were through a layer of polysilicon on top of single-crystal silicon or through a layer of a mixed III–V compound on top of gallium arsenide. Another configuration is an oxide or other masking layer over silicon or gallium arsenide. Since the solutions to be given here do not consider a moving boundary, they are an approximation to diffusion through a thermal oxide growing on silicon. Fig. 8.18 shows the two circumstances considered. In Fig. 8.18a, there is no segregation coefficient between the two media, and thus the concentration is continuous across the boundary. This situation is applicable to the polysilicon/Si and mixed III–V/GaAs cases. For an oxide layer, the concentration is usually not continuous, as is shown in Fig. 8.18b. The equations for case (a) are as follows (47):

$$D_1\frac{\partial^2 N_1}{\partial x^2} = \frac{\partial N_1}{\partial t} \qquad \text{for} - a < x < 0 \qquad 8.58a$$

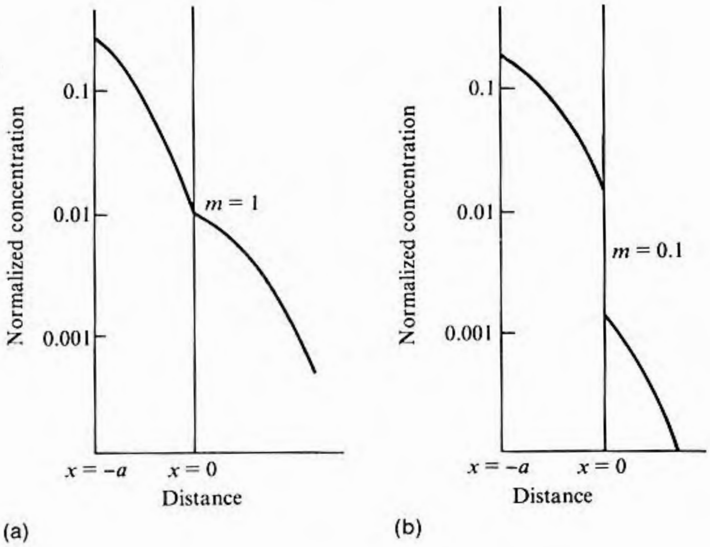
$$D_2\frac{\partial^2 N_2}{\partial x^2} = \frac{\partial N_2}{\partial t} \qquad \text{for } x > 0 \qquad 8.58b$$

The conditions are

$$J_1 = J_2 \qquad \text{at } x = 0$$

FIGURE 8.18

Diffusion through thin layer into infinitely thick layer of different material.



$$N_1(-a,t) = N_0 \qquad \text{where } a = \text{thickness of thin layer}$$

$$N_1(0,t) = N_2(0,t)$$

$$N_2(x,t) \rightarrow 0 \text{ as } x \rightarrow \infty$$

The solutions are

$$N_1(x,t) = N_0 \sum_{j=0}^{\infty} \left(\frac{1 - \mu}{1 + \mu}\right)^j \times \left[\operatorname{erfc}\frac{a(2j + 1) + x}{2\sqrt{D_1 t}} - \frac{1 - \mu}{1 + \mu}\operatorname{erfc}\frac{a(2j + 1) - x}{2\sqrt{D_1 t}}\right] \qquad 8.59$$

$$N_2(x,t) = \frac{2\mu N_0}{1 + \mu} \sum_{j=0}^{\infty} \left(\frac{1 - \mu}{1 + \mu}\right)^j \times \operatorname{erfc}\left[\frac{a(2j + 1)}{2\sqrt{D_1 t}} + \frac{x}{2\sqrt{D_2 t}}\right] \qquad 8.60$$

where $\mu = D_1/D_2$. If $a/2\sqrt{Dt} > 1$, the first term in each series is a good approximation. That is,

$$N_1(x,t) \approx N_0\left(\operatorname{erfc}\frac{a + x}{2\sqrt{D_1 t}} - \frac{1 - \mu}{1 + \mu}\operatorname{erfc}\frac{\mathrm{a} - x}{2\sqrt{D_1 t}}\right) \qquad 8.61a$$

$$N_2(x,t) \approx \frac{2\mu N_0}{1 + \mu}\operatorname{erfc}\left(\frac{a}{2\sqrt{D_1 t}} + \frac{x}{2\sqrt{D_2 t}}\right) \qquad 8.61b$$

In the event that the concentration is not continuous from medium 1 to medium 2, an additional boundary condition is needed:

$$N_2(0,t) = mN_1(0,t)$$

Note that m is the segregation coefficient and is discussed in Chapter 3. For $m < 1$ and N_2 not limited by solid solubility, the solutions for N_1 and N_2 analogous to Eq. 8.61 become (48)

$$N_1(x,t) \approx N_0\left(\operatorname{erfc}\frac{a + x}{2\sqrt{D_1t}} - \frac{m - \mu}{m + \mu}\operatorname{erfc}\frac{a - x}{2\sqrt{D_1t}}\right) \qquad 8.62\text{a}$$

$$N_2(x,t) \approx \frac{2m\mu}{m + \mu}N_0\operatorname{erfc}\left(\frac{a}{2\sqrt{D_1t}} + \frac{x}{2\sqrt{D_2t}}\right) \qquad 8.62\text{b}$$

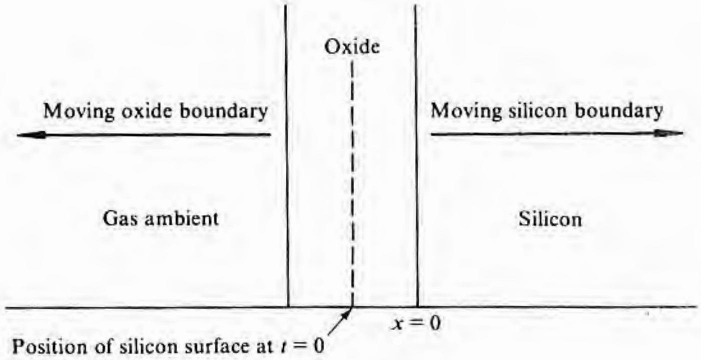
The problem of diffusion through a growing rather than a static oxide (a moving boundary problem) has also been solved (49), but the requirement for a profile under such conditions is seldom encountered. What is often needed is the amount of dopant diffusing through a simultaneously growing thermal oxide mask in order to see whether it will cause a surface problem. The likelihood of a problem can be determined by calculating $N_2(0)$ for any desired time from Eq. 8.62a and comparing it with the background doping. However, the source material reacting with the masking oxide may reduce its thickness by a substantial amount and thus increase the amount of dopant that diffuses through it, as is particularly true in the case of phosphorus diffusions.

8.3.10 Diffusion during Thermal Oxidation

This solution concerns the redistribution during oxidation of impurities already in a silicon wafer. Fig. 8.19 shows the geometry. During oxidation, the silicon wafer surface moves inward. Simultaneously,

FIGURE 8.19

Geometry for diffusion during silicon thermal oxidation. (Coordinates have been chosen so that $x = 0$ follows the silicon–oxide interface.)

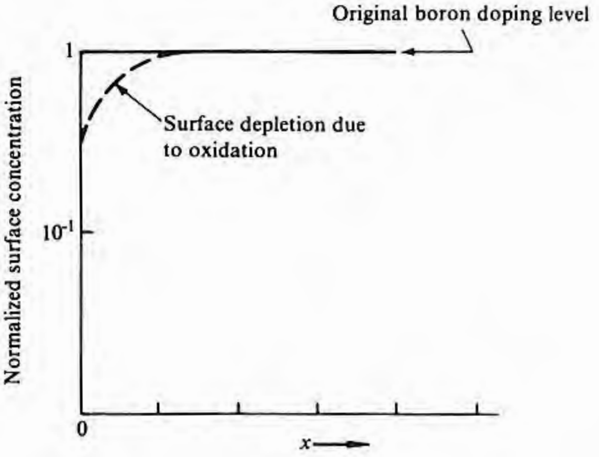


the oxide increases in thickness. The impurities originally in the silicon consumed by the oxide may be rejected and diffuse ahead of the silicon or be incorporated into the growing oxide, depending on whether the oxide–silicon segregation coefficient m is >1 or <1. (As defined earlier, m is the impurity concentration in silicon at the interface divided by the concentration in the oxide at the interface.) Solutions to the various boundary conditions that can arise are very complex (50), and most have been solved by numerical integration (51, 52).

Thus, if the silicon is initially uniformly doped, the oxidation will either pile up some of those impurities in the silicon next to the interface or else deplete the silicon and cause a concentration dip at its surface. Fig. 8.20 shows the kind of profile to be expected for the case of $m > 1$, where there is a depletion of impurity from the silicon surface. When there is a nonuniform distribution, as, for example, if an earlier diffusion or implant step had taken place, the same pileup or depletion occurs. A profile comparable to that of Fig. 8.20 is shown in Fig. 8.21 but represents a boron dopant after a predep and subsequent drive-in.[12] In the case of a very shallow predep or when an implant is used, care must be taken to ensure that the oxide layer is not grown thick enough to consume all of the impurities.

FIGURE 8.20

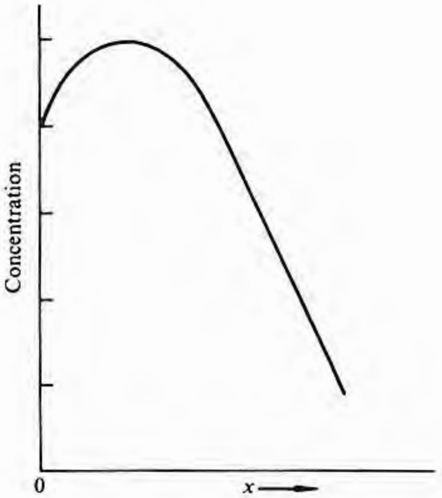
Depletion of boron concentration at surface due to thermal oxidation.



[12] A drive-in is used to intentionally lower the surface concentration after an initial predeposition diffusion. Such a lowering can occur in two ways. Since the drive-in diffusion source is limited to only those impurities introduced during the predep, continued diffusion causes the surface concentration to decrease as was shown in Fig. 8.11. In addition, in the case of boron, the growing oxide not only retains all of the boron in the silicon consumed but also acts as a sink so that boron from silicon adjacent to the oxide diffuses into it.

FIGURE 8.21

Boron diffusion profile after combined oxidation and drive-in.



8.3.11 Diffusion from Infinite Source into Finite Thickness

Most diffusions are intended to extend only a short distance into the wafer and thus the case described in section 8.3.1 is appropriate. However, in the case of fast-diffusing impurities such as the heavy metals or for long diffusion times and thin wafers, it is quite possible for a diffusion originating from one side to travel completely through the wafer. If the back of the wafer is considered to be impermeable (37),

$$N(x,t) = N_0\left[1 - \frac{4}{\pi}\left(e^{-y}\sin\frac{\pi x}{2a} + \frac{1}{3}e^{-9y}\sin\frac{3\pi x}{2a} + \ldots\right)\right] \qquad 8.63$$

where the thickness of the wafer is a and $y = \pi^2Dt/4a^2$.

Should the back of the wafer not be impermeable, as, for example, if it were coated with an oxide that acted as a sink for the diffusant, then a diffusion current would exist across the back surface given by

$$J = K[N(x = a,t) - N(+)] \qquad 8.64$$

where K is a rate constant and $N(+)$ is the concentration just outside the wafer boundary. For solutions of this case, see reference 53.

Occasionally, diffusion into a wafer from both sides is of interest. In this case, Eq. 8.63 is slightly modified:

$$N(x,t) = N_0\left[1 - \frac{4}{\pi}\left(e^{-y'}\sin\frac{\pi x}{a} + \frac{1}{3}e^{-9y'}\sin\frac{3\pi x}{a} + \ldots\right)\right] \qquad 8.65$$

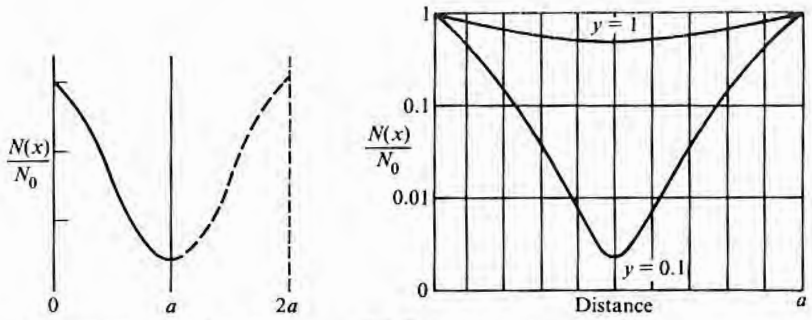
where y' is given by $y' = \pi^2 Dt/a^2$. These distributions are shown in Fig. 8.22.

8.3.12 Two-Dimensional Diffusion from High-Diffusivity Path

The idealized geometry for this case is shown in Fig. 8.23a. Solutions to this case were derived to cover diffusion along and out from metal grain boundaries and dislocations (54–56). When semiconductor processing was less developed, there was concern about enhanced diffusion along grain boundaries, twin boundaries, stacking faults, and dislocations, and the effect was studied in germanium and silicon (57–59). Grain boundaries exhibit a pronounced effect,

FIGURE 8.22

Diffusion profile when $x/2\sqrt{Dt}$ is comparable to wafer thickness.

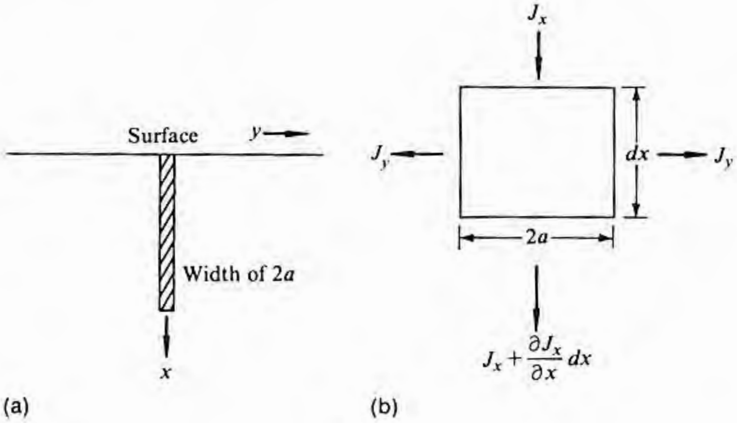


(a) From one side of wafer

(b) From both sides of wafer

FIGURE 8.23

(a) Geometry for diffusion from a high-diffusivity layer. (b) Fluxes into and out of an element of the high-diffusivity layer that are needed to develop continuity equation.



but now grain boundaries are not tolerated in the starting wafers used in IC processing, and any incoming starting slices having grain boundaries are rejected. No enhancement has been observed in either first- or second-order twins (also not tolerated in IC processing) or in stacking faults. Enhanced diffusion can apparently occur at a single dislocation, but not a high enough density of dislocations exists to materially affect overall diffusivity or the shape of the diffusion front. Enhanced diffusion occurs in polycrystalline material, and some of the newer structures using polysilicon-filled trenches may have trenches thin enough that solutions to be given here are useful. It is also observed in the defect network adjacent to the sapphire substrate in silicon-on-sapphire wafers (60).

It is assumed that the segregation at the boundary is 1 and that the surface concentration N_0 of each medium is the same. Let N and D be the concentration and diffusivity in the single crystal; N' and D', that in the high-diffusivity layer. At the boundary between the two media ($y \pm a$), $N = N'$ and $J(y) = J'(y)$. Outside the layer,

$$D\nabla^2 N = \frac{\partial N}{\partial t} \qquad 8.66$$

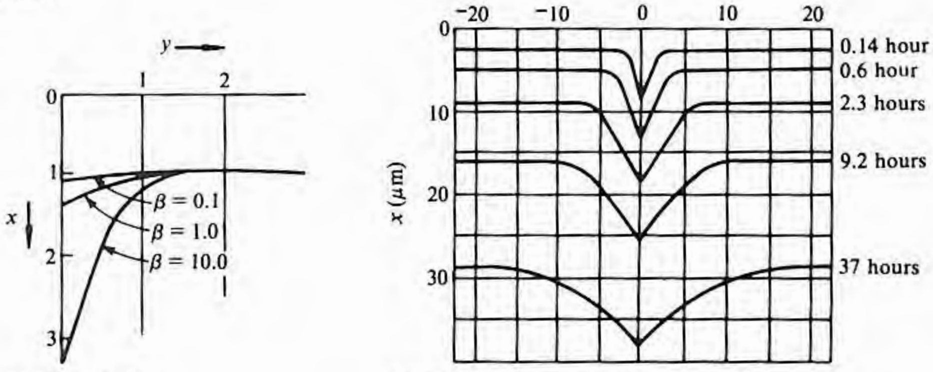
Using the fluxes for an element of the layer, as shown in Fig. 8.23b, and the continuity equation within the layer gives as an additional boundary at $y = \pm a$

$$D'\frac{\partial^2 N}{\partial x^2} - \frac{D}{a}\frac{\partial N}{\partial x} = \left(\frac{D'}{D} - 1\right)\frac{\partial N}{\partial t}$$

Eq. 8.66 has been solved analytically, but the expression is very complex and unwieldy (55). To simplify its use, tables of values for various parametric ratios have been computed (56). Calculated plots of $N(x,y)$ are shown in Fig. 8.24a for differing values of the parameter $\beta = [(D'/D) - 1]a/\sqrt{Dt}$. β can change by either changing D'/D or the diffusion time. In Fig. 8.24a, β changed because of a change in the D ratio since a change in time would have changed the depth of the diffusion front at large y. This change is shown in Fig. 8.24b, which is actual data from a silicon grain boundary diffusion (57). In this case, the time was changed but not the ratio of grain boundary to single-crystal diffusivities. The D'/D ratio for phosphorus has been experimentally measured as $\sim 10^4$–10^5 (57, 61). This data leads to the conclusion that with a close spacing of boundaries, a large increase in effective diffusivity would be expected. Depending on diffusion conditions, the interstitial sinks at closely spaced grain boundaries could also increase the overall diffusivity. Experimentally, it is observed that polycrystalline silicon does show enhanced D values. The effect is modeled in some programs by assuming styl-

FIGURE 8.24

Grain boundary isoconcentration profiles. (*Source:* (a) Adapted from R.T.P. Whipple, *Phil. Mag. Series 7, Vol. 45*, pp. 1225–1236, 1954. (b) Adapted from Van E. Wood et al., *J. Appl. Phys. 33*, pp. 3574–3579, 1962.)



(a) Calculated

(b) Measured on a silicon grain boundary

ized grains, bulk and grain boundary diffusion coefficients, and simultaneous flow through grains and along boundaries (62). Heavily doped amorphous silicon also shows enhanced diffusivity in the 500°C–600°C range (63) (at higher temperatures, the amorphous silicon recrystallizes).

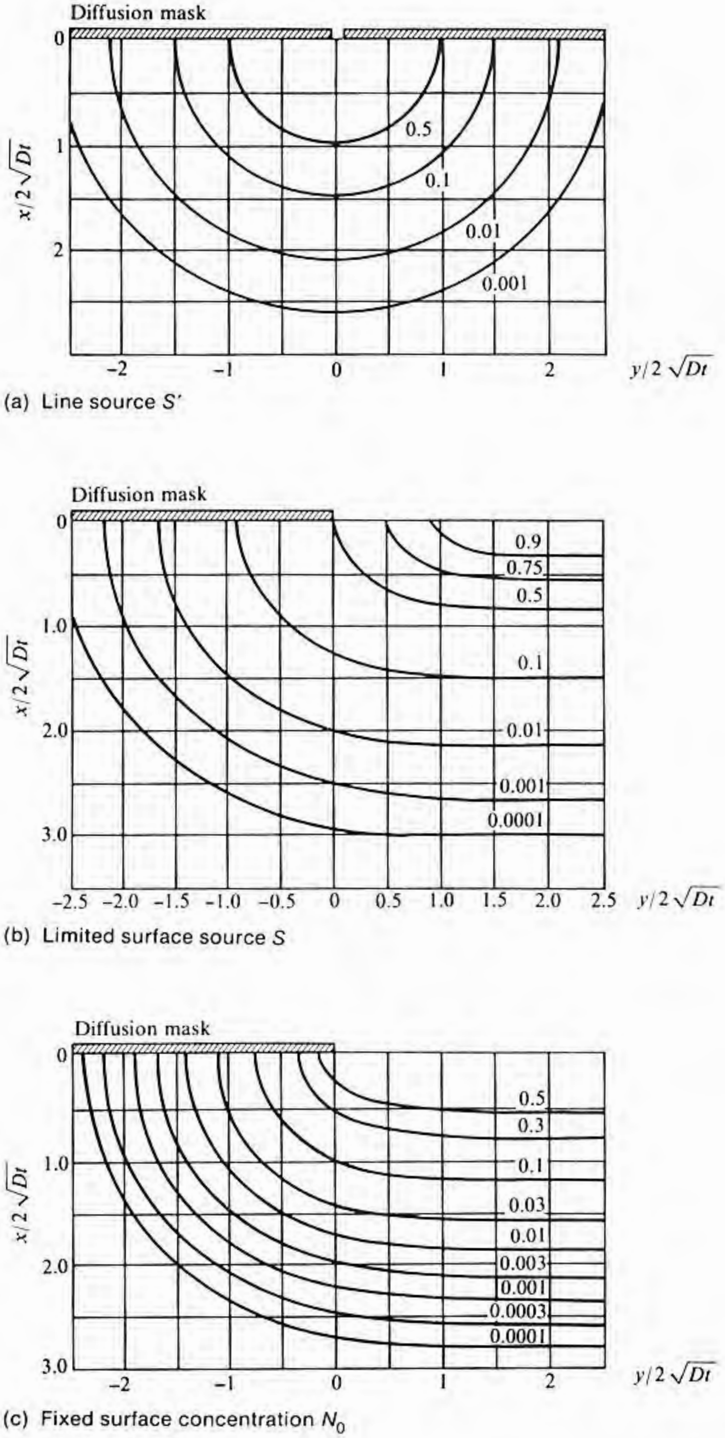
8.3.13 Two-Dimensional Solutions

Except for diffusion from the high-diffusivity layer, all of the solutions described thus far have only considered a planar diffusion front of infinite extent. In actual device fabrication, the diffusions do not extend to infinity but terminate near the edge of the mask used to define the diffusions. To calculate the behavior at the edge, two-dimensional solutions are required. The display method in this case is generally one of a series of isoconcentration lines as shown in Fig. 8.25 rather than a concentration–distance profile as used for one-dimensional cases. Fig. 8.25a is the case of diffusion from a limited source with the two mask edges infinitely close together, which gives a line source of S' atoms/cm. Fick's second law in two-dimensional spherical coordinates with axial symmetry is

$$D\frac{\partial^2 N}{\partial r^2} + \frac{D}{r}\frac{\partial N}{\partial r} = \frac{\partial N}{\partial t} \tag{8.67}$$

FIGURE 8.25

Isoconcentration contours at edge of diffusion mask. (*Source:* Curves b and c from D.P. Kennedy and R.R. O'Brien, *IBM J. Res. Develop.* 9, p. 179, 1965.)



(a) Line source S'

(b) Limited surface source S

(c) Fixed surface concentration N_0

and it has as a solution (37)

$$N(r,t) = \frac{S'}{2\pi Dt} e^{-r^2/4Dt} \qquad 8.68$$

This case is not actually encountered, and it is the only case that has radial symmetry about $r = 0$ (where diffusion along the surface equals that normal to the surface).

For the other extreme, the case of a limited surface concentration of S atoms/cm^2 and the two edges of the mask separated by an infinite distance rather than by an infinitely narrow strip, the solution near one mask edge in x,y coordinates is (64)

$$N(x,y,t) = \frac{S}{2\sqrt{\pi Dt}} e^{-x^2/4Dt} \left[1 + \text{erf}\left(\frac{y}{2\sqrt{Dt}} \right) \right] \qquad 8.69$$

where the coordinate axes are as shown in Fig. 8.25b. Note that diffusion out under the mask causes increased depletion of the surface near the mask edge and that any particular isoconcentration line extends further down beneath the open window than laterally beneath the mask. For intermediate cases where the mask opening (source) has a finite width w but the isoconcentration line of interest is at a depth much greater than w, Eq. 8.68 can be used with S' given by Sw (29).

The solution for a fixed surface concentration (not given) is considerably more complex and requires the use of hypergeometric series (64). A plot is shown in Fig. 8.25c. When the diffusion coefficient D is concentration dependent, it appears, both experimentally and from modeling (65), that the ratio of lateral to vertical motion is somewhat reduced from that predicted by Eq. 8.69.

8.3.14 Effect of Temperature Varying with Time

Often, the determination of a diffusion profile after several heat-treatment steps at different times and temperatures is necessary. Since only the Dt product occurs in the various diffusion equations, if the initial conditions do not change from one cycle to the next, then, even though a series of discrete times and temperatures are actually used, a single $(Dt)_{\text{eff}}$ is equivalent:

$$(Dt)_{\text{eff}} = \Sigma D_1 t_1 + D_2 t_2 + D_3 t_3 + \ldots \qquad 8.70$$

If the temperature varies continuously, the $\int D(T)dt$ is required. Examples of cases where this approach is applicable are diffusion from a step and diffusion from a limited source. If a diffusion were being made from an infinite source, and the surface concentration varied with temperature as is sometimes true, then Eq. 8.70 is not sufficient. However, that case can be treated by a series solution involving the various Dt's and their respective surface concentrations (66).

8.4 DIFFUSION PROFILE CALCULATIONS

In order to calculate a diffusion profile, a D value, the appropriate $N(x,t)$ function, and a time–temperature sequence are required. As was mentioned earlier, if the concentration N remains below n_i, then D for substitutional diffusion can generally be considered as independent of N. Fig. 8.7 showed n_i versus temperature for silicon, and Fig. 8.26 shows typical low concentration values of D for the common silicon substitutional dopants. These values may be substituted into the expressions given in the previous section to give profiles for the desired boundary conditions. For dopant concentrations high enough for the semiconductor to be extrinsic at the diffusion temperature, the enhancement described in section 8.2.7 is observed for most group IIIA and VA impurities. Experimental data are available for As, Sb, P, Ga, and B (3, 28, 67). Typically, they all look much like the profile shown in Fig. 8.27 for arsenic. It should be noted that while this profile has the enhancement plotted versus the concentration of arsenic atoms, some data are plotted versus the number of carriers. The effect of an enhanced diffusivity at high concentrations is to make the diffusion profile fuller at high concentrations as shown in Fig. 8.28. To account for this effect, the diffusion equations may be solved numerically, or a polynomial approximation may be used for the portion of the profile in the extrinsic region, or else an "average"[13] value of D may be used that will give the correct junction depth but an incorrect profile.

The arsenic profile in the extrinsic region can be reasonably approximated by (68)

$$\frac{N}{N_0} = 1 - 0.87Y - 0.45Y^2 \qquad 8.71$$

where $Y = (8N_0D^*t/n_i)^{-1/2}x$. While the boron profile has the same general features, it is best fitted by an expression of the form (69)

$$\frac{N}{N_0} = 1 - Y^{2/3} \qquad 8.72$$

This expression gives a very good fit to the experimental data, but no simple relation exists between Y and the diffusivity. Y is approximately given by

[13] This approach was widely used before a better understanding of the diffusion mechanisms was developed and led to D values that were dependent on background doping.

FIGURE 8.26

Diffusion coefficients for common dopants diffusing in intrinsic silicon. (*Source:* From data in Richard B. Fair, *Impurity Doping Processes in Silicon*, F.F. Wang, ed., North-Holland, New York, 1981.)

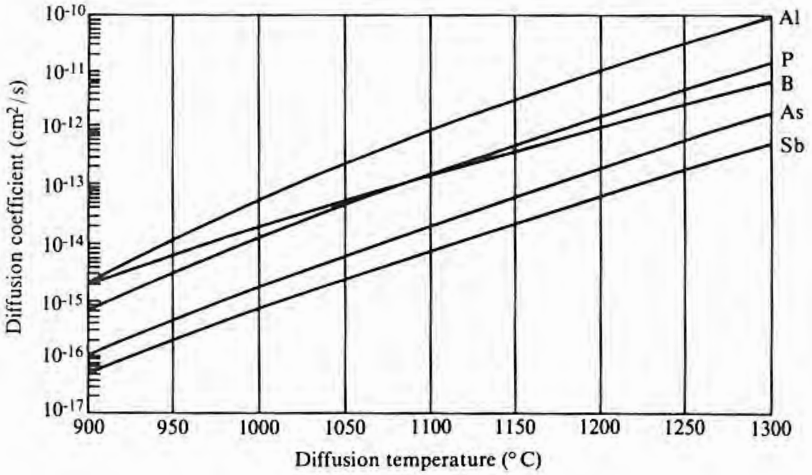
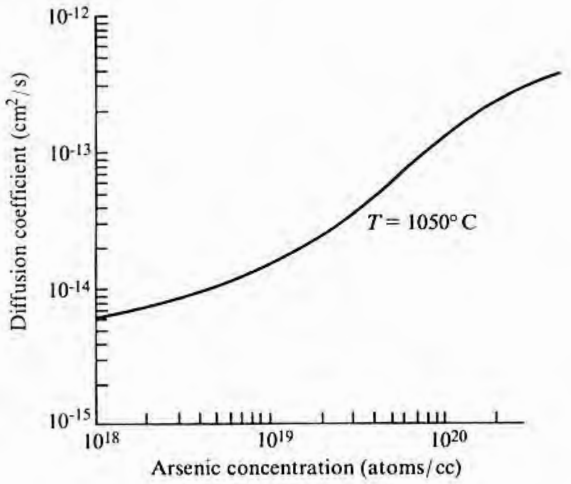


FIGURE 8.27

Arsenic diffusivity enhancement versus arsenic concentration. (*Source:* Adapted from Richard B. Fair and Joseph C.C. Tsai, *J. Electrochem. Soc.* *122*, p. 1689, 1975.)



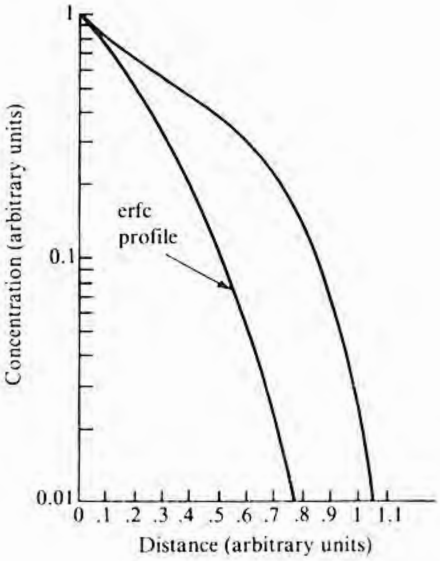
$$Y = \left(\frac{x^2}{6D_s t}\right)^{1/2} \qquad 8.73$$

where $D_s = D^* N_0/n_i$.

Diffusion in polycrystalline silicon has not been nearly as well characterized as diffusion in single-crystal material. In general, however, diffusivity is much higher in polycrystalline material. Data for arsenic show a D value with an activation energy approximately the same as that of single-crystal silicon diffusivity, but with a mag-

FIGURE 8.28

Effect of enhanced diffusivity in extrinsic region on diffusion profile.



nitude about 5 orders of magnitude greater (70). Phosphorus appears about 2 orders of magnitude higher (71); and boron, about 1 order higher (72).

8.4.1 Phosphorus Profile

The phosphorus profile (3), shown in Fig. 8.29, has peculiarities that do not allow its modeling by the polynomial approach just described. The flat top, shown by a dashed line, is sometimes found in other profiles and is really an artifact of measurement. It represents the maximum amount of electrically active phosphorus present. The solid line shows the total phosphorus content as determined by SIMS. In the region from maximum concentration down to approximately $N = 10^{20}$ atoms/cc, D is given reasonably well by D^* $(1 + N/n_i)^2$. That is, the double-charged vacancy term of Eq. 8.9 predominates. However, when the impurity concentration drops enough, the P-double vacancy complex dissociates, giving an excess of V vacancies, which then diffuse away. At the same time, the excess V^- provide an enhanced number of $\mathrm{P}^{+}V$ complexes and increase the diffusivity over that ordinarily observed during intrinsic diffusion by increasing D^* .

8.4.2 Silicon Interstitial Diffusants

Fig. 8.30 gives D values for a number of fast interstitial diffusants in silicon. However, since the diffusivity of most of these impurities has not been examined in depth, profiles calculated from these numbers may be grossly in error. As an example, consider gold, which has been studied extensively (5, 73). Gold is thought to diffuse both

FIGURE 8.29

Profile of phosphorus diffusion with high surface concentration. (*Source:* Adapted from R.B. Fair and J.C. Tsai, *J. Electrochem. Soc. 124*, p. 1107, 1977.)

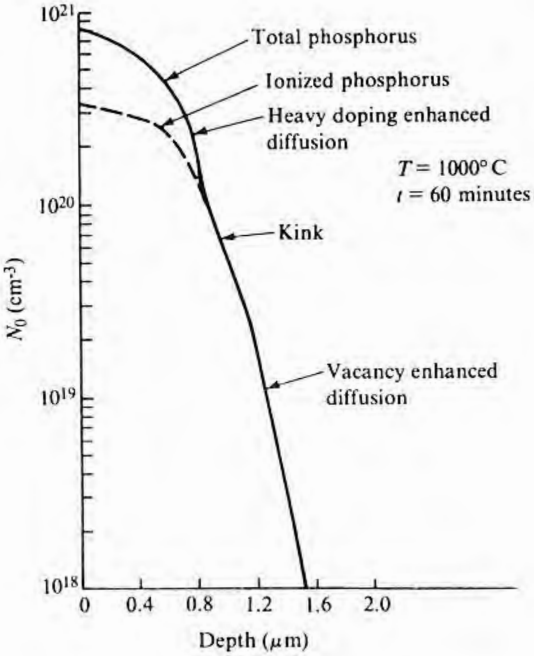
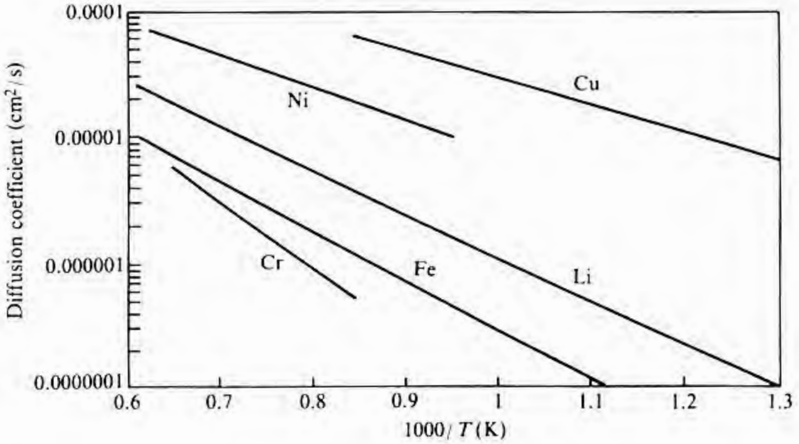


FIGURE 8.30

Diffusion coefficients of some fast diffusers in silicon. (*Source:* From data in Eicke R. Weber, *Appl. Phys. A30*, p. 1, 1983.)



substitutionally and interstitially, with $D_{\text{int}} >> D_{\text{sub}}$. D_{int} and D_{sub} are shown in Fig. 8.31 as the upper and lower curves. Experimentally, much of the reported data has followed one of the three lower curves (5) and has been interpreted as following either Eq. 8.14b (curve II) or Eq. 8.12 (curve I). Other data appear to be best explained by assuming that silicon self-interstitial diffusion is the limiting factor in gold diffusion as is described by Eq. 8.15 (6, 7).

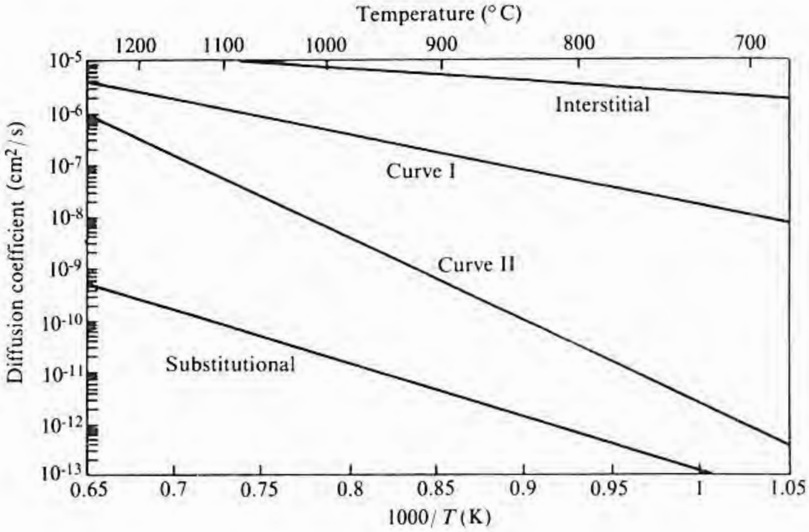


FIGURE 8.31

Diffusion coefficients of gold in silicon. (*Source:* Adapted from W.R. Wilcox and T.J. LaChapelle, *J. Appl. Phys. 35*, p. 240, 1964.)

8.4.3 Gallium Arsenide Profile

Gallium arsenide diffusivities are neither as well behaved nor as well understood as those of silicon. The donors are thought to diffuse substitutionally, and junction depths are linear with $\sqrt{t}$ (29). The acceptors apparently diffuse by an interstitial–substitutional mechanism (74) (see the discussion of zinc in gallium arsenide in section 8.2.3). The effective diffusion coefficients are generally concentration dependent as well as dependent on such things as the wafer gas environment, wafer defect density, and wafer capping (74, 75). Because of this variability, only two D values and activation energies are given in Table 8.2. The behavior of zinc (an acceptor) under carefully controlled conditions is shown in Fig. 8.32 (76). It is characterized by a flat top and a very abrupt falloff. Such curves are not described by any of the diffusion cases discussed, but a plot of the depth x_j versus $\sqrt{t}$ gives a straight line with a slope of D_{sur}, which is given by

$$D_{\text{sur}} = \frac{KN_{\text{sur}}^2}{(P_{\text{As4}})^{1/4}} \qquad 8.74$$

where N_{sur} is the dopant concentration at the surface, K is an equilibrium constant, and P_{As4} is the partial pressure of arsenic over the wafer.

8.4.4 Solid Solubility Data

In order to calculate profiles when using a fixed surface concentration, solubility versus temperature data are required as shown in Fig. 8.33 for the common substitutional dopants in silicon. Experimental gold solubility data are for substitutional atoms and mostly

FIGURE 8.32

(a) Diffusion profiles for zinc in gallium arsenide. (*Source:* R. Jett Field and Sorab K. Ghandhi, *J. Electrochem. Soc. 129*, p. 1567, 1982. Reprinted by permission of the publisher, The Electrochemical Society, Inc.). (b) x_j from part a plotted versus square root of diffusion time.

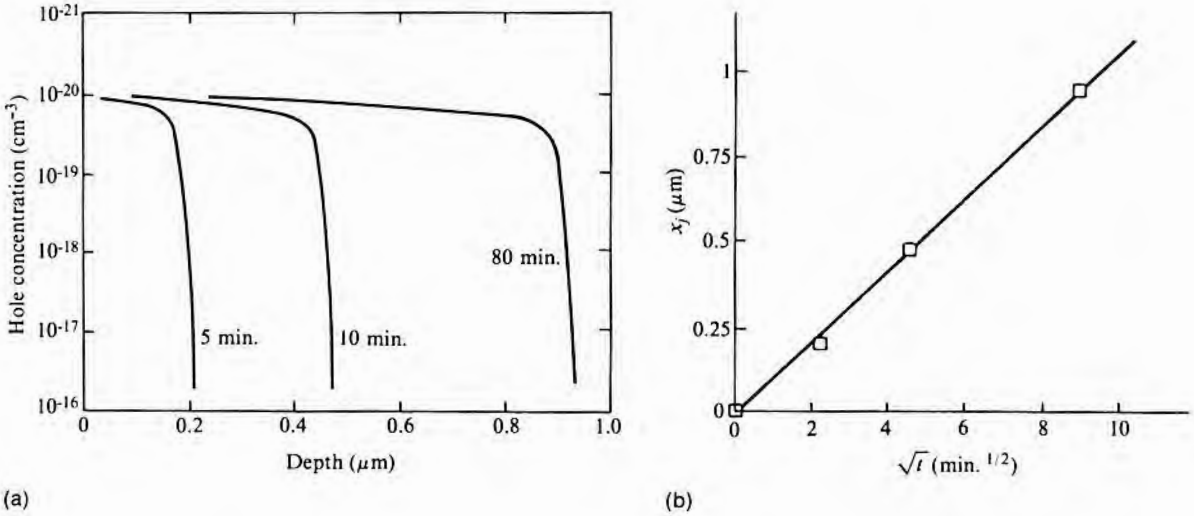
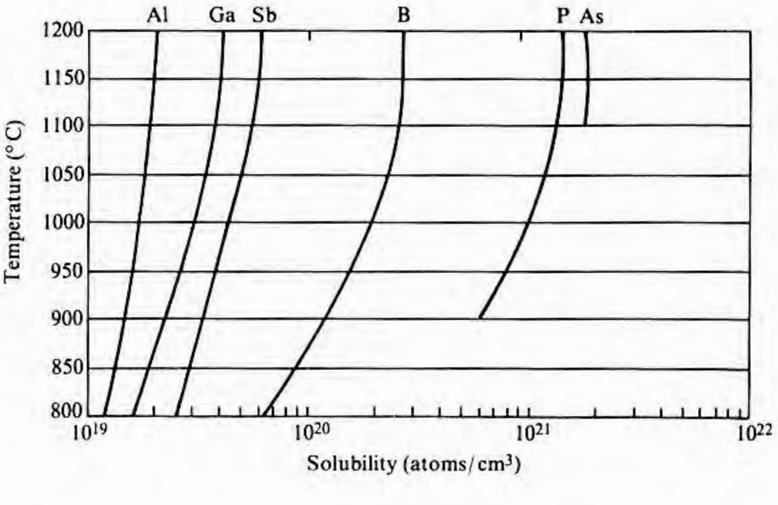


FIGURE 8.33

Solid solubility limits of various dopants in silicon. (*Source:* Adapted from F.A. Trumbore, *Bell Syst. Tech. J. 39*, p. 205, 1960, and G.L. Vick and K.M. Whittle, *J. Electrochem. Soc. 116*, p. 1142, 1969.)



for temperatures above 1200°C, but both interstitial and substitutional solubilities for lower temperatures have been calculated (73). Log solubility is linear in $1/T$ (K) and, for substitutional atoms, is $\sim 10^{17}$ atoms/cc at 1200°C and $\sim 5 \times 10^{13}$ atoms/cc at 700°C. Interstitial solubilities are $\sim 10^{16}$ atoms/cc at 1200°C and $\sim 10^{13}$ atoms/cc at 800°C. (The substitutional solubility is shown later in Fig. 8.39.)

In diffusion processes using ion implanting to put a fixed number of atoms near the surface as a source, solid solubility data are still required to make sure that the implant dose does not lead to precipitation.

8.5 DIFFUSION CHARACTERIZATION

The full characterization of a diffusion usually means a determination of the profile—that is, the determination of the impurity concentration versus distance from the surface—and an evaluation of the uniformity of the profile from point to point over the wafer surface. Unfortunately, such profiling is very time consuming and often requires complex equipment. Consequently, various alternative procedures have evolved. The two most common measurements are junction depth and sheet resistance, both of which are more closely related to device design and performance than the impurity profile itself. From these two measurements, the surface concentration can be inferred if N as a function of depth is known. When one is trying to control or change the junction depth x_j and the diffusion sheet resistance, the value of surface concentration becomes important and thus must be periodically evaluated. Diffusion uniformity over the wafer surface is generally specified in terms of sheet resistance uniformity.

8.5.1 Conversion from Dopant Concentration to Resistivity

During the course of characterization, it is sometimes the dopant concentration and sometimes the resistivity due to that dopant that is measured. Consequently, it is convenient to be able to quickly convert from one value to the other. The conductivity σ may be calculated from

$$\sigma = q\mu N_{ne} \qquad 8.75$$

where q is the electronic charge, μ is the carrier mobility, and N_{ne} is the net ionized impurity concentration.[14] For concentrations less than about 10^{17} atoms/cc, all of the dopant atoms will be ionized and $N_{ne} = N_{\text{net}}$. However, as the concentration increases, there will not

[14] The net ionized impurity concentration is determined from the net impurity concentration N_{net} given by $|N_A - N_D|$.

TABLE 8.3

Percent Ionization Versus Dopant Concentration

Concentration	10^{16}	10^{17}	10^{18}	10^{19}	10^{20}	10^{21}
Phosphorus (1, 2)	100		90	100*	90*	30*
Boron (3)	100	93	75			

*Experimental data.

1. S.S. Li and W.R. Thurber, *Solid-State Electronics 20*, p. 609, 1977.
2. R.B. Fair and J.C.C. Tsai, *J. Electrochem. Soc. 124*, p. 1107, 1977.
3. Sheng S. Li, *Solid-State Electronics 21*, p. 1109, 1978.

be complete ionization, and, in addition, there may be dopant–defect pairing that prevents part of the dopant from even being electrically active. Ionization data for boron and phosphorus are given in Table 8.3. For conversion of dopant concentration to resistivity, a set of curves as shown in Fig. 8.34 can be used instead of Eq. 8.75. Their nonlinearity is due to a combination of concentration-dependent mobility and the incomplete ionization shown in Table 8.3.

8.5.2 Sheet Resistance Measurement

Sheet resistance (R_s) of the diffused layer can be measured directly by a four-point probe (77) if a junction exists between the sheet being measured and the main body of the wafer. Such measurements are destructive only insofar as the wafer surface is either damaged or contaminated by the probe points. R_s is related to the doping of the diffused layer by

$$R_s = \frac{1}{q}\int_0^\infty \mu N_{ne}(x)dx \quad \Omega/\text{sq} \qquad 8.76$$

The average resistivity of a layer of thickness x_j is given by

$$\rho_{av} = R_s x_j \quad \Omega\text{-cm} \qquad 8.77$$

and the average conductivity is given by

$$\sigma_{av} = \frac{1}{R_s x_j} \qquad 8.78$$

When large quantities of data are desired, as, for example, when the uniformity of sheet resistance over a whole wafer is mapped, it may be more expeditious to put an array of metal patterns on the wafer surface and then use automatic wafer test equipment to collect the data. Specialized equipment is also available that steps probes over the wafer in a predetermined pattern and then displays the data in various formats such as contour maps, single-dimensional cross-sectional profiles, and percent deviation (78).

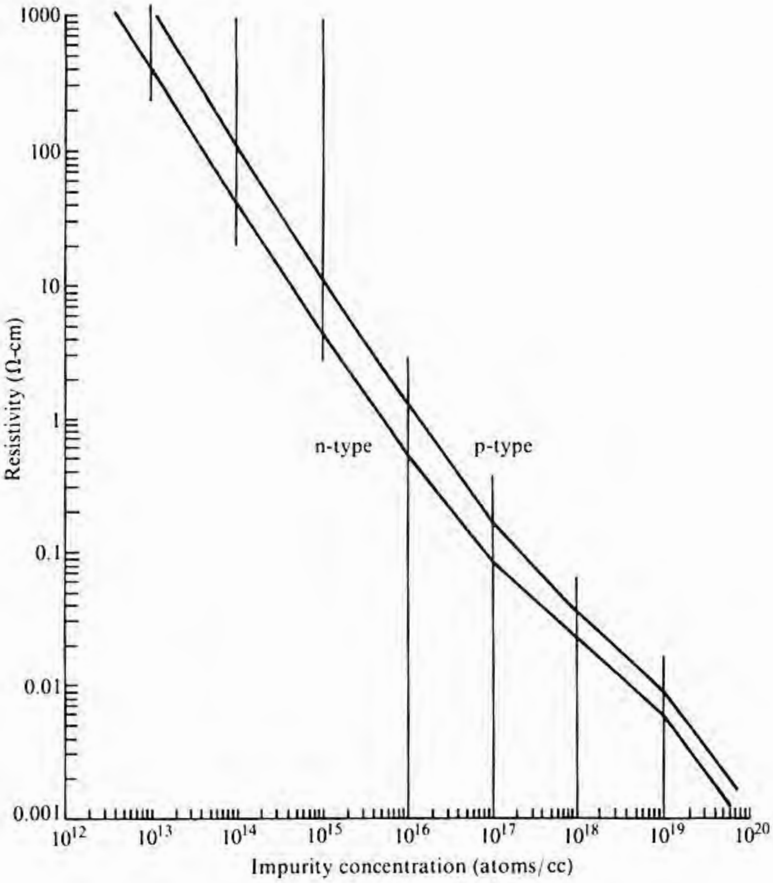


FIGURE 8.34

Resistivity versus impurity concentration for silicon. From data in S.M. Sze and J.C. Irving, *Solid State Electronics 11*, p. 599, 1968; W.R. Thurber et al., *J. Electrochem. Soc. 127*, p. 1807, 1980; and L.C. Linares and S.S. Li, *J. Electrochem. Soc. 128*, p. 601, 1981.

8.5.3 Determination of N_0

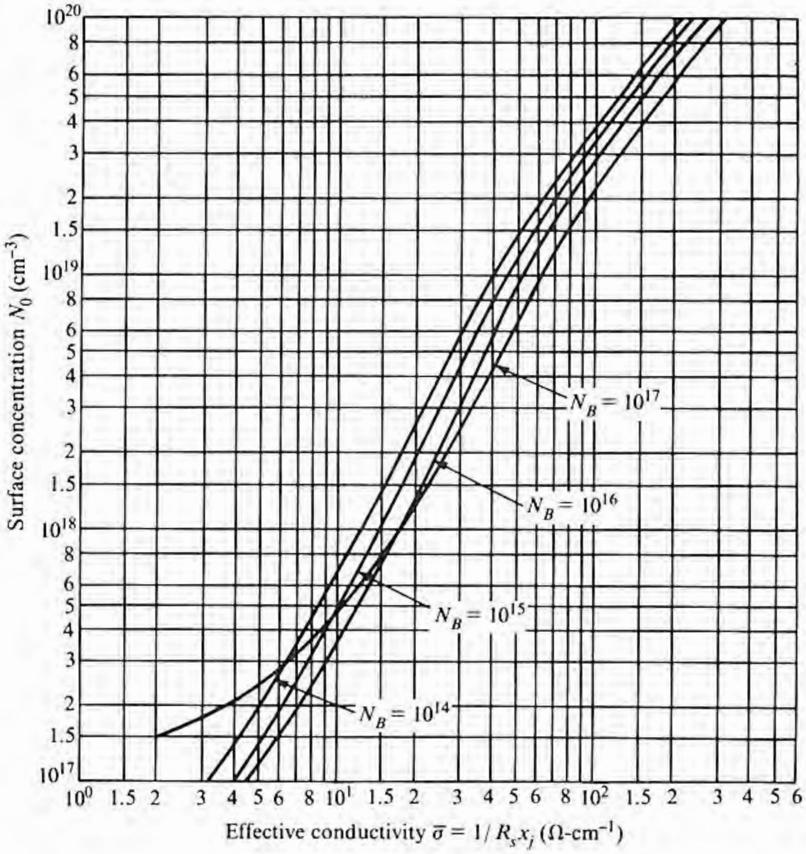
When the N_{ne} of Eq. 8.76 is known—for example,

$$N_{ne} \cong N_{\text{net}} = \left|N_0 \operatorname{erfc}\left(\frac{x}{2\sqrt{Dt}}\right) - N_B\right| \qquad 8.79$$

the sheet resistance can be calculated as a function of N_0 and N_B by using Eqs. 8.76 and 8.79, providing the mobility as a function of N_{ne} is known (see, for example, references 1 and 2 in Table 8.3 for data). By having made such calculations, curves such as those shown in Fig. 8.35 can be plotted for the range of values of interest and then N_0 read from them (79, 80). References 79 and 80 both have a selection of curves for error and Gaussian function distributions in silicon. However, reference 79 has used more recent electrical data. The Hall constant of the diffused layer can be measured, and, by using an analogous set of curves, the surface concentration can

FIGURE 8.35

Relation between average conductivity ($1/R_sx_j$) of n-type layers diffused into p-type wafer and surface concentration N_0. (Curves for four wafer background doping levels N_B are included.) (*Source:* Adapted from J.C. Irving, *Bell Syst. Tech. J. 41*, p. 387, 1962.)



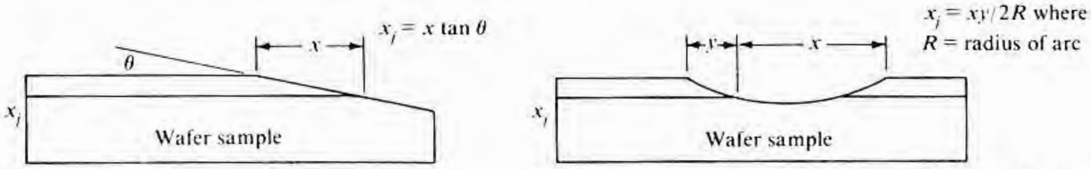
be determined (81). Since a Hall measurement, like the resistance measurement just discussed, includes only ionized impurities, conversion to total impurities requires a consideration of nonionized and precipitated impurities. In addition to these two rather specialized procedures, any of the profiling methods will give N_0 when extrapolated to $x = 0$.

8.5.4 Junction Depth Measurement

Junction depth measurement is ordinarily done by lapping and staining and is destructive. It is, however, quick and simple and requires only a small portion of a wafer. The general procedure is to expose a section through the wafer by some sort of bevel lapping in order to get mechanical magnification and then stain with a chemical solution that differentiates between p- and n-material. Fig. 8.36 shows two types of bevel that are often used. The beveling shown in part a is more appropriate for low angles and high magnification; that shown in part b is quicker and is normally used for in-line produc-

FIGURE 8.36

Use of beveling to magnify junction depth. (A chemical stain is used to define the position of x_j so that x and y can be measured.)



(a) Bevel sample

(b) Grind depression with spherical or cylindrical tool

tion checks. A typical stain for silicon is 1–3–10 etch, comprised of 1 ml HF, 3 ml HNO_3, and 10 ml acetic acid. The p-region will stain dark. A stain for gallium arsenide is 1 ml HF, 1 ml H_2O_2, and 10 ml H_2O applied under an intense white light. (For a more detailed description of the procedure, see Chapter 7 of reference 82.)

8.5.5 Profile Measurement

Diffusion profiling can be done by counting either the impurity atoms or the free carriers produced by the impurities. The oldest profiling method is a combination of sheet resistance measurements and sequential removal of material from the wafer surface (66). It can be shown that if R_1 and R_2 are respectively the sheet resistances measured when the surface is at two closely spaced depths x_1 and x_2 below the original surface, then the bulk resistivity of the thin layer Δx between x_1 and x_2 is given by

$$\rho = \frac{R_1 R_2 \Delta x}{R_2 - R_1} \qquad 8.80$$

The thin layers can be removed by etching the semiconductor (84) and, in the case of silicon, also by anodically oxidizing the surface and then stripping the oxide (85).

The value of $\rho(x)$ from Eq. 8.80 depends linearly on the value of Δx, which is difficult to measure accurately. It also depends on the difference between successive values of sheet resistance. To simplify the task of smoothing the data, the log of the surface conductance G (which is $1/R$) versus x can be plotted and smoothed as desired, without having the possibility of gross errors from an incorrect Δx or $R_n - R_{n+1}$ (86). The slope s of that curve for any value of x can be determined from the smoothed curve and is also given by

$$s = d(\log G)dx = \frac{1}{G}\frac{dG}{dx} \qquad 8.81$$

The bulk conductivity σ in terms of the measured sheet conductance G is given by

$$\sigma(x) = \frac{dG}{dx} \qquad 8.82$$

Combining these two equations gives

$$\sigma(x) = G(x)s(x) = \frac{1}{\rho(x)} \qquad 8.83$$

Rather than successively removing layers, the wafer can be beveled, and either a 4-point probe or a spreading resistance probe can be stepped down the bevel (86–88).

Capacitance–voltage profiling, using a metal Schottky barrier contact or an MIS capacitor, can profile from near the surface to the depth where avalanche breakdown occurs. By using a MOSFET transistor with back-gate bias, a depletion zone can be moved out to where avalanche occurs, and the profile can be determined to that point by measuring the MOS transistor voltages (89–91).

Secondary ion mass spectrometry (SIMS) is widely used to directly measure the dopant atom concentration as a function of depth. It is most applicable to shallow profiling, but the industry trend for several years has been toward shallower junctions. The SIMS technique consists of using an ion beam (usually oxygen or cesium) to sputter away the semiconductor and produce ions that can then be mass-analyzed (92). Table 8.4 shows the sensitivities that can be expected. The ease of ionization is ordinarily the largest factor in determining sensitivity. However, interference between the desired ion and ion complexes causes reduced sensitivity in some cases. This problem can sometimes be alleviated by measuring a different isotope. An example of the problem is Si–H, with masses

TABLE 8.4

SIMS Sensitivity to Dopant Atoms

Dopant	*Semiconductor*	*Sensitivity (atoms/cc)*
Boron	Silicon	10^{14}
Phosphorus	Silicon	2×10^{16}
Antimony	Silicon	10^{16}
Arsenic	Silicon	10^{16}
Gold	Silicon	10^{15}
Chromium	Gallium arsenide	2×10^{15}

of 29, 30, and 31, which interferes with the only stable isotope of phosphorus, ^{31}P. To eliminate the problem with ion complexes being formed during sputtering, an accelerator mass spectrometer has been substituted for the conventional mass spectrometer used in a normal SIMS instrument. The result is a lowering of the detection limits of phosphorus and arsenic by over an order of magnitude (93).

Rutherford backscattering (RBS) can be used for profiling and, in addition has the potential for separating interstitial and substitutional atoms. It is more applicable to atoms of higher mass number than the semiconductor they are in and has, for example, been used to study antimony profiles in silicon (92).

Before the advent of the newer techniques of SIMS and RBS, sectioning by etching, combined with radioactivity counting, was used to obtain profiles independently of the electrical properties. Either the dopant could be a radioactive isotope (radiotracer analysis), or else after diffusion, the sample could be subjected to neutron irradiation in order to produce a radioactive species (neutron activation analysis) (94). The radiotracer method also allowed profiles after an isoconcentration diffusion to be determined. In this kind of study, the radioactive source vapor pressure is the same as the equilibrium vapor pressure due to the nonradioactive dopant of the same species already present so that there is no net flow of impurity into or out of the wafer surface.

8.5.6 Determination of Diffusion Coefficient

When the diffusion coefficient is independent of concentration (Eq. 8.26 is applicable) and the solution for $N(x)$ is known, there are several ways to rather easily determine D (95). If a series of diffusions with differing t's are made from a fixed concentration source into wafers of opposite conductivity type with a concentration N_B, a plot of x_j versus $\sqrt{t}$ will give a straight line (Eq. 8.36). From the slope of the line, D can be determined. In principle, only one t and x_j need to be measured, but plotting a series of values allows data smoothing. If N_0 cannot be found by some method such as the use of initial sheet resistance and curves like those of Fig. 8.35, then two wafers with widely differing N_B's can be simultaneously diffused so that the two surface concentrations are the same. Dividing the expression for N_{B1} by that for N_{B2} (Eq. 8.32) eliminates N_0 and gives

$$\frac{N_{B1}}{N_{B2}} = \frac{1 - \operatorname{erf}(x_{j1}/2\sqrt{Dt})}{1 - \operatorname{erf}(x_{j2}/2\sqrt{Dt})} \qquad 8.84$$

from which D may be determined by successive approximations (96). Alternatively, the approximate expression for erfc(z) for large z from section 8.10 can be used to solve Eq. 8.84 (97):

$$D \cong \frac{(1/4t)(x_{j2}^2 - x_{j1}^2)}{\ln(N_{B1}x_2/N_{B2}x_1)} \qquad 8.85$$

It must be emphasized that if the profile is not an error function, interpreting the data in this fashion will lead to erroneous results. If, for example, the actual profile is like the one of Fig. 8.37, which is typical of high-concentration diffusions, not only will an N_0 determination from a sheet resistance and an x_j measurement be wrong, but the calculated D value will appear to depend on the background doping level (10). This kind of interpretation apparently was the reason for some early reports of such D value dependency.

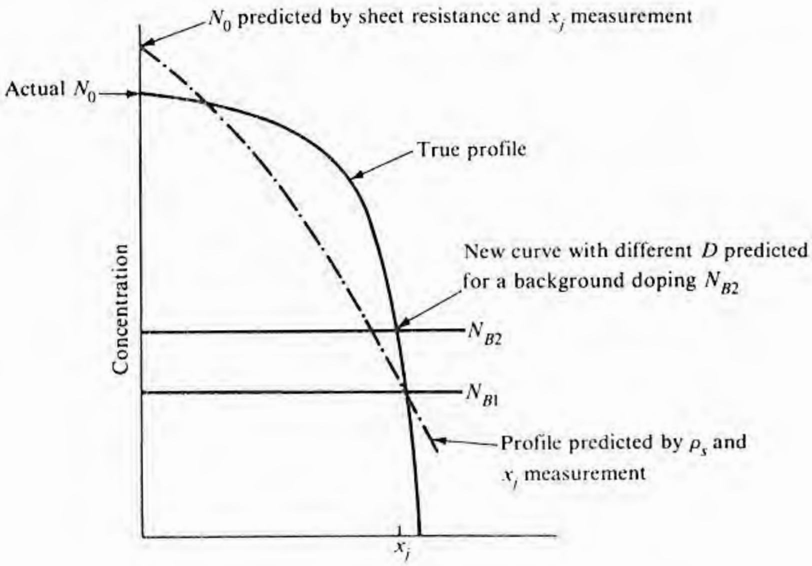
If diffusion is from a limited source instead of from an infinite source, if S is known—for example, by having ion implanted a known density of diffusant—and if the diffusion is for a long enough time for x_j to be much greater than the implant range R, then D can be determined from Eq. 8.41b. Diffusing into wafers with two different doping levels allows S to be eliminated in the same manner that N_0 was in Eq. 8.84. In this case, D is given by

$$D = \frac{(1/4t)(x_{j2}^2 - x_{j1}^2)}{\ln(N_{B1}/N_{B2})} \qquad 8.86$$

It should be noted that in Eqs. 8.85 and 8.86, D depends on the difference between the squares of the junction depth and thus is quite sensitive to errors in measurement of the x_j's. If a profile of N versus x over a wide range of x is made for a diffusion time t, a plot

FIGURE 8.37

Errors introduced during attempt to interpret a diffusion profile due to a concentration-dependent dopant in terms of a constant D profile.
(*Source:* Adapted from S.M. Hu in D. Shaw, ed., *Atomic Diffusion in Semiconductors*, Plenum Publishing Co., London, 1973.)



of log N versus x^2 should be a straight line with a slope of $-1/4Dt$. That this is true can be seen by taking the log of both sides of Eq. 8.39, which gives

$$\log N = \log\left(\frac{S}{\sqrt{\pi Dt}}\right) - \frac{x^2}{4Dt} \qquad 8.87$$

When D is concentration dependent, the methods just described are not appropriate for determining its value. For the conditions of diffusion from an infinite source and from a concentration step, the Boltzmann–Matano method is ordinarily used (1). This method is based on substituting a new variable $\zeta = x/\sqrt{t}$ into Fick's second law as written for a concentration-dependent D. This substitution gives an ordinary differential equation that can be solved analytically.[15] First, relate $\partial N/\partial t$ and $\partial N/\partial x$ to the new variable ζ. That is,

$$\frac{\partial N}{\partial t} = \left(\frac{\partial \zeta}{\partial t}\right)\frac{\partial N}{\partial \zeta} = -\left(\frac{x}{2t^{3/2}}\right)\frac{dN}{d\zeta} \qquad 8.88a$$

$$\frac{\partial N}{\partial x} = \left(\frac{\partial \zeta}{\partial x}\right)\frac{\partial N}{\partial \zeta} = \left(\frac{1}{t^{1/2}}\right)\frac{dN}{d\zeta} \qquad 8.88b$$

Substituting these values into

$$\frac{\partial N}{\partial t} = \frac{\partial D}{\partial x}\frac{\partial N}{\partial x} + D\frac{\partial^2 N}{\partial x^2} \qquad \text{(Fick's second law)}$$

gives

$$-\left(\frac{\zeta}{2}\right)\frac{dN}{d\zeta} = \frac{d(DdN/d\zeta)}{d\zeta} \qquad 8.89$$

By integrating Eq. 8.89 once and rearranging,

$$D(N) = -\left(\frac{1}{2}\right)\frac{d\zeta}{dN}\int_0^N \zeta dN \qquad 8.90$$

or

$$D(N) = -\left(\frac{1}{2t}\right)\left(\frac{dx}{dN}\right)_N \int_0^N x dN \qquad 8.91$$

[15] Boltzmann used the new variable as a means of solving the diffusion equation (*Wied. Ann. 53*, p. 959, 1894). Later, Matano used it to determine D (*Jap. J. Phys. 8*, p. 109, 1933).

Thus, if an N versus x profile is determined, dx/dN at a desired point on the graph can be determined, and the integration can be performed graphically. When the variable $\zeta = x/\sqrt{t}$ is not appropriate, for example, in diffusion from a limited source, often other relations will allow D to be determined in a similar fashion (98–101).

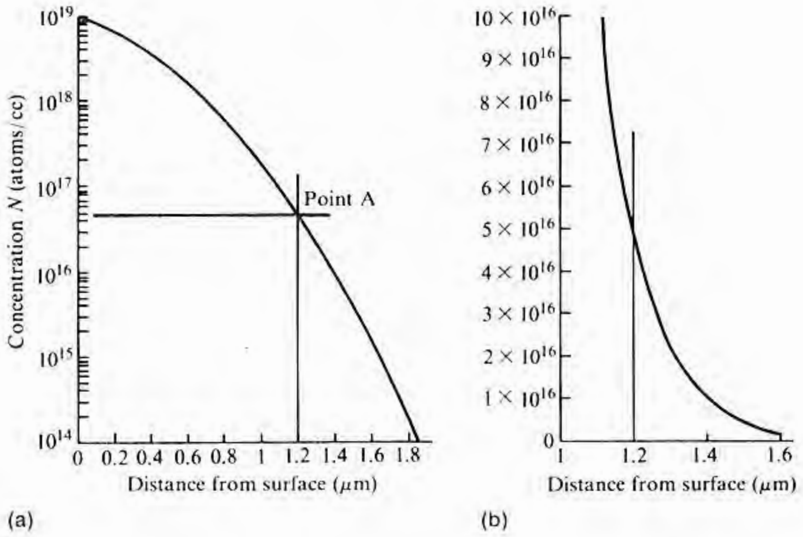
EXAMPLE □ Using the curves of Fig. 8.38, determine the diffusion coefficient using the Boltzmann–Mantano method. Since these curves have been plotted assuming a constant D, any choice of position should give the same value. The diffusion time was 10 hours.

Choose point A for evaluation. The value under the horizontal curve is the integral required and is about 6.2×10^{12} atoms/cm^2. The slope of the profile dx/dN at point A (measured from the linear plot) is about -3.3×10^{-22} cm^4/atom. The time is 7.2×10^4 s. Substituting these values into Eq. 8.91 gives $D = 2.8 \times 10^{-14}$ cm^2/s. The curve was plotted using a value of $D = 2.5 \times 10^{-14}$ cm^2/s, so the agreement is satisfactory. □

When one is examining the effects of variables such as crystallographic orientation and oxidation conditions on diffusion results, it is often difficult to compare results obtained from separate wafers. Consequently, devising structures and/or flows that enable comparisons to be made on the same wafer is often helpful. A simple example is the use of a nitride mask over part of a wafer during drive-in to study the effect of oxidation-produced defects on diffusion rate. More complex examples are the use of orientation-dependent

FIGURE 8.38

Profile to use with Boltzmann–Matano method of determining D. (Curve b is a linear plot to simplify the determination of dx/dN and the numerical integration of xdN.)



etch to expose both (111) and (100) faces on the same wafer (102) or the use of twinned or polycrystalline wafers to provide two or more orientations on the same planar surface.

8.6 DIFFUSION PROCESSES

Diffusion processes can be characterized in terms of the function of the diffused structure such as a base or a source/drain diffusion. From a processing standpoint, however, they can be better described in terms of high or low surface concentration, high or low diffusion temperature, and deep or shallow diffusion depths. The approximate limits of each of these factors as defined by common usage are shown in Table 8.5.

8.6.1 Depth

MOS circuits are characterized by shallow diffusions. MOS transistors require only a very thin layer for the channel, and any source–drain depth below it usually only adds unwanted capacitance. Thus, source/drain diffusions can be 1 μm or less in depth. (With very thin diffusions, great care must be exercised to ensure that the contacts do not alloy completely through the diffused layer.) The exception to shallowness is the formation of diffused wells in CMOS, but even then the well depth is usually only 5–6 μm. The collector isolation diffusion is the only deep bipolar diffusion. It must extend from the surface down to the epitaxial substrate, which may be up to 20 μm away for older designs. Truly deep diffusions are found in discrete devices like triacs, SCRs, and power rectifiers.

Deep diffusion processes will normally be at high temperature in order to reduce diffusion time and to provide the high surface concentration that is also usually required. For very deep power device p-diffusions, aluminum is often used because of its high diffusion coefficient. In order to control shallow diffusions, low temperatures are usually required. Since the total amount of impurity is small, an ion implantation predep, followed by a thermal drive-in, is usually used. In many cases, the implantation is followed only by an anneal step.

8.6.2 Surface Concentration

The surface concentration is normally set by the device requirements and is controlled by a combination of diffusion temperature or implant dose and total diffusion time. When a one-step diffusion is used, the solid solubility of the dopant at that temperature can be

TABLE 8.5

Diffusion Process Ranges

	Temperature (°C)	*Surface Concentration*	*Depth (μm)*
High	1150–1250	$>5 \times 10^{19}$	>15
Low	850–1000	$<5 \times 10^{19}$	<2

used to hold the surface concentration constant. As Fig. 8.33 showed, the solid solubilities versus temperature for common dopants in silicon change no more than a factor of 3 over the 800°C–1200°C temperature range. Thus, while the concentration can be fixed by solid solubility, it cannot be varied over a very wide range of values by changing temperature. To obtain a wider range, other methods must be used (none of which will provide higher concentrations). A direct approach, but one that does not give very good control, is just to reduce the dopant source flow if a gas or liquid source is being used. Another approach is to first grow a thin thermal oxide and let the glassy source layer form on top. If the diffusion time is very short, as for an emitter diffusion, the dopant can diffuse through the thin oxide, with the result that the concentration at the silicon surface is reduced (103). With longer times, the glassy source layer will dissolve the thin oxide and then again be in contact with the silicon surface. When a two-step diffusion (predep or ion implant followed by a drive-in) is used, the initial surface concentration is set by temperature or implant dose, but the final lower concentration is determined by the time and temperature of the subsequent drive-in. Alternatively, CVD diffusion source layers with lesser dopant concentrations can be deposited on the surface to obtain a given surface concentration even though the diffusion is still made at a higher temperature. (See section 8.9 on diffusion sources for more details on various diffusion sources.)

8.6.3 Lifetime Control

In most cases, a high semiconductor bulk lifetime is desired for reasons such as minimizing reverse junction leakage, improving bipolar transistor gain, and increasing refresh time in dynamic MOS memories. Some circuits, however, such as TTL logic, depend on a low lifetime in the collector regions of the transistors to improve switching time. Substitutional gold, whose deep levels provide for carrier recombination, is most often used (73, 104), although platinum is a possible alternative. For a given concentration, platinum lowers the lifetime more than gold does and, for a given lifetime, produces fewer deep-level centers and hence less generation leakage current (105, 106).

Both gold and platinum are introduced by diffusion, and both are fast diffusers. Thus, the length of time required to diffuse to the required region is generally quite short. The diffusion is complicated by the fact that the highest solubility regions for the killer impurity are the heavily doped ones, while the regions where the lifetime reduction is required are the lightly doped ones. Thus, heavy doping acts as a sink and "getters" the gold (and other lifetime killing impurities). Data indicate that for phosphorus concentrations over about 2 × 10^{19}, substantial gettering occurs (107, 108). Curves such as the one in Fig. 8.39 can be used to relate the amount of gold required in a region to give the desired device parameters (transistor switching time or diode recovery time). It is, however, difficult to estimate the amount that may be trapped in adjacent regions. Therefore, gold or platinum concentrations are difficult to predict. Usually, the procedure is to diffuse at a temperature at which the solid solubility of the substitutional gold or platinum is adequate to control lifetime and for a time adequate to ensure that the region of interest is saturated. Fig. 8.39 also relates gold solubility to processing temperature. After diffusion, the wafer is quenched rapidly enough to prevent precipitation. As will be discussed in a later section, since rapid cooldowns can generate excessive crystallographic damage, process trade-offs may be required.

Since diffusion times required are relatively short, the gold (or platinum) is introduced at the end of the diffusion cycles, and because of gettering by heavily doped regions, introduction is almost always directly into the collector region. The simplest procedure is to introduce the gold from the back of the wafer. If, however, a heavily doped layer exists between the back of the wafer and the collector–base junction, as for an n^+-epi substrate or a subcollector diffusion, that layer will interfere with the gold's diffusing to the required region. Sometimes, diffusion is through collector contact openings in the oxide, but if the concentration of gold is too high, the collector near the surface will be compensated, causing high contact resistance and poor device performance. To eliminate the problem, some designs open up special lifetime doping holes in the oxide.

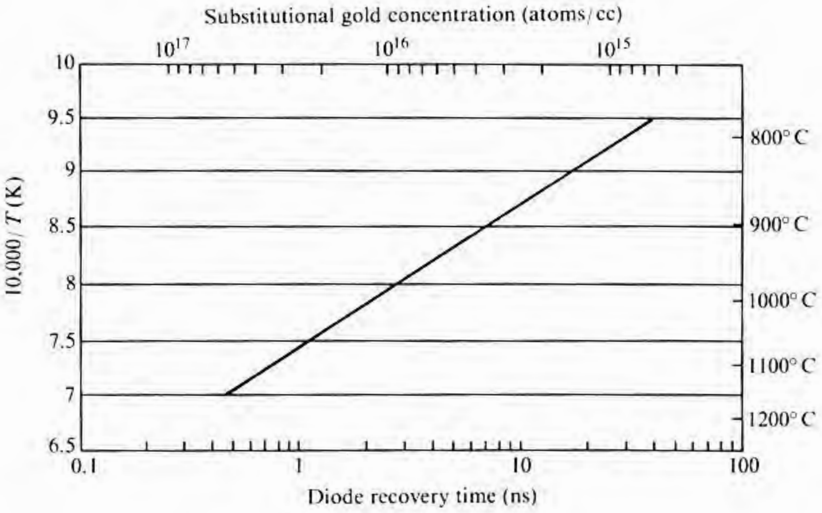


FIGURE 8.39

Relations between diode recovery time, gold concentration, and gold diffusion temperature. (*Source:* Adapted from W.M. Bullis, *Solid-State Electronics* 9, p. 143, 1966.)

Lifetime can also be controlled by nondiffusion processes such as high-energy particle irradiation (electrons and protons). Irradiation effects can usually be rather easily annealed out and hence are seldom used in production devices. Localized ion implanted argon can be used for selective lifetime control and is reported to provide a useful range of lifetimes even after 1200°C processing (109).

8.6.4 Gettering of Silicon

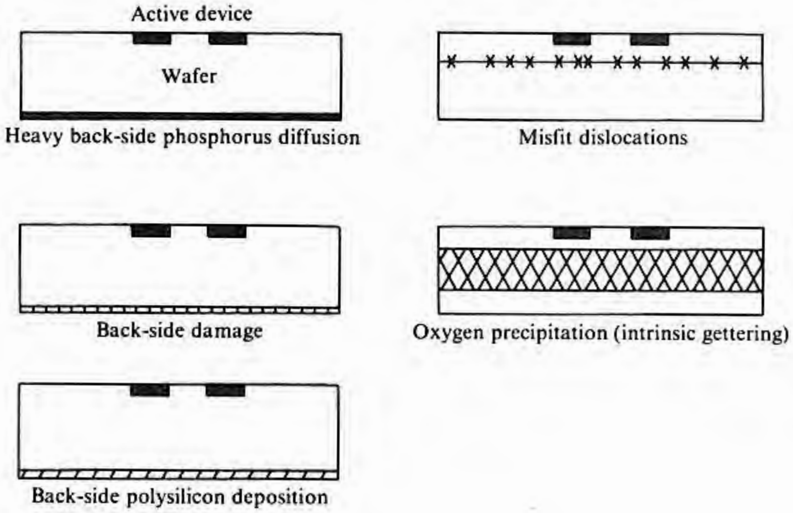
Gettering steps are used to remove lifetime reducing dopants (usually some of the heavy metals) from regions of the circuit where their presence would degrade performance. The impurities most likely to be found are gold, nickel, copper, and iron. However, all of the transition metals (Ti, V, Cr, Mn, Fe, Co, Ni) are reported to be deleterious (110). The most common effect is that of lifetime reduction, but gold, for example, reduces mobility in MOS structures (108, 110–113) and increases the resistivity for both n- and p-type as the substitutional gold concentration approaches that of the IIIA/VA dopant. In addition, any of the metals with high solubilities at processing temperature and very low solubility at room temperature are prone to form precipitates and cause excessive leakage or even direct shorts between transistor elements. It is also reported that metallic precipitates near the surface can reduce the thickness of thermal oxide grown over them (114). As noted in Chapter 3, the presence of these metals will cause S-pits after oxidation and lead to the subsequent formation of stacking faults. The stacking fault generation is particularly troublesome in the oxidation–diffusion–epitaxy sequence used to produce bipolar buried-collector layers. Much effort is directed toward maintaining a clean process (see Chapter 3), but, in most instances, some gettering in conjunction with the diffusion steps is still required during silicon IC fabrication. Gold is the earliest lifetime killer identified (115), the most common contaminant, and the one most studied.

Gettering processes depend either on enhanced solubility of the metal in heavily doped n-material or on providing sites for enhanced precipitation in regions where they will do no harm. Various gettering methods are listed in Table 8.6 and shown schematically in Fig. 8.40. While lifetime killers are characterized as fast diffusers, as both diffusion times and temperatures are reduced, it becomes more difficult for back-side gettering sinks to function. Consequently, those methods that provide gettering locations closer to the front surface and the active devices become more desirable.

Heavy n-layers on the back of the wafer (116–121) were the first gettering sinks used and are probably still the most common. They are the natural consequence of an npn transistor emitter diffusion or an n-channel source/drain diffusion and an unprotected back surface. Hence, their use for gettering requires no additional processing

FIGURE 8.40

Location on wafer of gettering region for various gettering methods. (Note that only misfit dislocations and oxygen precipitation allow gettering close to the active device junctions.)


Active device
Wafer
Heavy back-side phosphorus diffusion
Misfit dislocations
Back-side damage
Oxygen precipitation (intrinsic gettering)
Back-side polysilicon deposition

steps. Heavily n-doped regions have increased substitutional solubility for those lifetime killers that behave as acceptors when substitutional (for example, Au and Cu), so, in principle, heavily doped phosphorus, arsenic, or antimony layers should perform equally as well. Experimentally, however, it is found that phosphorus is much more effective than the others (118). This fact has been explained by the formation of phosphorus–metal pairs (118) and by the extra crystal defects formed when silicon phosphide precipitates (119). Regions of misfit dislocations induced into (100) wafers have also been shown to be effective gettering sites (122). Heavy boron doping should increase the solubility of interstitial gold and hence also getter. However, interstitial solubility is much less than substitutional solubility, and it is experimentally observed that p^+-gettering is not nearly as effective as n^+. Some evidence exists that, unlike phosphorus, the boron glassy source associated with p^+-diffusions collects some of the gold (100).

Mechanical back-side damage gettering has been studied at least since 1965 (123–126). By using controlled sandblasting or other forms of light abrasion, the results are reasonably reproducible, and indeed wafers with such damage are now standard items of commerce. This sort of damage is applicable to processes where the formation of a heavy phosphorus layer presents problems. Within limits, the more abrasion, the more pronounced the gettering effect. For process control of such mechanical damage, light scattering from the abraded surface as measured by an instrument such as a paint gloss meter can be used.

Localized laser melting can be used to generate the back-side damage and does not involve the dirty processing required of abrasion (127, 128). A pulsed laser with a wavelength absorbed by the wafer, such as Nd:YAG, is used to scan the surface, generally producing nonoverlapping rows of overlapping melted regions. The thermal shock associated with the rapid heating and cooling produces a network of dislocations for the gettering. Some minimum power per pulse is required to ensure that the damage is severe enough that it does not anneal out and is on the order of 15–20 J/cm^2 (128).

Back-side ion implantation gettering is now used with some regularity. By selective masking, it can also be used locally on the front of the wafer and possibly be closer to the volume needing gettering than if it were on the back-side. A wide range of procedures and results have been reported (129–135). Methods of annealing and the point in the process at which the implant is done affect the final results, as does the implant species and the wafer orientation. Argon is one of the more effective ion implant species, and some correlation exists between gettering efficiency and the misfit ratio of the implanted ion to silicon (130). Much more residual damage remains in (111) oriented silicon than in (100), and the gettering efficiency is correspondingly higher. Dosages for implanting into bare silicon range from 10^{14} and 10^{15} ions/cm^2, with an implant energy of 100–200 keV. When implanting through a thin oxide layer, 10^{16} atoms/cm^2 have been used (135).

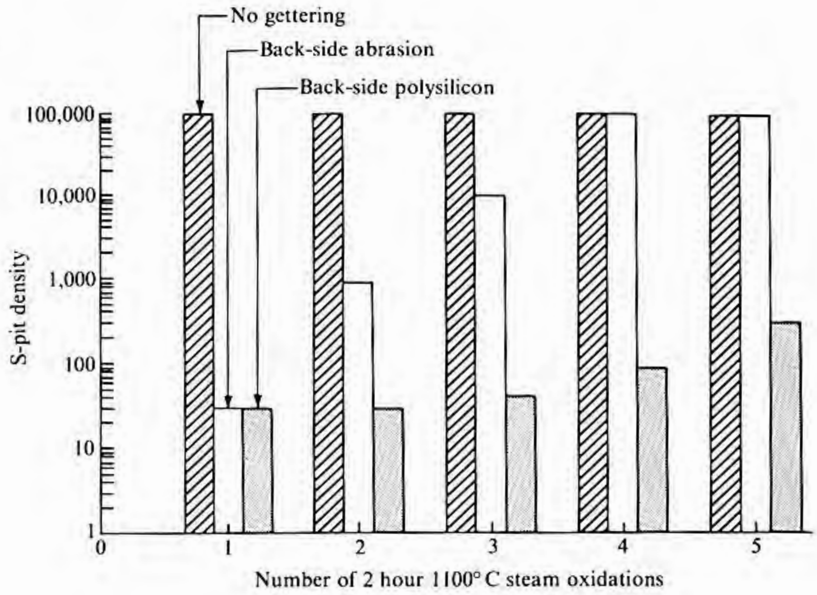
A silicon nitride layer deposited on the back of a wafer will provide enough stress during high-temperature processing to effectively getter (136, 137). Both the low-temperature plasma-deposited nitride and the higher-temperature CVD nitride appear effective. Thicknesses required are in the 1000–4000 Å range. Gettering is less for the thinner films, and wafer bowing can occur if the film is much above 4000 Å.

A thin layer of low-temperature polycrystalline silicon deposited on the back of the slice will provide for gettering sinks (138). The polycrystalline layers are deposited at low temperature, such as 650°C, to a thickness of about 0.5 μm. Even if the layers are thin enough to be fully oxidized during processing, stacking faults that continue to getter are propagated into the single-crystal silicon (138, 139). The gettering efficiency of a polysilicon layer is reported to be less easily annealed out during high-temperature cycling than the surface damage processes discussed earlier. The relative efficiency of back-side damage and a polysilicon layer in reducing S-pits as a function of the number of heat cycles is illustrated in Fig. 8.41.

A different kind of gettering, and one not included in Table 8.6, is the use of an atmosphere in the furnace tube during diffusion (or

FIGURE 8.41

General trend of back-side gettering efficiency versus amount of processing as measured by S-pit reduction. (The final S-pit value will depend not only on the gettering process but also on the initial contamination level.) (*Source:* Adapted from data in *Monsanto Applications Note*, AN9-7/82, revised January 1, 1983.)



oxidation) that will assist in transferring unwanted impurities from the wafer surface to the gas stream. It has been known since 1960 that the use of chlorine or a chlorine-bearing species such as PCl_3 or BCl_3 would improve lifetime (140). In the case of phosphorus and boron, the relative impact of the glassy layer, the heavy doping in the silicon, and the chlorine species in the tube ambient were not resolved. Since then, the use of HCl to clean tubes has become standard practice. The mechanism is one of forming chlorides volatile at the tube temperature, and the same mechanism can be used to clean wafers if etching of the surface can be prevented. Such conditions prevail during HCl oxidation, and substantial lifetime improvement is sometimes observed (141–143). In this case, the effectiveness will also depend on how easily the impurity to be gettered can diffuse through the protective thermal oxide and whether or not it is trapped at the interface. In the case of gold, it appears that there is a substantial pileup at the Si–SiO_2 interface (144) and that gold is not effectively gettered from wafers during an HCl oxidation (145).

Another gettering method not included in Table 8.6 and seldom used is the application of metallic layers to the silicon surface to either trap impurities or prevent them from entering the wafer (146, 147).

The misfit dislocations shown near the surface in Fig. 8.40 are normally introduced during an epitaxial operation and were dis-

TABLE 8.6

Gettering Methods

Item	*Comments*
Back-side phosphorus diffusion	
Back-side abrasion	Usually uses sandblasting.
Laser damage	Crystal damage introduced by local thermal shock.
Ion implant damage	Typically uses argon ions.
Silicon nitride deposition	~4000 Å layer deposited on the back of the wafer.
Back-side polysilicon deposition	Combination of polysilicon, residual oxide, and diffusion temperature introduces crystal damage.
Misfit dislocations	Must be introduced during an epitaxial step (see Chapter 7).
Oxygen precipitates	Requires correct oxygen level in crystal and a prescribed heat-treat cycle to be effective.

cussed in Chapter 7. They could also, in principle, be introduced by the high-energy implant of an appropriate impurity atom.

Intrinsic gettering generally refers to gettering by oxygen precipitates in the bulk of a silicon wafer. The term could, however, equally as well be applied to some other gettering mechanism originating with grown-in defects. The majority of silicon crystals are grown by the Teal–Little pulling process (often called CZ, or Czochralski) from a fused silica container. The molten silicon dissolves a small part of the fused silica, and some of the oxygen from it is incorporated into the single-crystal silicon. The amount depends on specific growth procedures but is in the 10^{18} atoms/cc range. As the crystal comes from the puller, most of the oxygen is dispersed as interstitial atoms and is electrically inactive. However, heat treatments such as are involved in diffusions can cause it to aggregate and eventually form rather large precipitates. The small aggregates are donors and can be sufficient to reduce high-resistivity material down to a few tenths Ω-cm n-type (148). Treatment in the 1300°C temperature range will cause a redissolving of the oxygen structures, but any crystallographic defects, such as stacking faults that formed because of the precipitates, will remain.

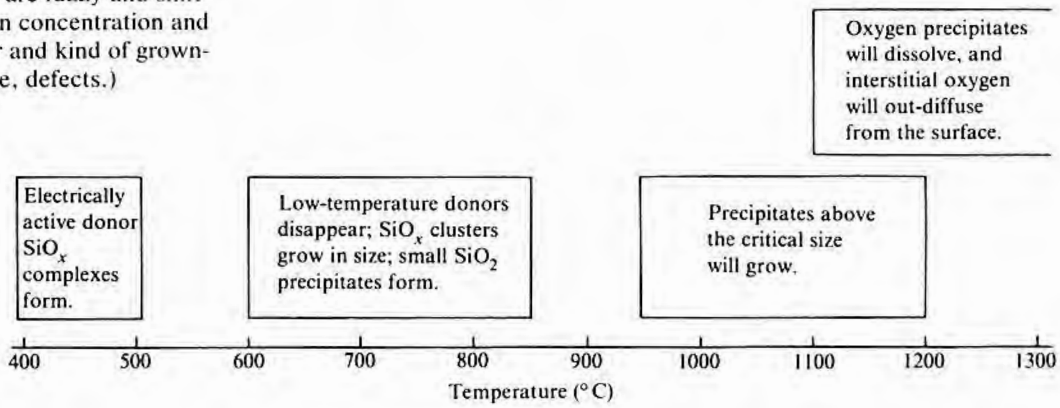
The oxygen-precipitation-induced defects have been shown to act as heavy-metal gettering sites (149), but if such sites are within the active volume of the IC devices, a yield degradation results. The idealized approach to intrinsic oxygen gettering is to first form a region from the wafer surface toward the wafer interior that is thick enough to contain the devices made in that wafer and has so little oxygen that it will not precipitate. Later, during subsequent processing, this region will remain free of oxygen-induced defects and thus not adversely affect devices made in it. The high-temperature operation to allow oxygen near the surface to out-diffuse and also dissolve any oxygen aggregates that may be in the wafer is referred to as a "denuding" step. After denuding, the wafer is subjected to a low-temperature heat cycle in order to allow precipitate nuclei to form. With this preparation, the dissolved oxygen will form large precipitates during the heat cycles associated with subsequent oxidations and diffusions. These large precipitates, in turn, produce crystal defects that act as gettering sites. This temperature sequence is often referred to as "high–low–high." In the interest of process time minimization, all or part of the required temperature steps may be combined with various regular processing steps so that an optimized high–low–high sequence may not exist. The approximate temperature ranges over which these various effects occur are shown in Fig. 8.42.

As will be discussed in section 8.7.4 on wafer warpage, interstitial oxygen improves silicon yield strength, while oxygen precipitates reduce it and sometimes lead to wafer slip during heatup and cooldown cycles. Thus, trade-offs between gettering and strength may be required in order to obtain maximum yield.

Under ordinary circumstances, an as-grown crystal is supersaturated with oxygen. It can be seen by noting that typical oxygen concentration specifications range from 25–40 ppma[16] and then com-

FIGURE 8.42

Behavior of oxygen in silicon as a function of processing temperature. (The temperature boundaries are fuzzy and shift with oxygen concentration and the number and kind of grown-in, or native, defects.)


Oxygen precipitates will dissolve, and interstitial oxygen will out-diffuse from the surface.
Electrically active donor SiO_x complexes form.
Low-temperature donors disappear; SiO_x clusters grow in size; small SiO_2 precipitates form.
Precipitates above the critical size will grow.
400
500
600
700
800
900
1000
1100
1200
1300
Temperature (° C)

[16] Since there are 5 × 10^{22} atoms/cc of silicon, 25–40 ppma converts to 1.25–2 × 10^{18} atoms/cc.

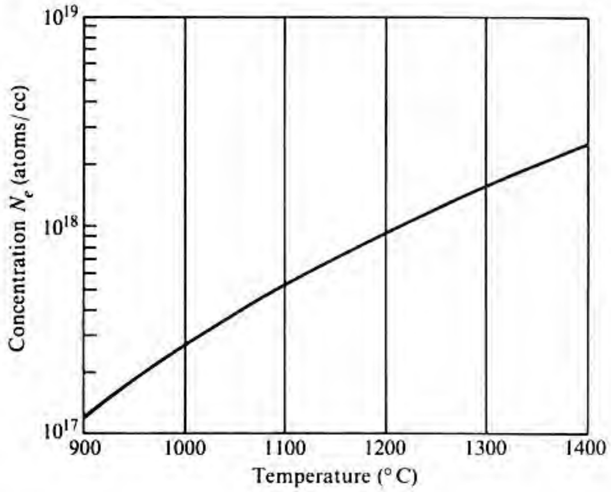


FIGURE 8.43

Oxygen interstitial solubility in silicon versus temperature. (*Source:* From data in R.A. Craven, in H.R. Huff and R.J. Kriegler, eds., *Semiconductor Silicon/81*, p. 254, Electrochemical Society, Pennington, N.J., 1981.)

paring those numbers with Fig. 8.43, which gives the equilibrium concentration N_e versus temperature (150). When cooled to room temperature, the oxygen is immobile and will remain dispersed so that the silicon is supersaturated. However, during high-temperature processing, part of the oxygen can diffuse out of the wafer and into the ambient, and part of that remaining can precipitate. If it is assumed that there is no barrier to oxygen leaving the wafer surface, the oxygen profile $N(x,t)$ is given by (151)

$$N = N_e + (N_0 - N_e)\text{erf}\left(\frac{x}{2\sqrt{Dt}}\right) \qquad 8.92$$

where N_0 is the initial concentration of oxygen in the wafer[17] and N_e is the equilibrium concentration at the heat-treat temperature. One reported value of the diffusion coefficient D is (152)

$$D(T) = 0.17e^{-2.54/kT} = 0.17e^{-29449/T} \quad \text{cm}^2\text{/s} \qquad 8.93$$

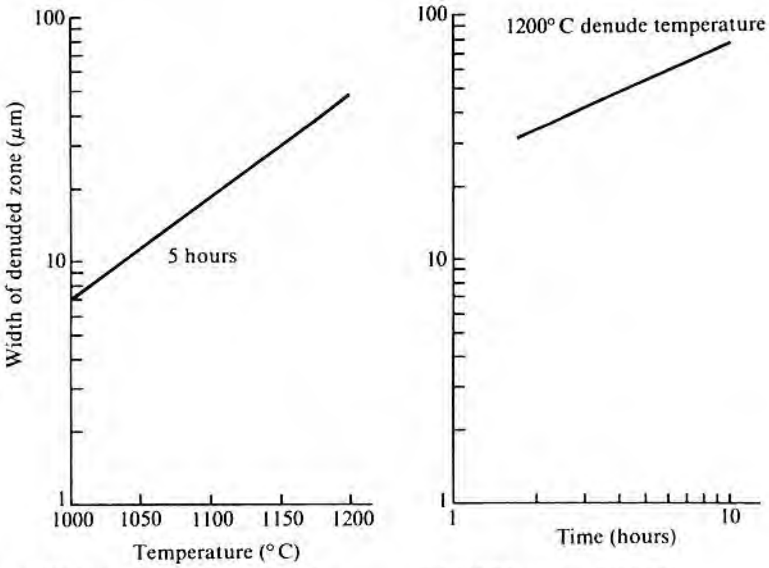
where T is the diffusion temperature in K.

Before Eq. 8.92 can be used to determine the width of a denuded zone, the concentration below which no precipitation will occur at the processing temperature must be determined. Unfortunately, determining this concentration is not easy since the value depends not only on oxygen supersaturation but also on the amount and kind of grown-in defects, the amount of impurities such as carbon that are present, and the prior heat-treat history. For a given wafer, however, the width should vary as the square root of denude time and exponentially with temperature. Typical plots are shown in Fig. 8.44.

The rate of nucleation depends in a complex way on interstitial

FIGURE 8.44

General behavior of oxygen-denuded region width.



(a) Width versus temperature for a fixed time

(b) Width versus time for a fixed temperature

oxygen concentration, temperature, oxygen diffusivity, dissolution enthalpy, and precipitate interfacial energy (153). A plot of nucleation rate versus anneal temperature for a given oxygen concentration is shown in Fig. 8.45.

In the intermediate temperature range, between low-temperature nucleation and high-temperature dissolution, the nuclei that are above a critical size will grow comparatively rapidly, while the smaller ones will dissolve. Fig. 8.46 shows the trend of critical size versus temperature for two different oxygen levels. Experimentally, it is observed that no optically observable precipitates form for $N_0 < \sim 6 \times 10^{17}$ atoms/cc (154). Apparently, below this concentration, the precipitate nuclei do not reach a size sufficiently large to continue growth when the temperature is increased to the point where

[17] The interstitial oxygen concentration can be measured nondestructively by IR absorption in the 1106 cm^{-1} band. (See, for example, B. Pajot, "Characterization of Oxygen in Silicon by Infrared Absorption: A Review," *Analusis* 5, p. 293, 1977.) Precipitated oxygen absorbs weakly in a different band, so the amount of precipitation can be inferred by the change in interstitial absorption. However, after the first diffusion, the IR absorption due to free carriers becomes so high that no further measurements can be made. By stopping the spectrometer aperture down to a few millimeters, the oxygen profile across starting slices and heat-treated wafers can be determined. Either selective etching combined with optical microscopy or X-ray topography can be used to observe precipitates directly.

FIGURE 8.45

Nucleation rate versus anneal temperature. (*Source:* Adapted from N. Inoue et al., in H.R. Huff and R.J. Kriegler, eds., *Semiconductor Silicon/81*, p. 282, Electrochemical Society, Pennington, N.J., 1981.)

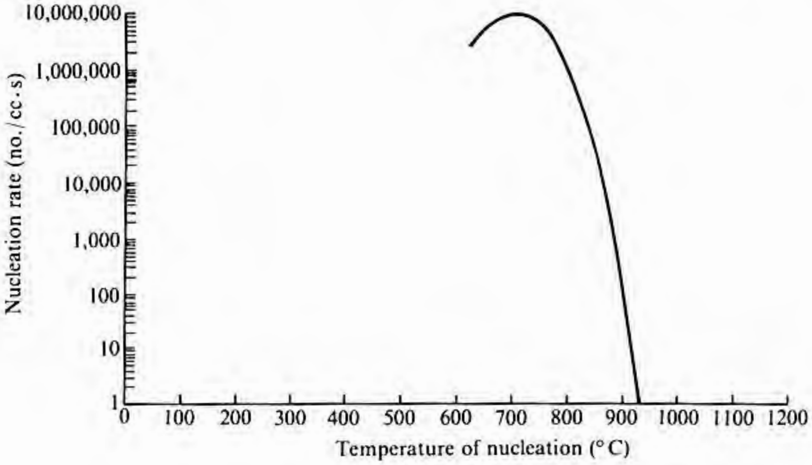
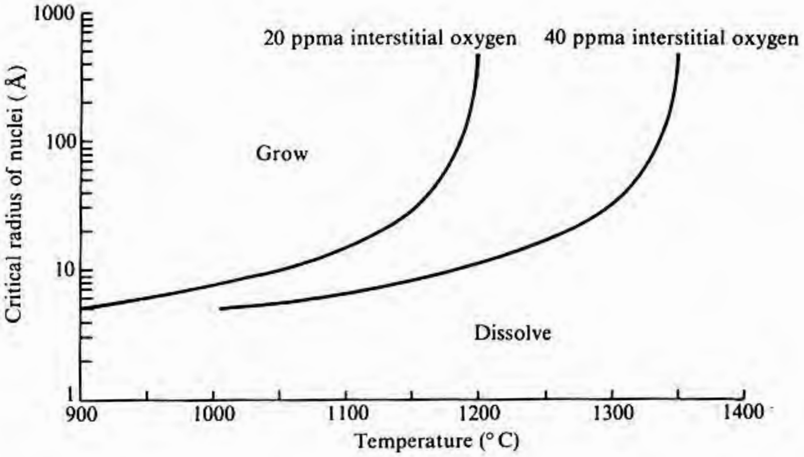


FIGURE 8.46

Effect of temperature on critical radius of oxygen precipitate nuclei. (*Source:* Adapted from R.A. Craven, in H.R. Huff and R.J. Kriegler, eds., *Semiconductor Silicon/81*, p. 254, Electrochemical Society, Pennington, N.J., 1981.)

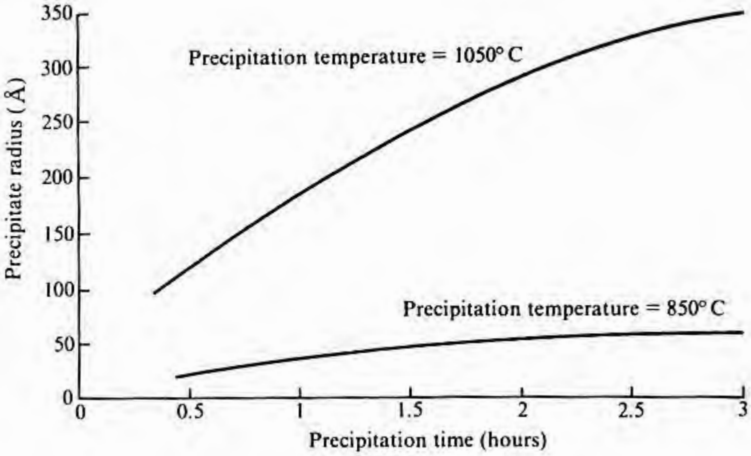


oxygen diffusion is large enough to allow growth. Even if there is sufficient oxygen, if the nucleation cycle time is insufficient for critically sized nuclei to form, or if the temperature is too high to nucleate at all, no precipitate growth will occur. When there are critically sized nuclei, the amount of oxygen precipitated will increase with the number of nuclei, the growth time and temperature, and the oxygen concentration. Fig. 8.47 shows how the precipitate size is thought to increase with time for a fixed temperature.

Because of the complex interactions between the variables involved in intrinsic gettering, it is difficult to experimentally devise an optimum process. In attempts to solve this problem, various computer modeling programs have been devised (155, 156). These

FIGURE 8.47

Calculated oxygen precipitate growth with time for two temperatures. (*Source:* Adapted from B. Rogers et al., in K.E. Bean and G.A. Rozgonyi, eds., *VLSI Science and Technology/84*, p. 74, Electrochemical Society, Pennington, N.J., 1984.)



programs require the starting material's oxygen content and proposed heat cycles as inputs and give the amount of precipitated oxygen as an output. It should be noted that if processing temperatures are reduced to about 800°C, it appears that there will be insuffficent precipitate formation for intrinsic oxygen gettering to be viable.

8.7 DIFFUSION-INDUCED DEFECTS

During the diffusion steps required to produce the desired impurity profile, a number of diffusion-related defects often appear. Some of these defects are listed in Table 8.7.

Chipped edges are likely to occur whenever wafers are handled, but during most processing, the wafers are either lying flat or else are sitting in plastic holders. During diffusion (and oxidation), the wafers are held in slots or grooves in carriers made of hard materials such as fused silica, silicon, or silicon carbide and are thus much more likely to be damaged without additional handling care. Further, the diffusion cycle sometimes produces a dopant-rich glass that flows at diffusion temperatures and collects where the wafers contact the carrier. The wafers may then stick to the carrier after cooldown and be chipped as they are removed from the carrier. Reducing the concentration of dopant reactant in the gas stream will generally prevent the formation of excessive glass.

Many of the gaseous doping compounds used for silicon will react with a clean silicon surface at high temperatures and cause pitting. To prevent pitting, an oxidizing ambient is generally used, and a thin layer of oxide is grown before the doping gases are admitted to the diffusion furnace tube.

TABLE 8.7

Diffusion Process Induced Defects

Item	*Comments*
Chipped wafer edges	Boat slots too small or source buildup on boat causing wafer sticking.
Wafer surface pitting	Flow of Cl- or Br-containing gas across hot wafer before protective oxide formed.
Localized unwanted diffusions	From initial pinholes in masking oxide or from oxide failure during diffusion.
Diffusion pipes	Filamentary diffusion paths connecting active elements (see Chapter 11 for a discussion of their effects).
Lifetime-killing impurities	May be brought in on wafer surface or be transported to hot wafer from contaminated diffusion boat or tube.
Precipitates	Due to choice of cooling cycle combined with an excess of impurity.
Strain-induced dislocations	From excessive concentration of diffusant.
Slip	From thermally induced stresses exceeding yield point of material.
Wafer warpage	Due to massive slip from stresses introduced during diffusion cycle.

8.7.1 Diffusion Flaws and Pipes

Either pinholes already in the masking oxide due to lithography problems or local failure of the mask during diffusion will cause small regions of unwanted impurities. In addition, large-scale mask failure could result from an incorrect choice of oxide thickness. The two most likely causes of local area flaws are phase separation (oxide flowers), which sometimes occurs in spin-on sources, and particulates containing a high concentration of dopant alighting on the wafer surface. Particulates high in dopant concentration may be brought in from the outside[18] or form in the diffusion tube. Phosphorus diffusions are particularly prone to generating products in the tube that cause oxide failures. The end-product of most phosphorus sources is P_2O_5, which tends to coat not only the wafers but the tube as well. The coating forms most heavily on cooler portions of the tube (near the end). Moisture from the air will interact with it while the wafers are being removed and gradually form a syrupy layer that

[18] According to legend, one of the earliest identified sources of unwanted phosphorus diffusions was phosphorus-rich lawn fertilizer tracked into the diffusion room.

can drop off onto the next set of wafers being loaded and subsequently dissolve the masking oxide during the diffusion cycle. Thin spots in the initial oxide due to surface contamination are also potential failure spots. For the case of phosphorus, $POCl_3$ is reported to produce a higher surface concentration and more defects than PBr_3 (157).

When a surface particle contains a higher concentration of dopant than the diffusion provides over the rest of the surface, the diffusion front[19] will be deeper under the particle and thus produce a diffusion "pipe" extending past the rest of the diffusion. If the diffusion covers a crystallographic defect such as a grain boundary, enhanced diffusion will occur along the boundary, as was discussed in section 8.3.12, and will also produce a diffusion pipe extending out from the main diffusion front. (The effect of pipes on devices is discussed in Chapter 11.)

8.7.2 Lifetime-Reducing Impurities and Precipitates

Contamination of the furnace by lifetime-reducing impurities (sometimes referred to as heavy metals or as transition metals) and methods of minimizing it were discussed in Chapter 3. Also discussed were various cleanup procedures designed to ensure that a clean wafer surface enters the diffusion or oxidation tube. Unfortunately, such extensive precautions do not always provide the required low level of contamination. Thus, various gettering steps are usually used during high-temperature processing. These steps were discussed in detail in the previous section. In some cases, however, the deliberate use of lifetime "killers" is necessary. The lifetime-reducing impurities usually have a much higher solid solubility at diffusion temperature than at room temperature so that if their concentration is high at diffusion temperature, either from intentional or unintentional doping, precipitation may occur during cooldown. Precipitates that are in space charge regions will cause excessive leakage (see Chapter 11). Precipitates near the surface can cause a thinning of thermal oxide grown over that region, and long conductive precipitates may even short elements together. Precipitate prevention is done either by having previously reduced the concentration below the precipitation level or by cooling the wafer so rapidly that there is not enough time for the impurities to diffuse to precipitate nuclei and cause growth. This method is used, for example, with the gold doping of a TTL circuit. Unfortunately, however, too rapid a cooldown can cause slip in the wafers, as will be discussed in section 8.7.4.

[19] Some chosen isoconcentration line is often referred to as a "front."

8.7.3 Strain-Induced Dislocations

Since dopant atoms are almost never close in size to the atoms they displace, in high concentrations, they can produce enough strain to generate misfit dislocations (158–161). If Vegard's law is followed,[20] the volume strain β is given by $(V_i - V)/V$, where V_i is the new volume after introduction of impurities. This can be rewritten as

$$\beta = \frac{\Delta V}{V} = f(\Gamma^3 - 1) \tag{8.94}$$

where ΔV is the change in volume, V is the original volume, and f is the atomic fraction of impurity present (N/N'), with N as the concentration of impurity atoms/cc and with N' as the number of host atoms/cc. Γ, the misfit ratio, is the ratio of the impurity covalent radius to that of the host. The linear strain ε is given by[21] the expression $[(V + \Delta V)/V]^{1/3} - 1$, or

$$\varepsilon = \frac{\Delta \ell}{\ell} = [1 + f(\Gamma^3 - 1)]^{1/3} - 1 \tag{8.95}$$

where ℓ is a linear dimension. Eq. 8.95 can be approximated by $f(\Gamma^3 - 1)/3 = \beta/3$.

If the new atom is smaller than the host, then the lattice will contract and ε will be negative. If the atom is larger, then the lattice will expand. When an impurity atom enters the lattice interstitially rather than substitutionally, the lattice will expand regardless of the relative radii, and Eq. 8.95 is not applicable. Table 8.8 lists some values of covalent radii and misfit ratios. Fig. 8.48 shows experimental strain data for boron, antimony, and phosphorus in silicon and compares them with the predictions of Eq. 8.95. It should be noted that as the strain increases, the bandgap will also change, causing n_i to change and, in turn, affecting D for cases where $N > n_i$ (162). Since the strain is in a thin layer on one surface, it will cause the wafer to bow, and indeed one method of estimating the amount of misfit-induced strain is to measure the amount of bowing on wafers whose elastic limits have not been exceeded. Bowing will not occur if a balanced diffusion occurs on the opposite side of the wafer, but most processes keep the back side protected by a thick oxide during diffusion.

TABLE 8.8

Covalent Radii

Element	*Radius (Å)*	*Misfit Ratio*
C	0.77	0.66
B	0.88	0.75
P	1.10	0.94
Si	**1.17**	**1.00**
As	1.18	1.01
Ge	1.22	1.04
Ga	1.26	1.08
Al	1.26	1.08
Sb	1.36	1.16
Sn	1.40	1.23

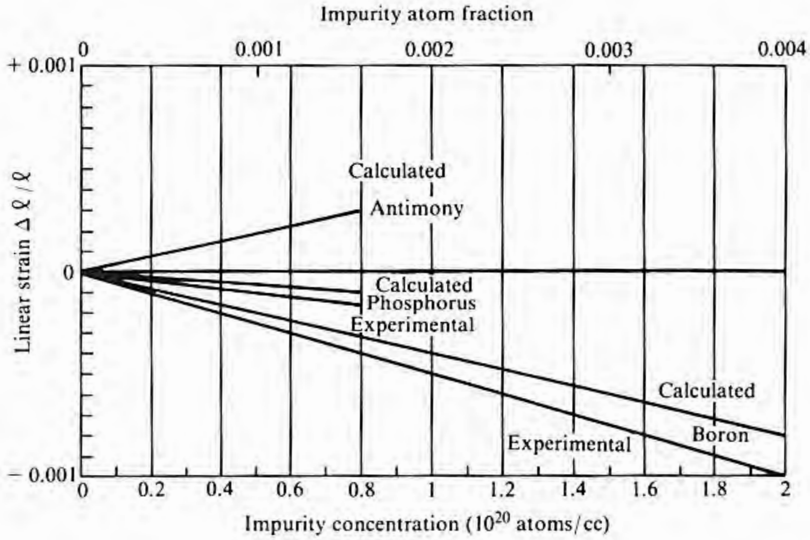
Source: Linus Pauling, *The Nature of the Chemical Bond*, Table 7–13, Cornell University Press, Ithaca, New York, 1960.

[20] Originally proposed for ionic crystals, Vegard's law is almost never followed for metallic solutions but, for at least some dopants in silicon, appears to apply reasonably well (see Fig. 8.48).

[21] The symbols σ and ε or T and S are commonly used for stress and strain, respectively. In this discussion, σ and ε will be used even though a conflict exists with earlier symbol definitions.

FIGURE 8.48

Linear strain versus amount of impurity. (*Source:* Adapted from G. Celotti et al., *J. Mat. Science 9*, p. 821, 1974; and F.H. Horn, *Phys. Rev. 97*, p. 1521, 1955.)



Analyses of the stresses involved as a function of depth and impurity profile are available (159), with the maximum stress σ_s occurring at the surface just as diffusion starts and given by

$$\sigma = \frac{Y\varepsilon}{1 - \nu} \qquad 8.96$$

or

$$\sigma = \frac{\beta Y}{3(1 - \nu)} \qquad 8.97$$

where Y is Young's modulus and ν is Poisson's ratio. The fraction of impurity at the surface and the f to be used in calculating β for Eq. 8.97, are given by $f = N_0/N'$ where N_0 is the surface concentration. Because of the stress relief afforded by dislocation production, the highest density of dislocations generally occurs not at the surface but somewhat below it.

EXAMPLE □ Estimate the value of σ, and decide whether or not dislocations will be induced from a 1100°C boron diffusion with a surface concentration of 10^{19} atoms/cc.

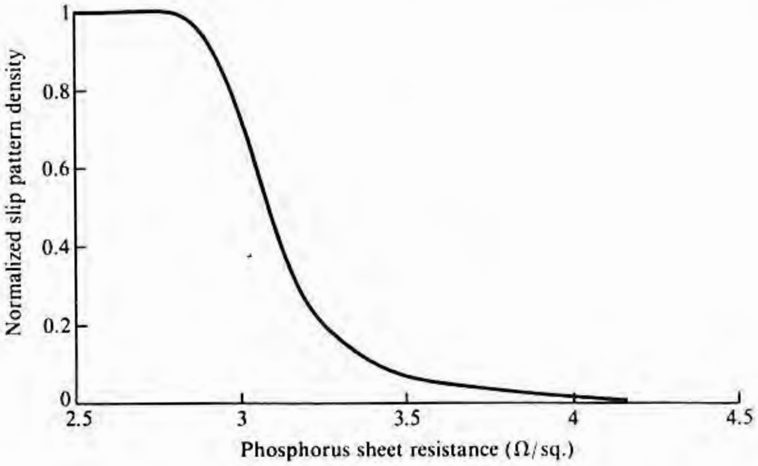
From Fig. 8.48, the strain induced by a boron diffusion with a concentration of 10^{19} atoms/cc is about 5×10^{-5}. Young's modulus and Poisson's ratio are both orientation and temperature dependent. As typical values, use $Y = 10^{12}$ dynes/cm^2 and $\nu = 0.3$. Thus, the stress σ is approximately 7×10^7 dynes/cm^2. Fig. 8.54, shown in the next section, gives the stress required for dislocation generation to become noticeable in silicon, and from it, one concludes that dislocations will probably not be generated. However, if a surface concentration of 10^{20} is used, ε becomes 0.0005, T increases to 7 × 10^8 dynes/cm^2, and from Fig. 8.54, at 1100°C, generation dislocation is very likely. □

High-concentration phosphorus diffusions, such as might be used for bipolar emitters, can cause widespread slip to occur in wafers. Its occurrence is apparently associated with the onset of precipitates (163). Data are shown in Fig. 8.49 for a $POCl_3$ source diffusion into (111) silicon at 1100°C. Background doping is 2 Ω-cm boron, and the junction depth is approximately 2 μm. Experimentally, it is observed that diffusion pipe losses increase drastically when there is heavy dislocation damage in the emitter. Hence, from the data of Fig. 8.49, the phosphorus emitter diffusion sheet resistance for the described set of diffusion conditions should be kept above about 4 Ω/sq. Since higher concentrations of phosphorus afford better gettering and usually give better emitter efficiency, the process engineer must balance these gains against losses from increased leakage due to the presence of excessive dislocations.

The strain introduced by a small atom can be compensated by simultaneously adding a larger one. Thus, germanium or tin, which would not be expected to appreciably affect the electrical properties of silicon, can be used to eliminate strain during a boron or phosphorus diffusion (164–167). Such a procedure will not, however, solve the problem of precipitates forming and producing additional mechanical damage. Arsenic atoms are closely matched to silicon

FIGURE 8.49

Slip density versus phosphorus diffusion sheet resistance for an 1100°C diffusion.
(*Source:* Adapted from R.A. McDonald et al., *Solid-State Electronics* 9, p. 807, 1966.)



and hence produce few misfit dislocations, but like the others, they do form precipitates at high doping levels.

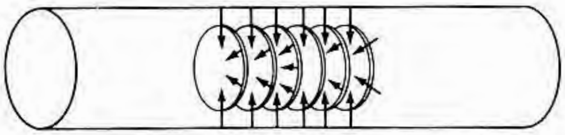
8.7.4 Permanent Wafer Warpage

The most common cause of wafer warpage is excessive thermal stresses occurring during wafer insertion and removal from diffusion or oxidation tube hot zones. The high thermal stresses are often enough to produce plastic flow and thus permanent deformation[22] of the wafer (168–172). These thermal stresses arise primarily because as the wafers are inserted or removed from the hot furnace, heating or cooling must be produced from the outside in as shown in Fig. 8.50. This effect can produce a transient temperature differential between the center and edge of the wafer of up to a few hundred degrees. In addition, the thermal mass of the boat[23] will prevent it from changing temperature as rapidly as the wafers and will thus keep the portion of a wafer in contact with it from heating or cooling as fast as the rest of the wafer. Fig. 8.51 shows some experimental data for 100 mm diameter wafers, taken radially along a direction not influenced by boat contact. The temperature depends on the wafer diameter and thickness, the spacing between wafers, and the boat design. The slower the insertion or withdrawal rate, the less the temperature differential, so that by decreasing the rate, ΔT can be reduced to any desired level. Since it is difficult to reproducibly withdraw wafers manually, programmable "push–pull" mechanisms are used. As can be seen from Fig. 8.51, lower furnace temperatures produce less temperature difference; therefore, as an alternative or sometimes as an adjunct, the furnace temperature can be gradually reduced or increased (ramped).

Aspects other than thermally induced damage may need to be

FIGURE 8.50

Heating from periphery because of external cylindrical heat source.



[22] This defect is separate from wafer bowing caused by mechanical surface damage, strain from low-concentration diffusions, or stresses from interconnect layers. Those defects seldom exceed the elastic limits of the silicon, so if they are removed, the silicon returns to its original shape.

[23] For a description of wafer carrier boats, see section 8.8.

considered in choosing push–pull or ramp rates, and the final choice must be based on the total effect. Slow withdrawal will allow precipitation if there is a high concentration of an impurity such as gold, which has a high diffusion coefficient and a low room-temperature solubility. When the gold is needed to kill lifetime, a slow withdrawal may be harmful in that it will allow the gold to precipitate rather than remain in the lattice interstitially. Even if a large amount of such an impurity is not present, a slow cool may allow it to diffuse to crystallographic defect sites, decorate them, and cause device failures. The processing window becomes progressively narrower as the wafer diameter increases, and it becomes quite difficult to choose a withdrawal rate for 150 mm wafers that will prevent thermally induced slip and yet not allow small haze-like precipitates to form on the surface. The electrical characteristics of a silicon–oxide interface depend on the kind of ambient used during the cooling cycle as well as on the rate. When a push–pull system is used, maintaining the desired ambient may be difficult, so the combination of a ramp-down to a low enough temperature to minimize slip, followed by a rapid withdrawal, may be the best compromise.

In order to estimate the temperature differential that can be tolerated, the strain due to the thermal expansion can be converted into mechanical stress, and the stress can be compared with the stress required for plastic flow at the temperature in question. With a radially symmetric temperature gradient from the center to the edge of a wafer, both radial and tangential stress will develop. The tangential stress will change sign in going from center to edge and will have a substantially larger absolute value at the edge than at the center. The radial component, equal to the tangential component in the center of the wafer, decreases to zero at the edge (169–172). An example of calculated stress, using the radial temperature distribution of Fig. 8.51, is shown in Fig. 8.52 (172). Since the maximum stress is at the periphery, that is where slip would be expected to begin. However, if the yield point of the silicon is less near the center than near the edge, then slip could occur elsewhere.

To evaluate the likelihood of plastic flow (slip), the applied stress must be resolved into shear stress lying in the slip plane and directed in the slip direction. For silicon, slip predominantly occurs between (111) planes in a [110] direction. The resolved shear stress is given by

$$\sigma_{\text{shear}} = \sigma_0 \cos \alpha \cos \beta \qquad 8.98$$

where σ_0 is the stress being resolved, α is the angle between σ_0 and the normal to the slip plane, and β is the angle between σ_0 and the slip direction.

EXAMPLE □ When σ_0 lies in the (100) plane with a [012] direction, what is the shear stress in the (111) plane in the [110] shear direction?

Using the expression

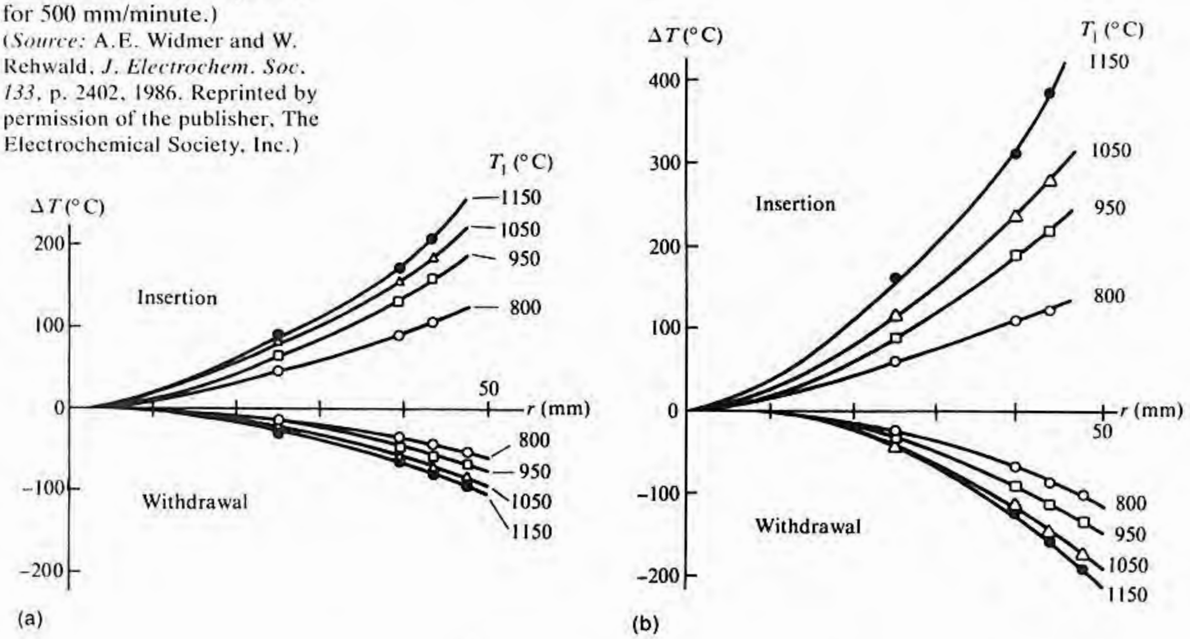
$$\cos\theta = \frac{hh' + kk' + \ell'}{\sqrt{(h^2 + k^2 + \ell^2)(h'^2 + k'^2 + \ell'^2)}}$$

where $hk\ell$ and $h'k'\ell'$ are the indices of any two planes, let $hk\ell$ = 012 and $h'k'\ell'$ = 111 to solve for cos α and $hk\ell$ = 012 and $h'k''$ = 110 to solve for cos β. Note that in the cubic system, the direction normal to a ($hk\ell$) plane is [$hk\ell$]. Alternatively, spherical coordinates of each of the direction vectors could be calculated with respect to the wafer surface (168). This allows for an easier choice of azimuthal angle of the applied stress. □

If the resolved shear stress is greater than the critical shear stress for the wafer material, then slip will occur. The radial dependence of the resolved stress (Eq. 8.98) is such that for (111) oriented silicon, 12 stress maxima occur. It is at those points on the periphery that slip is normally first seen, although, often, every other one will be much more pronounced and thus give the familiar six-point star pattern. On (100) silicon, there are 4 maxima. Fig. 8.53 shows a typical (111) slip pattern after being delineated by etching.

FIGURE 8.51

Temperature differential between center and various points along radius r of 100 mm wafer while load of wafers is inserted and withdrawn from hot diffusion tube. (T_1 is the temperature of the tube. The curves in part a are for insertion/removal rates of 100 mm/minute; the curves in part b, for 500 mm/minute.) (*Source:* A.E. Widmer and W. Rehwald, *J. Electrochem. Soc. 133*, p. 2402, 1986. Reprinted by permission of the publisher, The Electrochemical Society, Inc.)



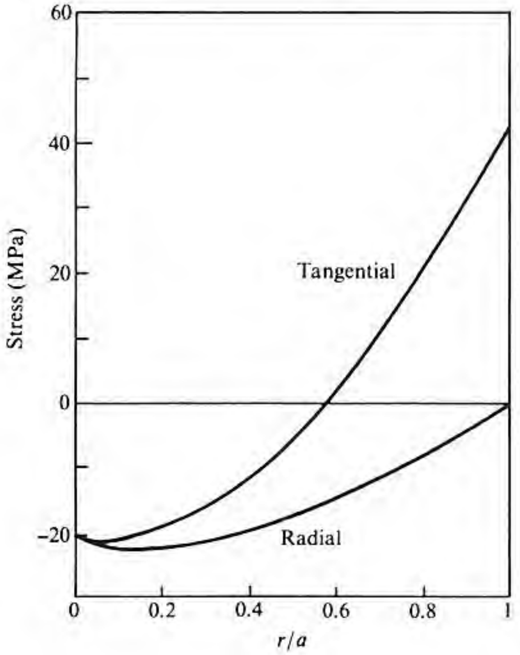
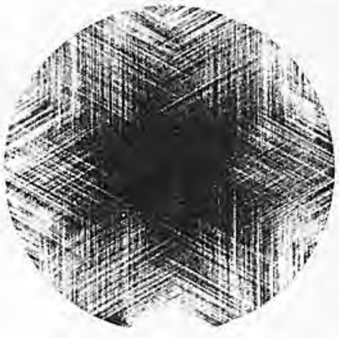


FIGURE 8.52

Calculated radial and tangential stresses induced in 100 mm silicon wafer when withdrawn from 1050°C furnace at rate of 50 cm/minute. (*Source:* A.E. Widmer and W. Rehwald, *J. Electrochem. Soc. 133*, p. 2403, 1986. Reprinted by permission of the publisher, The Electrochemical Society, Inc.)

FIGURE 8.53

A (111) oriented silicon wafer with severe slip introduced by a large thermal gradient. (The slip has been delineated by etching to show the emergence of dislocations.)



The value of the critical shear stress for a given material is quite variable and is still well described by a statement in A.H. Cottrell's book *Dislocations and Plastic Flow in Crystals*[24] written about 40 years ago: "Precise values for individual crystals have little significance apart from their order of magnitude because the critical stress not only depends upon the temperature and speed of straining but is also very structure sensitive. Variations in the purity of the material, the conditions of growth and the state of the surface of a crystal all greatly affect its shear strength." Higher temperatures reduce silicon yield strength in the general manner shown in Fig. 8.54 (173–175). Increasing the strain rate somewhat increases yield strength at a given temperature, thus making values dependent on the strain rate used by the measuring instrument. Interstitial oxygen increases silicon strength, while precipitated oxygen decreases it (176–180). Fig. 8.55 shows the general behavior. It should be noted that while the higher concentrations of interstitial oxygen enhance strength, they also have the potential to precipitate more during oxidation and diffusion cycles and thus reduce the strength below the level observed with no oxygen. Nitrogen, in concentrations of about 10^{15}

[24] Oxford University Press, London, 1953.

FIGURE 8.54

Estimated high-temperature yield stress for silicon.

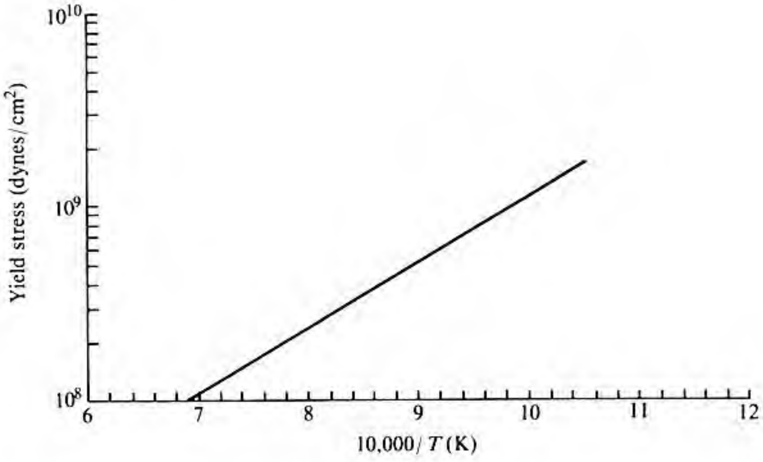
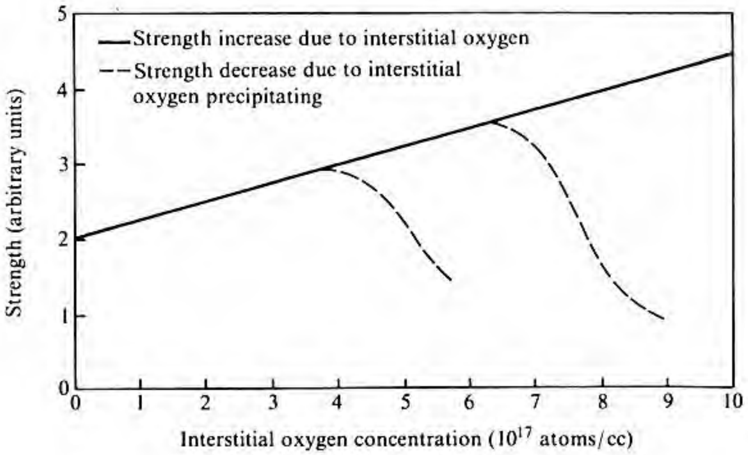


FIGURE 8.55

Effect of oxygen on wafer strength.

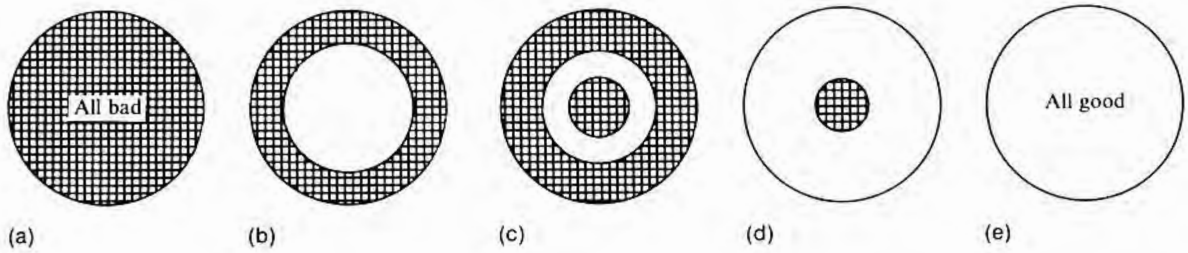


atoms/cc, and apparently interstitial, will increase yield strength and seems less likely to precipitate than oxygen (181, 182).

The wafer photographed for Fig. 8.53 was intentionally subjected to very severe thermal stresses in order to produce an exaggerated effect. In good IC processing, no discernible slip occurs. Marginal processing, however, produces a number of distinctive slip region patterns. Fig. 8.56 shows the shape of these patterns. They range from slip over the whole wafer (very rapid removal) to no slip. Reject chips very closely follow the pattern of slip, and the relation between a reject map and the slip pattern was used in early experiments to establish the correlation between slip and yield loss. The

FIGURE 8.56

Slip patterns observed in processed wafers. (The cross-hatching denotes area of high slip.)



pattern in part b occurs when the thermal differential is marginal. The ring of unslipped material in the pattern in part c apparently occurs from a combination of boron and oxygen precipitation in the center of the wafer, which reduces the yield point. The pattern in part d is on a wafer that did not have enough temperature differential to cause peripheral slip but that had enough oxygen precipitation in the middle to cause a reduction in yield strength and subsequent slip. Such central precipitation is usually associated with oxygen coring, in which a much higher than normal concentration of oxygen is incorporated into the central region of the crystal during growth.

8.8 DIFFUSION EQUIPMENT

Diffusions are carried out in diffusion furnaces that are comprised of a fused silica tube surrounded by a heater element. Temperatures are in the 900°C–1300°C range, with the trend being to lower temperatures. Historically, the tubes have been horizontal with the wafers held vertically. However, vertical-tube furnaces with the wafers held horizontally are also available. The diameter of the tube is a few centimeters larger in diameter than the wafers to be processed. Table 8.9 shows typical values. The length of the flat part of the furnace hot zone is about 100 cm and is heated by elements broken into at least three segments so that the two end zones can operate at higher power to compensate for heat losses out the ends of the tube. Most of the temperature controllers use microprocessors with appropriate control algorithms to match the thermal properties of the furnace. Some trade-offs must be made in the thermal design. Too little insulation will require excessive power, while too much insulation will cause cooldown time to be excessive. Attached to one end of the furnace proper is a gas cabinet to house controls for the gases (source cabinet).

TABLE 8.9

Diffusion Tube Diameters

Wafer Diameter (mm)	*Tube Diameter (mm)*
100	170/176*
125	184/190
150	215/224 or 225/235
200	250/275

*Inside diameter/outside diameter (ID/OD).

Source: Data courtesy of Thermco Systems, Inc.

8.8.1 Gas Flow Measurement and Control

In older plumbing systems and in laboratories, the rotameter is commonly used to measure gas flow rates. It is a simple and virtually foolproof instrument but is not very amenable to remote reading. It consists of a vertical glass tube with a tapered bore that linearly increases in diameter from bottom to top. Inside is a round ball that is free to move up and down and that fits rather tightly at the bottom of the tube. As the gas flow rate through the tube increases, the ball (float) will move up the tube, and its vertical position is a measure of flow.

Electronic mass flow meters depend on the fact that the temperature differential up and downstream from a heat exchanger supplying a constant amount of heat per unit time to the gas stream is proportional to the mass of gas per unit time flowing through the exchanger. The temperatures are sensed by resistance thermometers in a bridge network, and the difference is expressed in terms of a DC voltage. After multiplying by constants appropriate for the gas being measured, the voltage swing corresponding to zero to full flow is usually 0–5 V. This voltage can be read directly on a voltmeter, used to control the gas flow, and/or fed into an A/D converter for a computer input.

The gas flow can be varied by simple, manually operated needle valves, which are usually the valves used in conjunction with rotameter flow measurement. However, for automatic control, more sophisticated valves are required. They are still based on the needle valve concept but use piezoelectric elements or electrically heated thermal expansion elements to move the needle in and out of its seat. Combining the output of a mass flow meter with the input to an electrically operated valve provides an automatic mass flow controller that is widely used to control diffusion gas flows. On/off control is either by manual cutoffs or by air- or solenoid-operated valves. Solenoid valves do not supply as much closure force as those that are air operated. Thus, more problems with leakage may occur when they are used. When a valve is chosen, care must be taken to ensure that the valve seal is compatible with the gas being controlled.

8.8.2 Diffusion Tubes

Historically, the diffusion tube (the inside tube in a furnace, and the one containing the wafers and carrying the process gases) has been made of clear fused quartz.[25] Sometimes, between the diffusion tube and the furnace winding is a thick-walled tube, or liner, which is generally made of opaque fused SiO_2. The liner offers protection for the more expensive inner tube, provides additional thermal mass, and minimizes the propensity of the inner tube to sag at temperatures above about 1150°C. Problems with the fused quartz that have led to a search for alternative materials are the high-temperature sagging, the gradual devitrification with use at high temperature, and the relative ease with which some impurities can penetrate the quartz. Vapor-deposited polycrystalline silicon and sintered silicon carbide are two alternative materials that are partially replacing fused quartz. Both are now available in high purity, sag less at elevated temperature, and do not devitrify. They are, however, more expensive. Since both Si and SiC will react with metals at diffusion temperature, care must be taken to ensure that the heating coils do not become distorted and touch either of them; otherwise, alloying will take place, and both the heater and the tube will be ruined. Si and SiC are quite conductive at diffusion temperatures; thus, even if no reaction occurs, arcing between turns will occur if the heater and tube touch. In the case of SiC, this possibility can be minimized by using a tube coated with a nonconductor such as zirconia.

The inside walls of fused quartz tubes will react enough with the doping atmosphere to become a source. Thus, to prevent cross contamination, different tubes should be reserved for each of the dopants being used. The buildup of dopant on the tube walls will act as an additional source and may affect the diffusion.[26] A few dummy runs may be required to saturate the tube before reproducible results are obtained. Therefore, a tube also may not provide reproducible diffusions if it is used at two widely differing temperatures.

8.8.3 Wafer Diffusion Boats

The materials used for tubes—that is, SiO_2, Si, and SiC—are also used for the boats. The same care must be used in keeping separate

[25] Some usage refers to synthetically prepared SiO_2 that has been melted and fabricated as fused silica and to naturally occurring quartz that has been melted and fabricated as fused quartz. The terms are also sometimes used interchangeably. There are some differences in physical properties, with fused naturally occurring quartz generally having a slightly higher viscosity than the synthetic material.

[26] Sometimes the buildup in a single run is enough to cause the wafer boats to stick to the tube.

boats for each process as was used for tubes. In addition, cross contamination can also occur during the boat cleaning operation if boats from separate diffusion steps are etched in the same cleaning solution. Diffusion boats are designed to impede flow as little as possible and to have minimal thermal inertia.

8.8.4 Wafer Insertion

The original manual method for placing wafers in a conventional horizontal-tube furnace was to load each wafer individually into a slotted fused silica boat by tweezers and then slide the boatload of wafers into the furnace with a push rod. Difficulties with this procedure are that it is slow, tweezers introduce lifetime-killing contaminants and mechanical damage, the sliding boat generates particles, and the insertion and withdrawal rates are ill-defined. The cleanup step that precedes diffusion is almost always done with a number of wafers (25 or less, depending on diameter) in a carrier. Usually, the carrier is designed to withstand the cleaning solutions and is plastic. If the spacing of wafers in the cleanup boat matches the spacing for diffusion boats, conventional "flip–transfer" (dump–transfer) can be used to load the diffusion boat and eliminate the tweezer step. When spacings of the two boats are different, a slice-by-slice unload and reload machine can be used, but the slice pickups used can introduce damage or contamination. To provide control over insertion and withdrawal rates, paddles of fused silica or silicon carbide that either slide or ride on rollers driven by a "boat pusher" were introduced to carry the diffusion boats into the furnace. Since such sliding or rolling produces many defect-generating particles, cantilever boat loaders were developed (183, 184). The boat is supported on the end of a long arm that moves the wafers and boat(s) into the furnace without touching the walls. The arm may then set down the boat and withdraw, or it may remain in the furnace for the duration of the diffusion.

Vertical-tube furnace loading varies somewhat from one manufacturer to the next, but wafers are generally automatically transferred from cleanup carriers to a special fused quartz carrier that is part of the furnace.

8.8.5 Computer Control of Tube Functions

Studies made before the introduction of computer-controlled furnaces showed that most yield loss came from human error and that the numerical losses were several yield points. Some of these errors are listed in Table 8.10. The industrywide desire to eliminate such errors became the driving force for computer control. Several levels of control can be exercised, as shown in Fig. 8.57. The first level, usually done by an on-board microprocessor, is used to control the time sequencing of the various operations, the gas flows, temperature, temperature ramping, and wafer insertion and withdrawal rates. The next level, also usually done by the local microprocessor, includes things like tube diagnostics and data collection. Some examples of tube diagnostics are given in Table 8.11. Data collection includes records of specified and actual flow rates and temperatures.

The remainder of the functions shown in Fig. 8.57 are usually done by a higher-level (host) computer. Without computer control, one of the main causes of misprocessed diffusion lots is either the skipping of a diffusion or giving a lot the same diffusion twice. Thus, a system that keeps track of a given lot's correct location and gives a warning when it is logged into an incorrect step is very useful and is a natural adjunct to a lot tracking system. If control is totally local, a number of recipes in memory can be called up by an operator as required when a lot is started into a particular tube. With a host computer, the specifications for each device type at each furnace operation can be kept on file so that when a particular lot number is entered, the proper settings are automatically made. Automatic checks are also available to ensure that the lot is put in a tube compatible with the lot requirements. Finally, as part of an overall line-balancing program, lots can be automatically scheduled to particular tubes in order to maximize utilization and minimize cycle time.

TABLE 8.10

Major Causes of Diffusion Process Associated Yield Loss

1. Lot to wrong diffusion tube.
2. Error in selection of sequencing.
3. Wrong temperature or flow set.
4. Push–pull rates in error.
5. Boat loaded into wrong part of heat zone.
6. Lot skipped cycle or double cycled.
7. Tubes and boats not cleaned on proper schedule.
8. Failure to find plumbing leaks or cracked tubes.
9. Liquid sources run dry during processing.

FIGURE 8.57

Levels of computer control of diffusion tubes. (The functions in the two inner circles can be handled locally; the outer two levels require a host, or supervisory, computer.)

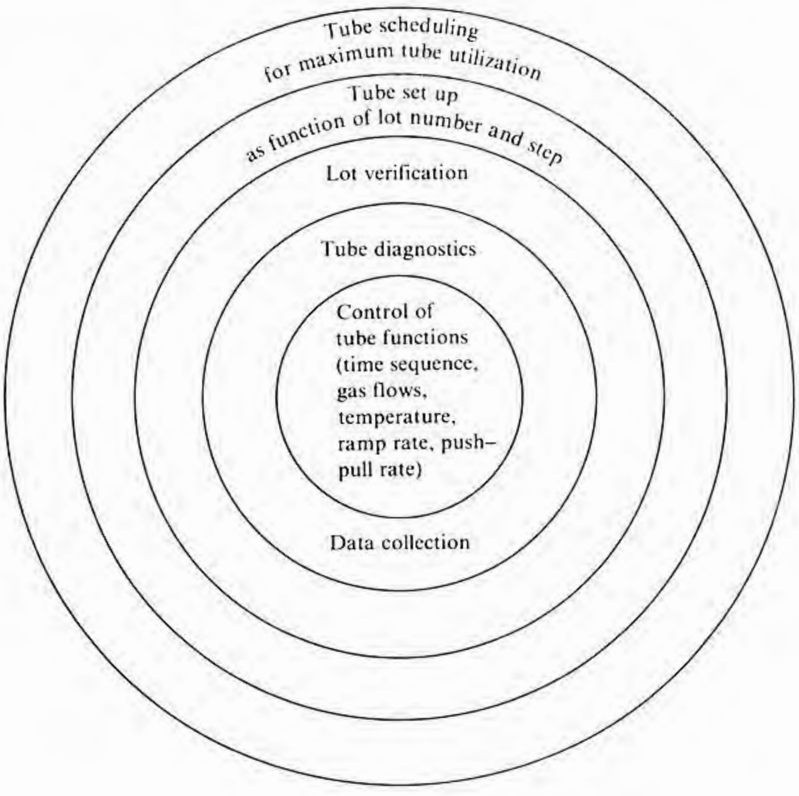


TABLE 8.11

Tube Diagnostics

Item	*Approach*
Plumbing leak	Close appropriate valves; measure residual gas flow with sensitive flow meter.
Furnace health	Check status of cooling water flow and temperature, air flow, line voltage, and so on.
Tube preconditioning	Check time since tube last used; cycle source if required.
Tube cleaning	Check time since last tube cleaning; give warning; prevent start-up if warning time exceeded.
Thermocouple drift	Will vary with control system used.
Mass flow meter clogging	Sequence valves to put two meters in series; compare readings.

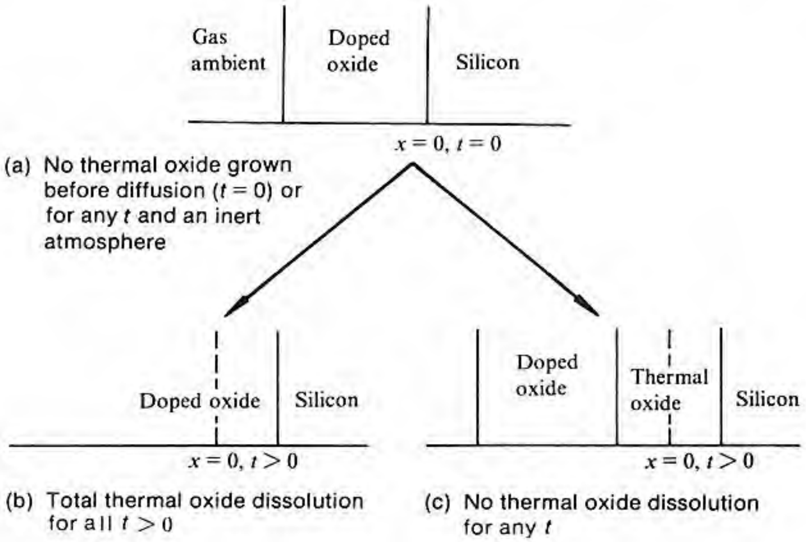
8.9 DIFFUSION SOURCES

Diffusion processes can be broadly categorized as open tube or closed tube. In the first catagory, the diffusion tube is operated at atmospheric pressure and a carrier gas flows through the tube during the diffusion operation. In closed-tube diffusion, the wafers and a solid that will provide impurity atoms for the diffusion are sealed in an evacuated ampule. Closed-tube diffusions are seldom used for silicon, but since they allow the vapor pressure to be controlled, they are often used for gallium arsenide.

In open-tube systems, the final source for silicon diffusions is generally a mixed oxide layer of silicon and the diffusant on the wafer surface. As shown in Fig. 8.58, as diffusion progresses, the oxide–silicon interface moves to the right. If the concentration of dopant in the mixed oxide is relatively low, the thermal oxide will remain between silicon and the mixed oxide (Fig. 8.58c). However, for high concentrations, the thermal oxide will be dissolved, and the doped oxide will contact the silicon surface (Fig. 8.58b). The doping layer may be derived from the reaction of a gaseous dopant species with oxygen and the thermal silicon oxide grown during diffusion or predeposited on the wafers before loading into the furnace. The gaseous dopant can be from an external gaseous source such as diborane (BH_3), from a vaporized liquid source such as BBr_3, or from a vaporizing solid in the furnace, such as an upstream boat of P_2O_5 or an adjacent wafer of boron nitride. Examples of predeposited sources are the separate low-temperature CVD deposition of silicon oxide with doping impurities incorporated into the oxide and the use of a liquid that when coated on the surface and allowed to dry leaves a residue of SiO_2 and dopant oxide. These sources are usually referred to as doped oxide and spin-on (sometimes paint-on) sources,

FIGURE 8.58

Behavior of doped oxide sources on surface of silicon during diffusion.



(a) No thermal oxide grown before diffusion ($t = 0$) or for any t and an inert atmosphere

(b) Total thermal oxide dissolution for all $t > 0$

(c) No thermal oxide dissolution for any t

respectively. Since the growth of an additional oxide on the wafer surface is not required with these sources, they can, in principle, be applied equally well to either silicon or gallium arsenide. Indeed, the doped oxide concept was first applied to gallium arsenide.[27]

The choice of the kind of source depends somewhat on personal preference, but important factors are the availability and purity of source material, the ease of use, the doping uniformity, the amount of damage introduced, and the surface concentration control. $POCl_3$ and BBr_3 are reasonably satisfactory liquid sources and are widely used. Liquid As and Sb sources have generally not worked well, and doped oxide sources are more common. For cases where maximum concentration is not desired, doped oxide sources are more applicable. Solid wafer sources are probably the most convenient to use but are not available for all dopants.

The source cabinet alluded to earlier that is part of the furnace console usually contains controls for all of the gases used in the furnace. In addition, if the source of the dopant is a liquid that must be vaporized prior to being introduced into the furnace, the liquid reservoir and vaporizer will be in it as well.

[27] There is little application for diffusion sources in the fabrication of gallium arsenide ICs. Ion implantation instead of a predep diffusion is normally used. Diffused junctions are used in discrete devices, however.

8.9.1 Liquid Sources

Typical liquid sources for silicon are boron tribromide (BBr_3) and phosphorus oxychloride ($POCl_3$). They are vaporized by bubbling an inert gas through the liquid. The concentration can be controlled by the liquid's temperature, which sets the vapor pressure, flow of gas through the liquid, and amount of dilution after vaporization. An alternative, and one that eliminates the problem of incomplete saturation in the bubbler, is to use mass flow meters to measure the ratio of gas going into the bubbler to the amount of dopant picked up. The input flow can then be varied as necessary to give the proper amount of dopant in the final input stream. Fig. 8.59 shows a diagram of a typical liquid source using the mass flow control (MFC). MFC–2 controls the total flow of nitrogen through the tube. The combination of MFC–3 and MFC–4 allows the amount of vapor to be controlled. The dopant can be vented until its flow has stabilized. Valves isolate the bubbler or bypass it if gas flow is needed to clear liquid from the lines. The flow constrictor is used to provide enough pressure drop to force nitrogen through the bubbler. The oxygen input is required to react with the incoming source gas and prevent pitting of the silicon surface.

The reactions of BBr_3, which has a boiling point of 90°C, are assumed to be

$$2BBr_3 + \text{heat} \rightarrow 2B + 3Br_2$$

$$4B + 3O_2 \rightarrow 2B_2O_3$$

$$B_2O_3 + SiO_2 \text{ on Si surface} \rightarrow \text{Mixed oxide on surface}$$

$$2B_2O_3 \text{ in mixed oxide} + 3Si \rightarrow 4B + 3SiO_2$$

$$\text{Excess B} + Si \rightarrow SiB_4 \quad \text{or} \quad SiB_6$$

If the silicon surface does not have some protective oxide, the bromine can etch it at diffusion temperature; if too much free boron is formed, it reacts with the silicon wafer to give a mixture of SiB_4 and SiB_6. The boron silicon compounds are difficult to remove and usually must be oxidized away. With a lower BBr_3 concentration in the gas stream, the insoluble layer can be prevented and the doped oxide removed with HF. At an intermediate set of conditions, a nitric acid soluble layer results. In addition, the presence of SiB_4 and SiB_6 increases the surface concentration and gives erratic and nonuniform surface concentration and sheet resistance. Increasing the oxygen decreases the likelihood of forming the unwanted layers, but more than a few percent will produce such thick oxide that surface concentration will begin to decrease. Spacing the wafers too closely will cause an increase in sheet resistance toward the center of the wafer because of restricted gas flow (185–187).

Methyl borate (188, 189), made by saturating methyl alcohol

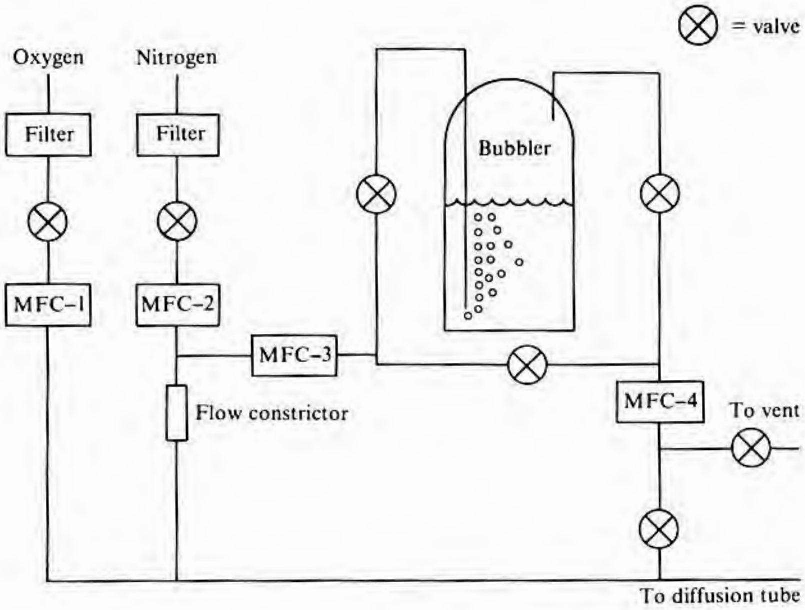


FIGURE 8.59

Piping diagram for liquid diffusion source.

with boric acid, was examined as a liquid source by 1961; however, it has never been widely used. In principle, the carbon in the compound should exit the tube as CO_2, but carbon does deposit on the cooler walls of the entry side of the diffusion tube.

The phosphorus liquid source in common use is $POCl_3$ (having a boiling point of 107°C), although some years ago, PBr_3 was also used (163, 190). As in the case of BBr_3, oxygen must also be included in the gas stream. In the case of PBr_3, the reaction products are presumably analogous to those occurring during the use of BBr_3, with P_2O_5 being the compound deposited on the surface. However, no evidence exists of the formation of a hard-to-remove surface phosphorus compound. It has been reported that with heavy-concentration phosphorus diffusions, silicon phosphide precipitates form in the bulk (191). $POCl_3$ apparently dissociates into several components, such as Cl_2, PCl_3, P, and P_2O_4. The P and P_2O_4 react with oxygen to form P_2O_5 (192). The mixed oxide that forms on the silicon surface prevents attack by the chlorine.

8.9.2 Gaseous Sources

Diborane (B_2H_6), phosphine (PH_3), and arsine (AsH_3) are gas phase dopants that have been used for silicon (189). They are particularly hazardous to health, and extreme caution should be exercised in their use. Like the vaporized liquid sources just discussed, they depend on a high-temperature reaction with oxygen to form an oxide deposit on the wafer surface. Oxygen is also needed to form an initial oxide on the wafer and prevent pitting. In principle, pitting should not occur, but evidence does exist that if free phosphorus is formed on the surface, it will occur. The plumbing for gas phase doping is very simple, requiring only a path for oxygen like the one shown in Fig. 8.59 and one each for nitrogen and the dopant.

Some potential chemical reactions involving phosphine are as follows:

Without oxygen

$$2PH_3 + 440°C \text{ heat} \rightarrow 3H_2 + 2 \text{ red phosphorus}$$

With oxygen

$$PH_3 + 2O_2 + 150°C \text{ heat} \rightarrow H_3PO_4$$

The H_3PO_4 can decompose through a series of steps and thus give $P_2O_5 + H_2O$. Alternatively,

$$2PH_3 + 4O_2 + 300°C \text{ heat} \rightarrow P_2O_5 + 3H_2O$$

However, in an experimental study of reaction products from the oxidation of phosphine, no water, only hydrogen, was found (193). In the case of the oxidation of diborane, both were observed. Regardless of the path followed, the final result is that with phosphine, B_2H_6 or AsH_3, an oxide of the dopant is formed on the surface. Stibine (SbH_3) could also, in principle, be used, but it is even more unstable than arsine and often partially decomposes in its storage cylinder.

Unwanted gaseous sources must also be considered. Since the incoming gases may be contaminated with very fine metal particles that are carried into the furnace, point-of-use gas filters should always be used. Some kinds of furnace tube materials may out-gas. An example is the sintered silicon carbide diffusion tubes introduced several years ago. Vapors from high-temperature heater components such as the heating element, thermocouples, and various impurities in the refractory insulation may penetrate quartz diffusion tubes, particularly after long usage and partial devitrification.

8.9.3 Planar Sources

Instead of bringing the dopant into the furnace via a stream, doping-wafers that are the same size as silicon wafers can be placed adjacent to them in carriers in the diffusion furnace. From the planar source wafers, a suitable dopant such as B_2O_3 can be transferred across the narrow spacing to the silicon. The earliest planar sources were slices of boron nitride (BN) (194). In a high-temperature oxygen ambient preoxidation step, their surfaces are oxidized to B_2O_3. The B_2O_3 will volatilize, diffuse across the narrow spacing between source and wafer, slightly react with the silicon, and form a coating on the wafer that has a lower vapor pressure than that of the B_2O_3 and that will act as a diffusion source. However, a little moisture either in the diffusion ambient or absorbed in the source will lead to the formation of HBO_2, which has a much higher vapor pressure. It will then react with the silicon wafer surface such that (195)

$$2HBO_2 + 2Si \rightarrow 2SiO_2 + 2B + H_2$$

forming a mixed oxide on the surface. Variable moisture will lead to sheet resistance variability, but by keeping the BN dry and then deliberately adding water, generally formed by reacting hydrogen with oxygen in the tube ("hydrogen injection"), reproducibility is improved (196). The B_2O_3 formed on the BN is gradually depleted so that the doping wafers must be periodically reoxidized. Also, since the B_2O_3 layer is hygroscopic, the wafers must be stored between runs at about 350°C in dry nitrogen. To minimize the chance of contamination from impurities in the BN, sources made of B_2O_3 mixed with SiO_2 and other oxides are also available (197). A similar approach has been used to make planar sources for phosphorus (198, 199) and arsenic (200).

8.9.4 Doped Oxide Sources

Instead of forming a doping layer on the wafer during the diffusion cycle, the layer can be predeposited. Either doped spin-on glasses (201–205) or CVD oxides have been used (201). Methods of producing these films were discussed in Chapter 4. Spin-on sources are commercially available for both silicon and gallium arsenide dopants and can be applied with modified photoresist spinners. This type of source is widely used for antimony buried-layer diffusions in silicon.

CVD oxides are not as convenient as spin-on glasses since a CVD reactor and substantially more processing are required and thus are seldom used. They do, however, allow more leeway in the choice of dopants. Typically, a mixed oxide consisting of SiO_2 and an oxide of the dopant are co-deposited. Examples of doping oxides are P_2O_5, B_2O_3, As_2O_3, SnO, and ZnO (206–208, 76).

8.9.5 Closed-Tube Sources

Closed-tube diffusions seal both the wafer to be diffused and the source in a capsule, usually of fused silica (209–212). The tube is evacuated at room temperature, but at operating temperature, substantial pressure due to the partial pressures of the doping agent constituents can exist. Generally, no buildup of a doping layer on the wafer surface will occur; that is, the dopant transfer is directly from the gas phase into the wafer and allows control of surface concentration over a wide range. For diffusion into wafers that tend to dissociate at diffusion temperature, such as GaAs, a sealed tube containing arsenic as well as a doping source affords a way of maintaining a high enough pressure to prevent wafer degradation. By using a long tube and a two-zone furnace, relatively independent control of diffusion temperature and vapor pressure from the additive can be maintained.

If chips are diffused individually, as was sometimes done in the early 1960s, large quantities can be sealed and diffused economically. However, when wafers are used rather than chips, as is now done, and when wafer diameters are quite large, closed-tube diffusions are inconvenient and very expensive.

8.10 THE ERROR FUNCTION

8.10.1 Error Function Algebra

The error function erf(z) (1), given by the integral $(2/\sqrt{\pi})\int_0^z e^{-u^2}du$, can also be evaluated from the series

$$\frac{2}{\sqrt{\pi}}\left[z - \frac{z^3}{3 \cdot 1!} + \frac{z^5}{5 \cdot 2!} \cdots \frac{(-1)^n z^{2n+1}}{(2n+1) \cdot n!}\right.$$

$$\mathrm{erfc}(z) = \frac{2}{\sqrt{\pi}}\int_z^\infty e^{-u^2}du = 1 - \mathrm{erf}(z)$$

$$\mathrm{erf}(-z) = -\mathrm{erf}(z)$$

$$\mathrm{erf}(0) = 0$$

$$\mathrm{erf}(\infty) = 1$$

$$\mathrm{erf}(z) \cong \frac{2z}{\sqrt{\pi}} \qquad \text{for } z << 1$$

$$\mathrm{erf}(z) \cong 1 - \left(\frac{1}{z\sqrt{\pi}}\right)e^{-z^2} \qquad \text{for } z >> 1$$ [28]

$$\frac{d[\mathrm{erf}(z)]}{dz} = \left(\frac{2}{\sqrt{\pi}}\right)e^{-z^2}$$

$$\int_0^\infty \mathrm{erfc}(z)dz = \frac{1}{\sqrt{\pi}}$$

8.10.2 Calculation of Error Function Values

Abbreviated sets of error function and Smith function values are given in Tables 8.12 and 8.13. The highest surface concentration encountered in diffusion problems will be in the range of 10^{20} atoms/cm^3, and a typical wafer background level will be 5×10^{14} atoms/

[28] Functional form only; not accurate enough for most calculations (see expression in the next section).

TABLE 8.12

Error Function erf(z)

z	erf(z)	z	erf(z)	z	erf(z)	z	erf(z)
0.00	0.000 000	0.88	0.786 687	1.76	0.987 190	2.64	0.999 811
0.02	0.022 565	0.90	0.796 908	1.78	0.988 174	2.66	0.999 831
0.04	0.045 111	0.92	0.806 768	1.80	0.989 091	2.68	0.999 849
0.06	0.067 622	0.94	0.816 271	1.82	0.989 943	2.70	0.999 866
0.08	0.090 078	0.96	0.825 424	1.84	0.990 736	2.72	0.999 880
0.10	0.112 463	0.98	0.834 232	1.86	0.991 472	2.74	0.999 893
0.12	0.134 758	1.00	0.842 701	1.88	0.992 156	2.76	0.999 905
0.14	0.156 947	1.02	0.850 838	1.90	0.992 790	2.78	0.999 916
0.16	0.179 012	1.04	0.858 650	1.92	0.993 378	2.80	0.999 925
0.18	0.200 936	1.06	0.866 144	1.94	0.993 923	2.82	0.999 933
0.20	0.222 703	1.08	0.873 326	1.96	0.994 426	2.84	0.999 941
0.22	0.244 296	1.10	0.880 205	1.98	0.994 892	2.86	0.999 948
0.24	0.265 700	1.12	0.886 788	2.00	0.995 322	2.88	0.999 954
0.26	0.286 900	1.14	0.893 082	2.02	0.995 719	2.90	0.999 959
0.28	0.307 880	1.16	0.899 096	2.04	0.996 086	2.92	0.999 964
0.30	0.328 627	1.18	0.904 837	2.06	0.996 423	2.94	0.999 968
0.32	0.349 126	1.20	0.910 314	2.08	0.996 734	2.96	0.999 972
0.34	0.369 365	1.22	0.915 534	2.10	0.997 021	2.98	0.999 975
0.36	0.389 330	1.24	0.920 505	2.12	0.997 284	3.00	0.999 977 91
0.38	0.409 009	1.26	0.925 236	2.14	0.997 525	3.02	0.999 980 53
0.40	0.428 392	1.28	0.929 734	2.16	0.997 747	3.04	0.999 982 86
0.42	0.447 468	1.30	0.934 008	2.18	0.997 951	3.06	0.999 984 92
0.44	0.466 225	1.32	0.938 065	2.20	0.998 137	3.08	0.999 986 74
0.46	0.484 655	1.34	0.941 914	2.22	0.998 308	3.10	0.999 988 35
0.48	0.502 750	1.36	0.945 561	2.24	0.998 464	3.12	0.999 989 77
0.50	0.520 500	1.38	0.949 016	2.26	0.998 607	3.14	0.999 991 03
0.52	0.537 899	1.40	0.952 285	2.28	0.998 738	3.16	0.999 992 14
0.54	0.554 939	1.42	0.955 376	2.30	0.998 857	3.18	0.999 993 11
0.56	0.571 616	1.44	0.958 297	2.32	0.998 966	3.20	0.999 993 97
0.58	0.587 923	1.46	0.961 054	2.34	0.999 065	3.22	0.999 994 73
0.60	0.603 856	1.48	0.963 654	2.36	0.999 155	3.24	0.999 995 40
0.62	0.619 411	1.50	0.966 105	2.38	0.999 237	3.26	0.999 995 98
0.64	0.634 586	1.52	0.968 413	2.40	0.999 311	3.28	0.999 996 49
0.66	0.649 377	1.54	0.970 586	2.42	0.999 379	3.30	0.999 996 94
0.68	0.663 782	1.56	0.972 628	2.44	0.999 441	3.32	0.999 977 34
0.70	0.677 801	1.58	0.974 547	2.46	0.999 497	3.34	0.999 997 68
0.72	0.691 433	1.60	0.976 348	2.48	0.999 547	3.36	0.999 997 983
0.74	0.704 678	1.62	0.978 038	2.50	0.999 593	3.38	0.999 998 247
0.76	0.717 537	1.64	0.979 622	2.52	0.999 634	3.40	0.999 998 478
0.78	0.730 010	1.66	0.981 105	2.54	0.999 672	3.42	0.999 998 679
0.80	0.742 101	1.68	0.982 493	2.56	0.999 706	3.44	0.999 998 855
0.82	0.753 811	1.70	0.983 790	2.58	0.999 736	3.46	0.999 999 008
0.84	0.765 143	1.72	0.985 003	2.60	0.999 764	3.48	0.999 999 141
0.86	0.776 100	1.74	0.986 135	2.62	0.999 789	3.50	0.999 999 257

(*continues*)

TABLE 8.12 (*continued*)

z	erf(z)	z	erf(z)	z	erf(z)	z	erf(z)
3.52	0.999 999 358	3.68	0.999 999 805	3.84	0.999 999 944		
3.54	0.999 999 445	3.70	0.999 999 833	3.86	0.999 999 952		
3.56	0.999 999 521	3.72	0.999 999 857	3.88	0.999 999 959		
3.58	0.999 999 587	3.74	0.999 999 877	3.90	0.999 999 965		
3.60	0.999 999 644	3.76	0.999 999 895	3.92	0.999 999 970		
3.62	0.999 999 694	3.78	0.999 999 910	3.94	0.999 999 975		
3.64	0.999 999 736	3.80	0.999 999 923	3.96	0.999 999 979		
3.66	0.999 999 773	3.82	0.999 999 934	3.98	0.999 999 982		

Note: For a more complete table, see L.J. Comrie, "Chambers Six Figure Mathematical Tables," vol. 2, W. & R. Chambers, Ltd., Edinburgh, 1949, or "Tables of the Error Function and Its Derivative," National Bureau of Standards Applied Mathematical Series, no. 41, Oct. 22, 1954.

TABLE 8.13

Smith Function

β \ α	*0.1*	*0.2*	*0.4*	*0.6*	*0.8*	*1.0*	*1.2*	*1.4*	*1.6*	*1.8*	*2.0*	*2.5*	*3.0*	*4.0*	*5.0*	β / α
0.1	0.09015	0.08155	0.06672	0.05459	0.04467	0.03655	0.02990	0.02446	0.02002	0.01638	0.01340	0.00811	0.00491	0.00180	0.00066	0.1
0.2	0.17838	0.16119	0.13162	0.10748	0.08777	0.07167	0.05853	0.04779	0.03903	0.03187	0.02603	0.01568	0.00945	0.00343	0.00125	0.2
0.4	0.34254	0.30837	0.24993	0.20259	0.16422	0.13314	0.10794	0.08752	0.07097	0.05756	0.04668	0.02766	0.01640	0.00577	0.00204	0.4
0.6	0.48366	0.43290	0.34692	0.27814	0.22308	0.17900	0.14368	0.11538	0.09268	0.07448	0.05988	0.03475	0.02021	0.00688	0.00236	0.6
0.8	0.59940	0.53264	0.42100	0.33317	0.26398	0.20940	0.16628	0.13219	0.10519	0.08379	0.06680	0.03806	0.02180	0.00724	0.00244	0.8
1.0	0.69176	0.60975	0.47475	0.37066	0.29013	0.22765	0.17903	0.14109	0.11141	0.08814	0.06985	0.03931	0.02231	0.00733	0.00246	1.0
1.2	0.76448	0.66808	0.51232	0.39486	0.30574	0.23772	0.18553	0.14529	0.11412	0.08989	0.07098	0.03969	0.02244	0.00735	0.00246	1.2
1.4	0.82144	0.71164	0.53781	0.40979	0.31449	0.24286	0.18855	0.14706	0.11517	0.09051	0.07134	0.03979	0.02247	0.00735	0.00246	1.4
1.6	0.86601	0.74388	0.55469	0.41865	0.31914	0.24530	0.18983	0.14774	0.11552	0.09070	0.07144	0.03981	0.02247	0.00735	0.00246	1.6
1.8	0.90095	0.76759	0.56562	0.42369	0.32147	0.24638	0.19033	0.14797	0.11563	0.09075	0.07147	0.03982	0.02247	0.00735	0.00246	1.8
2.0	0.92838	0.78491	0.57254	0.42646	0.32258	0.24682	0.19051	0.14804	0.11566	0.09076	0.07147	0.03982	0.02247	0.00735	0.00246	2.0
2.5	0.97404	0.81009	0.58029	0.42887	0.32335	0.24707	0.19059	0.14807	0.11567	0.09076	0.07147	.....	.....	.....	.....	2.5
3.0	0.99920	0.82094	0.58234	0.42928	0.32343	0.24708	0.19059	0.14807	0.11567	0.09076	0.07147	.....	.....	.....		3.0
∞	1.02843	0.82795	0.58291	0.42933	0.32343	0.24709	0.19059	0.14807	0.11567	0.09076	0.07147	0.03982	0.02247	0.00735	0.00246	∞

Source: From R.C.T. Smith, "Conduction of Heat in the Semi-Infinite Solid with a Short Table of an Important Integral," *Australian J. Phys. 6*, pp. 127–130, 1953.

cm^3. To plot a diffusion profile over this range will require an N variation of about 6 orders of magnitude. Thus, from Eq. 8.32, for example, the value of the error function must range from 1 to 10^{-6}. The z range corresponding to this is 0.0 to 3.5 and is covered in Table 8.12. If the series expansion from section 8.10.1 is used to calculate erf(z), it should be remembered that it is given by small differences of large numbers and that single-precision BASIC calculations will not provide the required accuracy. A better way, if errors of no more than $\pm 1.5 \times 10^{-7}$ are acceptable, is, for positive z, to use the approximate expression[29]

$$\mathrm{erf}(z) = 1 - (a_1T + a_2T^2 + a_3T^3 + a_4T^4 + a_5T^5)e^{-z^2}$$

where $T = 1/(1 + Pz)$. P and the a_i values are as follows:

$$P = 0.3275911$$

$$a_1 = 0.254829592$$

$$a_2 = -0.284496736$$

$$a_3 = 1.421413741$$

$$a_4 = -1.453152027$$

$$a_5 = 1.061405429$$

SAFETY — CHAPTER 8

Several materials used for diffusion sources are toxic and should therefore be handled with caution. In addition, some safeguards must be included in the facility. Some hazardous materials currently in common use are listed in Table 8.14. In addition, the quantities of nitrogen used in a diffusion area, if vented into an unexhausted room, as, for example, during a power failure, can cause asphyxiation. Of the materials listed in the table, diborane and phosphine are the most toxic. They have, respectively, a TLV of 0.1 ppm and 0.3 ppm and an IDLH of 40 ppm and 200 ppm.[30] Once applied, spin-on

TABLE 8.14

Toxic Materials Used in Diffusion

Item	*Use*
Arsine	Gaseous arsenic source
Diborane	Gaseous boron source
Phosphine	Gaseous phosphorus source
Phosphorus oxychloride	Liquid phosphorus source
Boron tribromide	Liquid boron source
Arsenic spin-on	Liquid spin-on source
Antimony spin-on	Liquid spin-on source

[29] Cecil Hastings, Jr., *Approximations for Digital Computers*, Princeton University Press, Princeton, N.J., 1955.

[30] Recommended threshold limit value (TLV) and immediately dangerous to life or health (IDLH).

materials should present little hazard, but the spin applicator should be carefully shielded and ventilated in order to remove toxic vapors. (See also the safety discussion given earlier in Chapter 4.)

KEY IDEAS CHAPTER 8

□ Diffusion is a mechanism by which impurities migrate from high- to low-concentration regions.

□ If the species is diffusing in silicon and normally resides at a substitutional position in the crystal lattice, it generally moves by jumping into an adjacent vacant lattice space.

□ If the diffusing species normally resides in space between lattice sites (interstices), it generally moves by jumping to an adjacent interstice.

□ The frequency of jumps from position to position is thermally activated.

□ Simple theory describes diffusion in terms of Fick's two laws:

$$J = -D\frac{\partial N}{\partial x}$$

$$\frac{\partial N}{\partial t} = D\frac{\partial^2 N}{\partial x^2}$$

where J is the diffusing atoms' flux, D is the diffusion coefficient, N is the concentration of diffusant, t is the time, and x is the distance.

□ D usually increases exponentially with temperature.

□ When D is not independent of concentration, Fick's laws must be modified to include $\partial D/\partial x$.

□ Solutions of Fick's laws using the appropriate boundary conditions are used to describe the diffusion profiles observed after semiconductor diffusions.

□ The solutions generally involve either erf or Gaussian distributions.

□ Diffusion coefficients are experimentally measured and are available for dopants of interest.

□ Interstitial atoms diffuse much more rapidly than substitutional atoms.

□ Diffusion processes usually are in two steps, with a shallow predep using a diffusion source followed by a drive-in. Particularly in the case of gallium arsenide, the diffused predep step has generally been superseded by an ion implant step.

□ The same formalism used to describe a specific diffusion process can also be used to describe impurity redistributions (diffusion that occurs during other high-temperature processing steps such as oxidation).

□ The most common measurements used to characterize diffusions are the sheet resistance and the junction depth.

□ Gettering is a process for collecting lifetime-degrading impurities present in wafers and storing them in designated regions of the wafer.

□ Gettering steps are often combined with diffusion.

□ Common gettering processes use a heavy back-side concentration of phosphorus, mechanically damaged wafer back side, oxygen precipitation in the middle of the wafer, or misfit dislocations at epi–substrate interfaces.

□ If wafers are heated or cooled too rapidly, slip will be generated. Very heavy surface concentrations of impurities will also generate slip.

□ Diffusion sources are described as liquid, gaseous, doped oxide, or closed-tube, with liquid as the one most commonly used with silicon.

CHAPTER 8 PROBLEMS

1. Express Fick's first law in words. Show how Eq. 8.5 can be derived from Eq. 8.3.
2. Describe the mechanism of diffusion by interstitialcy and compare it to diffusion by interstitial motion. Will an increase in the number of vacancies affect either type of diffusion?
3. What is the distinction between the concentration N_D of an n-type impurity in a wafer and the concentration n of majority carriers in the same wafer? Using the n_i versus temperature curve of Fig. 8.7, calculate the carrier concentration n versus temperature over the range from room temperature to 1100°C when $N_D = 10^{18}$ atoms/cc.
4. Assuming a constant phosphorus surface concentration of 10^{19} atoms/cc and a concentration-independent diffusion constant, calculate the time to produce a junction at a depth of 15 μm in 1 Ω-cm p-type silicon when (a) the diffusion temperature is 1250°C and when (b) it is 1050°C.
5. A sheet of atoms of density 10^{14} atoms/cc is diffused at a temperature that gives a D value of 5×10^{-14} cm²/s. Plot the impurity profile after diffusing for 30 minutes. Plot the surface concentration versus time for a period of 1 hour.
6. Show that for diffusion from a step, regardless of the relative magnitude of the diffusion coefficients of the impurities on each side of the step, the position where the two concentrations are equal moves more and more into the lightly doped side of the step as the diffusion time increases.
7. A 5μm thick, 1 Ω–cm n-type epitaxial layer is grown on a 10 Ω-cm p-type substrate. A p-type ring is to be diffused into the layer so that the ring will electrically isolate the n-region inside the ring from the n-region outside it. Show what the cross section will look like after diffusion. If the p-diffusant is boron, what is the minimum diffusion time at 1150°C required for isolation, assuming no substrate up-diffusion and no autodoping? Use D values from Table 8.2.
8. The material on one side of an abrupt silicon junction is 1 Ω-cm n-type. On the other side, it is 1 Ω-cm p-type. What is the impurity concentration in atoms/cc on each side? After a 2 hour diffusion, will the junction location have moved into the region initially n-type or into the region initially p-type? Justify your answer.
9. A p-type diffusion profile has the shape given by the following points. When the diffusion was into 1 Ω-cm material for 1 hour, the junction depth was 0.98 μm. When an identical diffusion was made in 10 Ω-cm material, the junction depth was 1.14 μm. Determine the diffusion coefficient.

Normalized Concentration	*Depth (μm)*
1.0	0
4.6×10^{-1}	0.2
6.6×10^{-2}	0.5
3.3×10^{-3}	0.8
2.5×10^{-4}	1.0
5.2×10^{-5}	1.1

10. What is the primary advantage of using a heavy phosphorus back-side wafer diffusion for gettering? What is the primary advantage of using misfit dislocations at an epi–substrate interface for gettering? What is a possible major disadvantage of the misfit dislocation approach?

11. Given a (001) oriented wafer, consider the quadrant included between the [010] and $[\bar{1}00]$ directions. If slip occurs between $(\bar{1}11)$ planes in the $[\bar{1}\bar{1}0]$ direction, plot the normalized resolved shear stress as a constant magnitude stress changes direction from $[\bar{1}\bar{1}0]$ to $[0\bar{1}0]$. Based on this plot, sketch where slip might first be expected to appear on a (100) wafer.

CHAPTER 8 REFERENCES

1. Paul G. Shewmon, *Diffusion in Solids*, McGraw-Hill Book Co., New York, 1963.
2. John R. Manning, *Diffusion Kinetics for Atoms in Crystals*, D. Van Nostrand Co., Princeton, N.J., 1968.
3. R.B. Fair and J.C. Tsai, "A Quantitative Model for the Diffusion of Phosphorus in Silicon and the Emitter Push Effect," *J. Electrochem. Soc. 124*, pp. 1107–1118, 1977.
4. Paul Fahey, "New Models of Dopant Diffusion Appropriate for VLSI Fabrication Processes," pp. 216–229, in Kenneth E. Bean and George A. Rozgonyi, eds., *VLSI Science and Technology/84*, Electrochemical Society, Pennington, N.J., 1984.
5. W.R. Wilcox and T.J. LaChapelle, "Mechanism of Gold Diffusion into Silicon," *J. Appl. Phys. 35*, pp. 240–246 and references therein, 1964.
6. U. Gosele et al., "Mechanism and Kinetics of the Diffusion of Gold in Silicon," *Appl. Phys. 23*, pp. 361–368, 1980.
7. M. Hill et al., "Diffusion of Silicon in Gold," *J. Electrochem. Soc. 129*, pp. 1579–1587, 1982.
8. L.R. Weisberg and J. Blanc, "Diffusion with Interstitial-Substitutional Equilibrium. Zinc in GaAs," *Phys. Rev. 131*, pp. 1548–1552, 1963.
9. F.M. Smits, "Formation of Junction Structures by Solid State Diffusion," *Proc. IRE 46*, pp. 1049–1061, 1958.
10. S.M. Hu, "Diffusion in Silicon and Germanium," pp. 217–350, in D. Shaw, ed., *Atomic Diffusion in Semiconductors*, Plenum Publishing Co., London, 1973. (This is an extensive review of diffusion processes and includes almost 400 references.)
11. S.M. Hu, "Formation of Stacking Faults and Enhanced Diffusion in the Oxidation of Silicon," *J. Appl. Phys. 45*, pp. 1567–1573, 1974.
12. Dimitri A. Antoniadis et al., "Boron in Near-Intrinsic <100> and <111> Silicon under Inert and Oxidizing Ambients—Diffusion and Segregation," *J. Electrochem. Soc. 125*, pp. 813–819, 1978.
13. D.A. Antoniadis et al., "Oxidation-Enhanced Diffusion of Arsenic and Phosphorus in Near-Intrinsic <100> Silicon," *Appl. Phys. Lett. 33*, pp. 1030–1033, 1978.
14. R. Francis and P.S. Dobson, "The Effect of Oxidation on the Diffusion of Phosphorus in Silicon," *J. Appl. Phys. 50*, pp. 280–284, 1979.
15. A. Miin-Ron Lin et al., "The Oxidation Rate Dependence of Oxidation-Enhanced Diffusion of Boron and Phosphorus in Silicon," *J. Electrochem. Soc. 128*, pp. 1131–1137, 1981.
16. S.M. Hu, "Kinetics of Interstitial Supersaturation during Oxidation of Silicon," *Appl. Phys. Lett. 43*, pp. 449–451, 1983.
17. R.M. Harris and D.A. Antoniadis, "Silicon Self-Interstitial Supersaturation during Phosphorus Diffusion," *Appl. Phys. Lett. 43*, pp. 937–939, 1983.
18. S.M. Hu, "Kinetics of Interstitial Supersaturation and Enhanced Diffusion in Short-Time/Low-Temperature Oxidation of Silicon," *J. Appl. Phys. 57*, pp. 4527–4533, 1985.
19. Scott T. Dunham and James D. Plummer, "Point-Defect Generation during Oxidation of Silicon in Dry Oxygen. I. Theory," *J. Appl. Phys. 59*, pp. 2541–2550, 1986.
20. Scott T. Dunham and James D. Plummer, "Point-Defect Generation during Oxidation of Silicon in Dry Oxygen. II. Comparison to Experiment," *J. Appl. Phys. 59*, pp. 2551–2561, 1986.
21. K. Taninguchi et al., "Oxidation Enhanced Diffusion of Boron and Phosphorus in (100) Silicon," *J. Electrochem. Soc. 127*, pp. 2243–2248, 1980.
22. J.F. Nye, *Physical Properties of Crystals*, Oxford University Press, London, 1960.
23. J.F. Shepard et al., "Study of a Liquid Boron Dif-

fusion Source for Silicon," *Electrochem. Soc. Ext. Abst.*, abst. no. 196, pp. 87–89, Fall 1966.

24. G.N. Wills, "The Orientation Dependent Diffusion of Boron in Silicon under Oxidizing Conditions," *Solid-State Electronics 12*, pp. 133–134, 1969.
25. L.E. Katz, "Orientation Dependent Diffusion Phenomena," in Charles P. Marsden, ed., *Silicon Device Processing*, NBS Special Pub. 337, pp. 192–199, 1970.
26. W.G. Allen, "Effect of Oxidation on Orientation Dependent Boron Diffusion in Silicon," *Solid-State Electronics 16*, pp. 709–717, 1973.
27. C. Hill, "Measurement of Local Diffusion Coefficients in Planar Device Structures," pp. 988–998, in Howard R. Huff and Rudolph J. Kriegler, eds., *Semiconductor Silicon/81*, Electrochemical Society, Pennington, N.J., 1981.
28. Richard B. Fair, Chap. 7, "Concentration Profiles of Diffused Dopants in Silicon," in F.F. Wang, ed., *Impurity Doping Processes in Silicon*, North-Holland, New York, 1981.
29. Sorab K. Ghandhi, *VLSI Fabrication Principles*, John Wiley & Sons, New York, 1983.
30. J.J. Wortman et al., "Effect of Mechanical Stress on p-n Junction Device Characteristics," *J. Appl. Phys. 35*, pp. 2122–2131, 1964.
31. R.B. Fair, "Modeling Anomalous Phenomena in Arsenic Diffusion in Silicon," pp. 963–987, in H.R. Huff and R.J. Kriegler, eds., *Semiconductor Silicon/81*, Electrochemical Society, Pennington, N.J., 1981.
32. J. Kasahara et al., "The Effect of Stress on the Redistribution of Implanted Impurities in GaAs," *J. Electrochem. Soc. 130*, pp. 2275–2279, 1983.
33. Y. Todokoro and I. Teramoto, "The Stress Enhanced Diffusion of Boron in Silicon," *J. Appl. Phys. 49*, pp. 3527–3529, 1978.
34. A.F.W. Willoughby, "Interactions between Sequential Dopant Diffusions in Silicon—A Review," *J. Phys. D: Appl. Phys. 10*, pp. 455–480, 1977.
35. L.E. Miller, "Uniformity of Junctions in Diffused Silicon Devices," pp. 303–322, in Harry C. Gatos, ed., *Properties of Elemental and Compound Semiconductors*, Interscience Publishers, New York, 1960.
36. S.M. Hu and S. Schmidt, "Interactions in Sequential Diffusion Processes in Semiconductors," *J. Appl. Phys. 39*, pp. 4272–4283, 1968.
37. W. Jost, *Diffusion in Solids, Liquids, and Gases*, Academic Press, New York, 1952.
38. Ron C. Wackwitz, Texas Instruments Incorporated, unpublished work.
39. R.C.T. Smith, "Conduction of Heat in the Semi-Infinite Solid with a Short Table of an Important Integral," *Australian J. Phys. 6*, pp. 127–130, 1953.
40. W.R. Runyan, *Silicon Semiconductor Technology*, McGraw-Hill Book Co., New York, 1965.
41. C.C. Allen and W.R. Runyan, "An Epitaxial Grown–Diffused Silicon Transistor," *IEEE Trans. on Electron Dev. ED-10*, pp. 289–290, 1963.
42. Warren R. Rice, "Diffusion of Impurities during Epitaxy," *Proc. IEEE 52*, pp. 284–295, 1964.
43. A.S. Grove, A. Roder, and C.T. Sah, "Impurity Distribution in Epitaxial Growth," *J. Appl. Phys. 36*, pp. 802–810, 1965
44. F.M. Smits and R.C. Miller, "Rate Limitation at the Surface for Impurity Diffusion in Semiconductors," *Phys. Rev. 104*, pp. 1242–1245, 1956.
45. R.C. Miller and F.M. Smits, "Diffusion of Antimony out of Germanium and Some Properties of the Antimony–Germanium System," *Phys. Rev. 107*, pp. 65–70, 1957.
46. T.I. Kucher; "The Problem of Diffusion in an Evaporating Solid Medium," *Soviet Phys.—Solid State 3*, pp. 401–404, 1961.
47. R.B. Allen et al., "Effect of Oxide Layers on the Diffusion of Phosphorus into Silicon," *J. Appl. Phys. 31*, pp. 334–337, 1960.
48. C.T. Sah et al., "Diffusion of Phosphorus in Silicon Oxide Film," *J. Phys. Chem. Solids 11*, pp. 288–298, 1959.
49. J.C.C. Tsai, "The Simultaneous Diffusion of Donor and Acceptor Impurities into Silicon," Ph.D. thesis, Ohio State University, 1962.
50. A.S. Grove et al., "Redistribution of Acceptor and Donor Impurities during Thermal Oxidation of Silicon," *J. Appl. Phys. 35*, pp. 2695–2701, 1964.
51. Taketoshi Kato and Yoshio Nishi, "Redistribution of Diffused Boron in Silicon by Thermal Oxidation," *Jap. J. Appl. Phys. 3*, pp. 377–383, 1964.
52. R.H. Krambeck, "Numerical Calculation of Im-

purity Redistribution during Thermal Oxidation of Semiconductors," *J. Electrochem. Soc. 121*, pp. 588–591, 1974.

53. E.D. Fabricius, "Diffusion of Impurities in a Thin Semiconductor Slab," *J. Appl. Phys. 33*, pp. 753–754, 1962.

54. J.C. Fisher, "Calculation of Diffusion Penetration Curves for Surface and Grain Boundary Diffusion," *J. Appl. Phys. 22*, pp. 74–77, 1951.

55. R.T.P. Whipple, "Concentration Contours in Grain Boundary Diffusion," *Phil. Mag. Series 7, Vol. 45*, pp. 1225–1236, 1954.

56. Van E. Wood et al., "Theoretical Solutions of Grain-Boundary Diffusion Problems," *J. Appl. Phys. 33*, pp. 3574–3579, 1962.

57. H.J. Queisser et al., "Diffusion along Small Angle Grain Boundaries in Silicon," *Phys. Rev. 123*, pp. 1245–1254, 1961.

58. A. Goetzberger and H. Queisser, "Structural Imperfections in Silicon p-n Junctions," Shockley Transistor Corp. Interim Rpt. No. 1, Contract AF19(604)8060, August 1961.

59. H. Queisser, "Failure Mechanisms in Silicon Semiconductors," Shockley Transistor Corp. Final Rpt., Contract AF30(602)2556, January 1963.

60. T.I. Kamins and S.Y. Chiang, "Lateral Dopant Diffusion in Implanted Buried-Oxide Structures," *Appl. Phys. Lett. 47*, pp. 1197–1199, 1985.

61. P.H. Holloway, "Grain Boundary Diffusion of Phosphorus in Polycrystalline Silicon," *J. Vac. Sci. Technol. 21*, pp. 19–22, 1982.

62. Kouichi Sakamoto et al., "Complete Process Modeling for VLSI Multilayer Structures," *J. Electrochem. Soc. 132*, pp. 2457–2462, 1985.

63. R.G. Elliman et al., "Diffusion and Precipitation in Amorphous Si," *Appl. Phys. Lett. 46*, pp. 478–480, 1985.

64. D.P. Kennedy and R.R. O'Brien, "Analysis of the Impurity Atom Distribution near the Diffusion Mask for a Planar p-n Junction," *IBM J. Res. Develop. 9*, pp. 179–186, 1965.

65. D.D. Warner and C.L. Wilson, "Two Dimensional Concentration Dependent Diffusion," *Bell Syst. Tech. J. 59*, pp. 1–41, 1980.

66. R.C. Wackwitz, "Analytical Solution of the Multiple-Diffusion Problem," *J. Appl. Phys. 33*, pp. 2909–2910, 1962.

67. R.B. Fair and J.C.C. Tsai, "The Diffusion of Ion-Implanted Arsenic in Silicon," *J. Electrochem. Soc. 122*, pp. 1689–1696, 1975.

68. Y. Nakajima et al., "Simplified Expression for the Distribution of Diffused Impurity," *Jap. J. Appl. Phys. 10*, pp. 162–163, 1971.

69. Richard B. Fair, "Boron Diffusion in Silicon—Concentration and Orientation Dependence, Background Effects, and Profile Estimation," *J. Electrochem. Soc. 122*, pp. 800–805, 1975.

70. B. Swaminathan et al., "Diffusion of Arsenic in Polycrystalline Silicon," *Appl. Phys. Lett. 40*, pp. 795–798, 1982.

71. T.I. Kamins et al., "Diffusion of Impurities in Polycrystalline Silicon," *J. Appl. Phys. 43*, pp. 81–91, 1972.

72. C.J. Coe, "The Lateral Diffusion of Boron in Polycrystalline Silicon and Its Influence on the Fabrication of Sub-Micron MOSTs," *Solid-State Electronics 20*, pp. 985–992, 1977.

73. W.M. Bullis, "Properties of Gold in Silicon," *Solid-State Electronics 9*, pp. 143–168 and various included references, 1966.

74. C.W. Farley and B.G. Streetman, "Simulation of Anomalous Acceptor Diffusion in Compound Semiconductors," *J. Electrochem. Soc. 134*, pp. 453–458 and references therein, 1987.

75. I.K. Naik, "Annealing Behavior of GaAs Ion Implanted with p-Type Dopants," *J. Electrochem. Soc. 134*, pp. 1270–1275, 1987.

76. R. Jett Field and Sorab K. Ghandhi, "An Open-Tube Method for Diffusion of Zinc into GaAs," *J. Electrochem. Soc. 129*, pp. 1567–1579, 1982.

77. F.M. Smits, "Measurement of Sheet Resistivity with the Four-Point Probe," *Bell Syst. Tech. J. 37*, pp. 711–718, 1958.

78. David S. Perloff et al., "Four-Point Resistance Measurements of Semiconductor Doping Uniformity," *J. Electrochem. Soc. 124*, pp. 582–590, 1977.

79. J.C. Irving, "Resistivity of Bulk Silicon and Diffused Layers in Silicon," *Bell Syst. Tech. J. 41*, pp. 387–410, 1962.

80. Gerhard Backenstoss, "Evaluation of the Surface Concentration of Diffused Layers in Silicon," *Bell Syst. Tech. J. 37*, pp. 699–710, 1958.

81. O.N. Tufte, "The Average Conductivity and Hall Effect of Diffused Layers in Silicon," *J. Electrochem. Soc. 109*, pp. 235–238, 1962.

82. W.R. Runyan, *Semiconductor Measurements and Instrumentation*, McGraw-Hill Book Co., New York, 1975.

83. C.S. Fuller and J.A. Ditzenberger, "Diffusion of Donor and Acceptor Elements in Silicon," *J. Appl. Phys. 27*, pp. 544–553, 1956.

84. P.A. Illes and B. Leibenhaut, "Diffusant Impurity Concentration Profiles in Thin Layers on Silicon," *Solid-State Electronics 5*, pp. 331–339, 1962.

85. E. Tannenbaum, "Detailed Analysis of Thin Phosphorus Diffused in p-Type Silicon," *Solid-State Electronics 2*, pp. 123–132, 1961.

86. Stacy B. Watelski et al., "A Concentration Gradient Profiling Method," *J. Electrochem. Soc. 112*, pp. 1051–1053, 1965.

87. T.H. Yeh, "Current Status of the Spreading Resistance Probe and Its Application," pp. 111–122, in Charles P. Marsden, ed., *Silicon Device Processing*, NBS Special Pub. 337, 1970.

88. O. Kudoh et al., "Impurity Profiles within a Shallow p-n Junction by a New Differential Spreading Resistance Method," *J. Electrochem. Soc. 123*, pp. 1751–1754, 1976.

89. Masami Konaka et al., "Non-Destructive Determination of Impurity Concentration in Silicon Epitaxial Layer Using Metal–Silicon Schottky Barrier," *Jap. J. Appl. Phys. 7*, pp. 790–791, 1968.

90. K.H. Zaininger and F.P. Heiman, "The *C–V* Technique as an Analytical Tool—Parts I and II," *Solid State Technology*, pp. 49–56 (May), pp. 46–55 (June), 1970.

91. M.G. Buehler, "The D.C. MOSFET Dopant Profile Method," *J. Electrochem. Soc. 127*, pp. 701–704, 1980.

92. C.W. White and W.H. Christie, "The Use of RBS and SIMS to Measure Dopant Profile Changes in Silicon by Pulsed Laser Annealing," *Solid State Technology*, pp. 109–116, September 1980.

93. J.M. Anthony et al., "Super SIMS for Ultrasensitive Impurity Analysis," *Proc. Materials Research Society Symposium 69*, pp. 311–316, 1986.

94. Philip F. Kane and Graydon B. Larrabee, *Characterization of Semiconductor Materials*, McGraw-Hill Book Co., New York, 1970.

95. T.H. Yeh, "Experimental Methods for Determining Diffusion Coefficients in Semiconductors," pp. 155–215, in D. Shaw, ed., *Atomic Diffusion in Semiconductors*, Plenum Publishing Co., London, 1973.

96. Howard Reiss and C.S. Fuller, "Diffusion Processes in Germanium and Silicon," pp. 222–268, in N.B. Hannay, ed., *Semiconductors*, Reinhold Publishing Corp., New York, 1959.

97. B.I. Boltaks, *Diffusion in Semiconductors*, Academic Press, New York, p. 149, 1963.

98. M. Ghezzo, "Diffusion from a Thin Layer into a Semi-Infinite Medium with Concentration Dependent Diffusion Coefficient," *J. Electrochem. Soc. 119*, pp. 977–979, 1972.

99. D. Anderson and K.O. Jeppson, "Nonlinear Two-Step Diffusion in Semiconductors," *J. Electrochem. Soc. 131*, pp. 2675–2679, 1984.

100. Dan Anderson and Kjell O. Jeppson, "Evaluation of Diffusion Coefficients from Non-Linear Impurity Profiles," *J. Electrochem. Soc. 132*, pp. 1409–1412, 1985.

101. R. Ghez et al., "The Analysis of Diffusion Data by a Method of Moments," *J. Electrochem. Soc. 132*, pp. 2759–2761, 1985.

102. C. Hill, "Measurement of Local Diffusion Coefficients in Planar Device Structures," pp. 988–998, in Howard R. Huff et al., eds., *Semiconductor Silicon/81*, Electrochemical Society, Pennington, N.J., 1981.

103. B.L. Morris and L.E. Katz, "Reduction of Excess Phosphorus and Elimination of Defects in Phosphorus Emitter Diffusions," *J. Electrochem. Soc. 125*, pp. 762–765, 1978.

104. W.R. Thurber and W.M. Bullis, "Resistivity and Carrier Lifetime in Gold-Doped Silicon," Final Rpt., PRO Y–71–71–906, AFCRL–72–0076 and the extensive references, January 1972.

105. K.P. Lisiak and A.G. Milnes, "Platinum as a Lifetime-Control Deep Level Impurity in Silicon," *J. Appl. Phys. 46*, pp. 5229–5235, 1975.

106. R. Saito et al., "Dual Diffusion of Gold and Platinum into Silicon," *J. Electrochem. Soc. 132*, pp. 225–229, 1985.

107. S.F. Cagnina, "Enhanced Gold Solubility Effect in Heavily n-Type Silicon," *J. Electrochem. Soc. 116*, pp. 498–502, 1969.

108. P.C. Parekh, "Gettering of Gold and Its Influence on Some Transistor Parameters," *Solid-State Electronics 13*, pp. 1401–1406, 1970.

109. A. Mogro-Campero and R.P. Love, "Localized

Lifetime Control by Argon Ion Implantation into Silicon," *Electrochem. Soc. Ext. Abst. 83–1*, abst. no. 319, pp. 502–503, 1983.

110. Eicke R. Weber, "Transition Metals in Silicon," *Appl. Phys. A 30*, pp. 1–22, 1983.

111. Paul Richmond, "The Effect of Gold Doping upon the Characteristics of MOS Field-Effect Transistors with Applied Substrate Voltage," *Proc. IEEE 56*, pp. 774–775, 1968.

112. D.R. Lamb et al., "The Effect of Gold Doping on the Threshold Voltage, Hall Mobility, Gain, and Current Noise of MOS Transistors," *Int. J. Electronics 30*, pp. 141–147, 1971.

113. Takashi Nishioka et al., "MOSFET's on Au-Diffused High Resistivity Si Substrates," *IEEE Trans. on Electron Dev. ED-29*, pp. 1507–1510, 1982.

114. Kouichirou Hondo et al., "Breakdown in Silicon Oxides—Correlation with Cu Precipitates," *Appl. Phys. Lett. 45*, pp. 270–271, 1984.

115. A. Goetzberger and W. Shockley, "Metal Precipitates in Silicon p-n Junctions," *J. Appl. Phys. 31*, pp. 1821–1824, 1960.

116. J.W. Adamic and J.F. McNamara, "A Study of the Removal of Gold from Silicon Using Phosphorus and Boron Glass Getters," *Electrochem. Soc. Ext. Abst. 13–2*, abst. no. 153, pp. 94–95, October 1964.

117. M. Ghezzo, "Vapor Deposition of Phosphosilicate Glasses from Mixtures of SiH_4, O_2 and $POCl_3$," *J. Electrochem. Soc. 119*, pp. 1428–1430, 1972.

118. R.L. Meek et al., "Diffusion Gettering of Au and Cu in Silicon," *J. Electrochem. Soc. 122*, pp. 786–796, 1975.

119. O. Paz et al., "$POCl_3$ and Boron Gettering of LSI Si Devices: Similarities and Differences," *J. Electrochem. Soc. 126*, pp. 1754–1761, 1979.

120. Livio Balsi et al., "Heavy Metal Gettering in Silicon-Device Processing," *J. Electrochem. Soc. 127*, pp. 164–169, 1980.

121. A. Ourmaszd and W. Schroter, "Phosphorus Gettering and Intrinsic Gettering of Nickel in Silicon," *Appl. Phys. Lett. 45*, pp. 781–783, 1984.

122. G.A. Rozgonyi et al., "Elimination of Oxidation-Induced Stacking Faults by Preoxidation Gettering of Silicon Wafers," *J. Electrochem. Soc. 122*, pp. 1725–1729, 1975.

123. E.J. Mets, "Poisoning and Gettering Effects in Silicon Junctions," *J. Electrochem. Soc. 112*, pp. 420–425, 1965.

124. D.I. Pomerantz, "A Cause and Cure of Stacking Faults in Silicon Epitaxial Layers," *J. Appl. Phys. 38*, pp. 5020–5026, 1967.

125. G.A. Rozgonyi et al., "The Identification, Annihilation, and Suppression of Nucleation Sites Responsible for Epitaxial Stacking Faults," *J. Electrochem. Soc. 123*, pp. 1910–1915, 1976.

126. C.L. Reed and K.M. Mar, "The Effects of Abrasion Gettering on Silicon Material with Swirl Defects," *J. Electrochem. Soc. 127*, pp. 2058–2062, 1980.

127. C.W. Pearce and V.J. Zalackas, "A New Approach to Lattice Damage Gettering," *J. Electrochem. Soc. 126*, pp. 1436–1437, 1979.

128. Y. Hayafuji et al.,"Laser Damage Gettering and Its Application to Lifetime Improvement in Silicon," *J. Electrochem. Soc. 128*, pp. 1975–1980, 1981.

129. T.M. Buck et al., "Gettering Rates of Various Fast-Diffusing Metal Impurities at Ion-Damaged Layers on Silicon," *Appl. Phys. Lett. 21*, pp. 485-487, 1972.

130. T.E. Seidel et al., "Direct Comparison of Ion-Damage Gettering and Phosphorus-Diffusion Gettering of Au in Si," *J. Appl. Phys. 46*, pp. 600–609, 1975.

131. T.W. Sigmon et al., "Ion Implant Gettering of Gold in Silicon," *J. Electrochem. Soc. 123*, pp. 1116–1117, 1976.

132. M.R. Poponiak et al., "Argon Implantation Gettering of Bipolar Devices," *J. Electrochem. Soc. 124*, pp. 1802–1805, 1977.

133. H.J. Geipel and W.K. Tice, "Reduction of Leakage by Implantation Gettering in VLSI Circuits," *IBM J. Res. Develop. 24*, pp. 310–317, 1980.

134. James A. Topich, "Reduction of Defects in Ion Implanted Bipolar Transistors by Argon Backside Damage," *J. Electrochem. Soc. 128*, pp. 866–870, 1981.

135. K.D. Beyer and T.H. Yeh, "Impurity Gettering of Silicon Damage Generated by Ion Implantation through SiO_2 Layers," *J. Electrochem. Soc. 129*, pp. 2527–2530, 1982.

136. P.M. Petroff et al., "Elimination of Process-Induced Stacking Faults by Preoxidation Gettering

of Si Wafers," *J. Electrochem. Soc. 123*, pp. 565–570, 1976.

137. M.C. Chen and V.J. Silvestri, "Post-Epitaxial Polysilicon and Si_3N_4 Gettering in Silicon," *J. Electrochem. Soc. 129*, pp. 1294–1299, 1982.

138. W.T. Stacy et al., "The Microstructure of Polysilicon Backsurface Gettering," *Electrochem. Soc. Ext. Abst. 83–1*, abst. no. 310, pp. 484–485, 1983.

139. W.J.M.J. Jousquin and M.J.E. Ulenaers, "Oxidation-Induced Defects at the Poly/Mono Silicon Interface," *J. Electrochem. Soc. 131*, pp. 2380–2386, 1984.

140. M. Waldner and L. Sivo, "Lifetime Preservation in Diffused Silicon," *J. Electrochem. Soc. 107*, pp. 298–301, 1960.

141. R.S. Ronen and P.H. Robinson, "Hydrogen Chloride and Chlorine Gettering as an Effective Technique for Improving Performance of Silicon Devices," *J. Electrochem. Soc. 119*, pp. 747–752, 1972.

142. D.R. Young and C.M. Osburn, "Minority Carrier Generation Studies in MOS Capacitors on n-Type Silicon," *J. Electrochem. Soc. 120*, pp. 1578–1581 and references therein, 1973.

143. P.D. Esqueda and M.B. Das, "Dependence of Minority Carrier Bulk Generation in Silicon MOS Structures on HCl Concentration in an Oxidizing Ambient," *Solid-State Electronics 23*, pp. 741–746, 1980.

144. S.D. Brotherton, "Electrical Properties of Gold at the Silicon–Dielectric Interface," *J. Appl. Phys. 42*, pp. 2085–2094, 1971.

145. Thomas A. Baginski and Joseph R. Monkowski, "The Role of Chlorine in the Gettering of Metallic Impurities from Silicon," *J. Electrochem. Soc. 132*, pp. 2031–2033, 1985.

146. S.J. Silverman and J.B. Singleton, "Technique for Preserving Lifetime in Diffused Silicon," *J. Electrochem. Soc. 105*, pp. 591–594, 1958.

147. N. Momma et al., "Gettering of Gold and Copper in Silicon during Gallium Diffusion," *J. Electrochem. Soc. 125*, pp. 963–968, 1978.

148. P. Rava et al., "Thermally Activated Oxygen Donors in Si," *J. Electrochem. Soc. 129*, pp. 2844–2849 and references therein, 1982.

149. C.Y. Tan et al., "Intrinsic Gettering by Oxide Precipitation Induced Dislocations in Czochralski Si," *Appl. Phys. Lett. 30*, pp. 175–176, 1977.

150. Robert A. Craven, "Oxygen Precipitation in Czochralski Silicon," pp. 254–271, in Howard R. Huff and Rudolph J. Kriegler, eds., *Semiconductor Silicon/81*, Electrochemical Society, Pennington, N.J., 1981.

151. John Andrews, "Oxygen Out-Diffusion Model for Denuded Zone Formation in Czochralski-Grown Silicon with High Interstitial Oxygen Content," *Electrochem. Soc. Ext. Abst. 83–1*, abst. no. 271, pp. 415–416, 1983.

152. M. Stavola et al., "Diffusivity of Oxygen at the Donor Formation Temperature," *Appl. Phys. Lett. 42*, pp. 73–75, 1983.

153. N. Inoue et al., "Oxygen Precipitation in Czochralski Silicon—Mechanism and Application," pp. 382–393, in Howard R. Huff and Rudolph J. Kriegler, eds., *Semiconductor Silicon/83*, Electrochemical Society, Pennington, N.J., 1981.

154. Charles W. Pearce, "Defect Contamination Control in Silicon Wafer Processing," *Proc. Third Annual Microelectronics Measurement Techniques*, pp. v27–v52 and references therein, 1981.

155. R.A. Hartzell et al., "A Model That Describes the Role of Oxygen, Carbon, and Silicon Interstitials in Silicon Wafers during Device Processing," *Proc. Materials Research Society Symposium 36*, pp. 217–222, 1985.

156. B. Rogers et al., "Computer Simulation of Oxygen Precipitation and Denuded Zone Formation," pp. 74–84, in Kenneth E. Bean and George Rozgonyi, eds., *VLSI Science and Technology/84*, Electrochemical Society, Pennington, N.J., 1984.

157. S. Blackstone et al., "Microdefects during Phosphorus Diffusion," *J. Electrochem. Soc. 129*, pp. 667–668, 1982.

158. H.J. Queisser, "Slip Patterns on Boron-Doped Silicon Surfaces," *J. Appl. Phys. 32*, pp. 1776–1780, 1961.

159. S. Prussin, "Generation and Distribution of Dislocations by Solute Diffusion," *J. Appl. Phys. 32*, pp. 1876–1881, 1961.

160. J.E. Lawrence, "Diffusion Induced Stress and Lattice Disorders in Silicon," *J. Electrochem. Soc. 113*, pp. 819–824, 1966.

161. K.G. McQuhae and A.S. Brown, "The Lattice Contraction Coefficient of Boron and Phospho-

rus in Silicon," *Solid-State Electronics 15*, pp. 259–264, 1972.

162. R.B. Fair, "The Effect of Strain-Induced Band-Gap Narrowing on High Concentration Phosphorus Diffusion in Silicon," *J. Appl. Phys. 50*, pp. 860–868, 1979.
163. R.A. McDonald et al., "Control of Diffusion Induced Dislocations in Phosphorus Diffused Silicon," *Solid-State Electronics 9*, pp. 807–812, 1966.
164. T.H. Yeh and M.L. Joshi, "Strain Compensation in Silicon by Diffused Impurities," *J. Electrochem. Soc. 116*, pp. 73–77, 1969.
165. T.H. Yeh et al., "Diffusion of Tin into Silicon," *J. Appl. Phys. 39*, pp. 4266–4271, 1968.
166. Y. Yukimoto et al., "Effect of Tin on the Diffusion of Impurities in Transistor Structure," pp. 692–697, in Howard R. Huff and Ronald R. Burgess, eds., *Semiconductor Silicon/73*, Electrochemical Society, Pennington, N.J., 1973.
167. Satoru Matsumoto et al., "Effects of Diffusion-Induced Strain and Dislocation on Phosphorus Diffusion in Silicon," *J. Electrochem. Soc. 125*, pp. 1840–1845, 1978.
168. Kenji Morizane and Paul S. Gleim, "Thermal Stress and Plastic Deformation in Thin Silicon Slices," *J. Appl. Phys. 40*, pp. 4104–4107, 1969.
169. S.M. Hu, "Temperature Distribution and Stresses in Circular Wafers in a Row during Radiative Cooling," *J. Appl. Phys. 40*, pp. 4413–4423, 1969.
170. B. Leroy and C. Plougonven, "Warpage of Silicon Wafers," *J. Electrochem. Soc. 127*, pp. 961–970, 1980.
171. D. Thebault and L. Jastrzebski, "Review of Factors Affecting Warpage of Silicon Wafers," *RCA Review 41*, pp. 592–611, 1980.
172. A.E. Widmer and W. Rehwald, "Thermoplastic Deformation of Silicon Wafers," *J. Electrochem. Soc. 133*, pp. 2403–2409, 1986.
173. G.L. Pearson et al., "Deformation and Fracture of Small Silicon Crystals," *Acta Metallurgica 5*, pp. 181–191, 1957.
174. W.D. Sylwestrowicz, "Mechanical Properties of Single Crystals of Silicon," *Phil. Mag. Series 8, Vol. 7*, pp. 1825–1845, 1962.
175. Masato Imai and Koji Suminc, "In Situ X-Ray Topograph Study of the Dislocation Mobility in High-Purity and Impurity Doped Silicon Crystals," *Phil. Mag. Series A, Vol. 47*, pp. 599–621, 1983.
176. J.R. Patel and A.R. Chaudhuri, "Oxygen Precipitation Effects on the Deformation of Dislocation-Free Silicon," *J. Appl. Phys. 33*, pp. 2223–2224, 1962.
177. S.M. Hu and W.J. Patrick, "Effect of Oxygen on Dislocation Movement in Silicon," *J. Appl. Phys. 46*, pp. 1869–1883, 1975.
178. K. Yasutake et al., "Mechanical Properties of Heat-Treated Czochralski-Grown Silicon Crystals," *Appl. Phys. Lett. 73*, pp. 789–791, 1980.
179. Yojiro Kondo, "Plastic Deformation and Preheat Treatment Effects in CZ and FZ Silicon Crystals," pp. 220–231, in Howard R. Huff et al., eds., *Semiconductor Silicon/81*, Electrochemical Society, Pennington, N.J., 1981.
180. Ichiro Yonenga and Koji Sumino, "Mechanical Strength of Silicon Crystals as a Function of the Oxygen Concentration," *J. Appl. Phys. 56*, pp. 2346–2350, 1984.
181. T. Abe et al., "Impurities in Silicon Single Crystals," pp. 54–71, in Howard R. Huff et al., eds., *Semiconductor Silicon/81*, Electrochemical Society, Pennington, N.J., 1981.
182. Koji Sumino et al., "Effects of Nitrogen on Dislocation Behavior and Mechanical Strength in Silicon Crystals," *J. Appl. Phys. 54*, pp. 5016–5020, 1983.
183. Joe Lambert and Chris Bayne, "A Suspended Boat Loader Based on the Cantilever Principle," *Semiconductor International*, pp. 150–155, 1983.
184. Arthur Waugh and Bryan D. Foster, "Design and Performance of Silicon Carbide Cantilever Paddles in Semiconductor Diffusion Furnaces," *American Ceramic Society Bul. 64*, pp. 550–554, 1985.
185. Pravin C. Parekh and David R. Goldstein, "The Influence of Reaction Kinetics between BBr_3 and O_2 on the Uniformity of Base Diffusion," *Proc. IEEE 57*, pp. 1507–1512, 1969.
186. G.M. Oleszek and W.M. Whittemore, "The Effect of Process Variables on the Open-Tube Diffusion of Boron into Silicon from Boron Tribromide," pp. 490–501, in Rolf R. Haberecht and Edward L. Kern, eds., *Semiconductor Silicon/69*, Electrochemical Society, New York, 1969.

187. P. Negrini et al., "Boron Predeposition in Silicon Using BBr_3," *J. Electrochem. Soc. 125*, pp. 609–613, 1978.

188. Gary Calson, Texas Instruments Incorporated, unpublished work.

189. R.M. Burger and R.P. Donovan, eds., *Fundamentals of Silicon Integrated Device Technology*, Prentice-Hall, Englewood Cliffs, N.J., 1967.

190. W Greig et al., "Diffusion Technology for Advanced Microelectronics Processing," pp. 168–174, in Charles P. Marsden, ed., *Silicon Device Processing*, NBS Special Pub. 337, 1970.

191. P.F. Schmidt and R. Stickler, "Silicon Phosphide Precipitates in Diffused Silicon," *J. Electrochem. Soc. 111*, pp. 1188–1189, 1964.

192. P.C. Parekh, "On the Uniformity of Phosphorus Emitter Concentration for Shallow Diffused Transistors," *J. Electrochem. Soc. 119*, pp. 173–177, 1972.

193. K. Strater and A. Mayer, "The Oxidation of Silane, Phosphine and Diborane during Deposition of Doped Oxide Diffusion Sources," pp. 469–480, in Rolf R. Haberecht and Edward L. Kern, eds., *Semiconductor Silicon/69*, Electrochemical Society, New York, 1969.

194. N. Goldsmith et al., "Boron Nitride as a Diffusion Source for Silicon," *RCA Review 28*, pp. 344–350, 1967.

195. David Rupprecht and Joseph Stach, "Oxidized Boron Nitride Wafers as an In-Situ Boron Dopant for Silicon Diffusions," *J. Electrochem. Soc. 120*, pp. 1266–1271, 1973.

196. J. Stach and J. Kruest, "A Versatile Boron Diffusion Process," *Solid State Technology*, pp. 60–67, October 1976.

197. J.J. Steslow et al., "Advances in Solid Planar Dopant Sources for Silicon," *Solid State Technology*, pp. 31–34, January 1975.

198. N. Jones et al., "A Solid Planar Source for Phosphorus Diffusion," *J. Electrochem. Soc. 123*, pp. 1565–1569, 1976.

199. R. Wheeler and J.E. Rapp, "Improved Transistor Characteristics Using Solid Planar Diffusion Sources for Emitter Diffusion," *Solid State Technology*, pp. 203–205, August 1985.

200. R.E. Tressler et al., "Present Status of Arsenic Planar Diffusion Sources," *Solid State Technology*, pp. 165–171, October 1984.

201. M.L. Barry, "Diffusion from Doped Oxide Sources," pp. 175–181, in Charles P. Marsden, ed., *Silicon Device Processing*, NBS Special Pub. 337 and references therein, 1970.

202. K. Reindl, "Spun-On Arsenosilica Films as Sources for Shallow Arsenic Diffusions with High Surface Concentration," *Solid-State Electronics 16*, pp. 181–189, 1973.

203. B.H. Justice et al., "Diffusion Processing of Arsenic Spin-On Diffusion Sources," *Solid State Technology*, pp. 39–42, July 1978.

204. T.C. Chandler et al., "Debris-Induced Effects from Spin-On Diffusion Sources," *J. Electrochem. Soc. 126*, pp. 2216–2220, 1979.

205. B.H. Justice et al., "A Novel Spin-On Dopant," *Solid State Technology*, pp. 153–159, October 1984.

206. D.M. Brown et al., "Characteristics of Doped Oxides and Their Use in Silicon Device Fabrication," *J. Crystal Growth 17*, pp. 276–287, 1972.

207. M. Gezzo and D.M. Brown, "Arsenic Glass Source Diffusion in Si and SiO_2," *J. Electrochem. Soc. 120*, pp. 110–116, 1973.

208. B. Jayant Baliga and Sorab K. Ghandhi, "Planar Diffusion in Gallium Arsenide from Tin-Doped Oxides," *J. Electrochem. Soc. 126*, pp. 135–138, 1979.

209. R. Gereth et al., "Solid–Solid Vacuum Diffusion Processes in Silicon," *J. Electrochem. Soc. 120*, pp. 966–971, 1973.

210. A. Kostka et al., "A Physical and Mathematical Approach to Mass Transport in Capsule Diffusion Processes," *J. Electrochem. Soc. 120*, pp. 971–974, 1973.

211. K.K. Shih, "High Surface Concentration Zn Diffusion in GaAs," *J. Electrochem. Soc. 123*, pp. 1737–1740, 1976.

212. O. Hasegawa and R. Namazu, "Zn Diffusion into GaAs by a Two-Temperature Method," *Appl. Phys. Lett. 36*, pp. 203–205, 1980.

CHAPTER 9

Ion Implantation

9.1 INTRODUCTION

Ions that have been accelerated by several kilovolts have enough energy to penetrate a solid surface and, unlike the ions of a typical diffusion process, can do so even when the solid is at room temperature. Such an operation is referred to as ion implantation and offers an alternative to thermal diffusion for introducing impurities into a semiconductor surface. The penetration depth is quite small, and unlike diffusion, ion implantation produces the maximum concentration beneath the wafer surface, as is illustrated in Fig. 9.1. The depth of the concentration peak increases as the accelerating voltage increases, and the total number of ions injected is proportional to the beam current and implant time. The depth spread depends inversely on the ratio of the mass of the host atom to the implanted ion.

As-implanted dopants are generally not in the proper lattice position and are mostly electrically inactive. In addition, the implant operation can generate substantial damage to the host crystal lattice. Most of the crystal damage and the electrical inactivity can be corrected by appropriate high-temperature anneals. Thus, while an implant can be, and generally is, done near room temperature, before it is useful, the semiconductor must be subjected to a high temperature heat treatment.

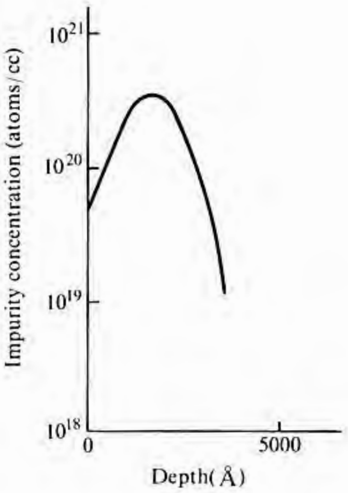
Applications of ion implanting to the semiconductor industry can grouped into the following categories:

1. A source of doping atoms.
2. A method of introducing gettering damage into the semiconductor.
3. The introduction of a layer of different composition into the wafer.
4. A means of supplying a known quantity of atoms for subsequent study.

Use as a source of doping atoms is the oldest and the most important implant application. As discussed in Chapter 8, in most cases, im-

FIGURE 9.1

Typical implanted impurity profile showing that, unlike a diffused profile, the concentration peak can be well below surface.



plantation can be used directly as an alternative to a diffusion predeposition. However, if no clear-cut advantage exists, diffusion is preferred since it is a less-expensive step. Indeed, for several years, this was the case for all diffusions, and high-volume ion implant operations were restricted to MOS transistor threshold adjusts, in which there is no reasonable alternative. Then, as shallower diffusions and lower-temperature processing became necessary, ion implantation became more appropriate than diffusion for some kinds of bipolar and MOS transistors. In the case of gallium arsenide, where substantial decomposition may occur during a diffusion predep, ion implantation is used almost exclusively.

Implantation can introduce an appreciable amount of damage, even to the point of producing amorphous layers in many cases. In conventional applications, the damage presents a problem and must be removed. However, some applications, such as gettering and the enhanced mixing of atoms at interfaces, depend on the damage.

By implanting a high concentration of oxygen or nitrogen in a silicon wafer, a discrete silicon oxide or nitride layer can be formed beneath the surface. If the implanting is done in a manner that does not amorphize the single-crystal silicon above the oxide or nitride, the result is a buried insulating layer surrounded with single-crystal silicon.

Ion implantation affords a relatively simple means of placing a known number of atoms in a wafer. Such a capability is very useful in determining, for example, the effect of a specific amount of impurity on lifetime.

9.2 IMPLANT DEPTH (RANGE)

As a low-energy ion moves into a solid, it will follow a zigzag path as it is deflected from one atom to another by nuclear collisions (see Fig. 9.2). Very-high-energy ions will initially lose much of their energy by electron interactions and will travel a relatively straight path until enough energy is lost for nuclear collisions to begin absorbing energy, at which time the zigzag path will begin. The total length of this path is called the range. However, it is the depth below the surface that is of practical interest, and that distance is referred to as the projected range—that is, the range projected onto the x axis as shown in Fig. 9.2. The amount of lateral travel (spreading of the beam) from the point of entry is given by projecting the range onto the y axis. Unlike that of diffusion, lateral penetration here is considerably less than penetration in the direction of the beam. Since each ion is subjected to a somewhat different set of conditions, the actual range will be different for each ion. The projected range just referred to is the arithmetic mean (average) of the projected range of a large number of implanted ions. To a first approximation, the scatter can be described by the standard deviation of the projected ranges.

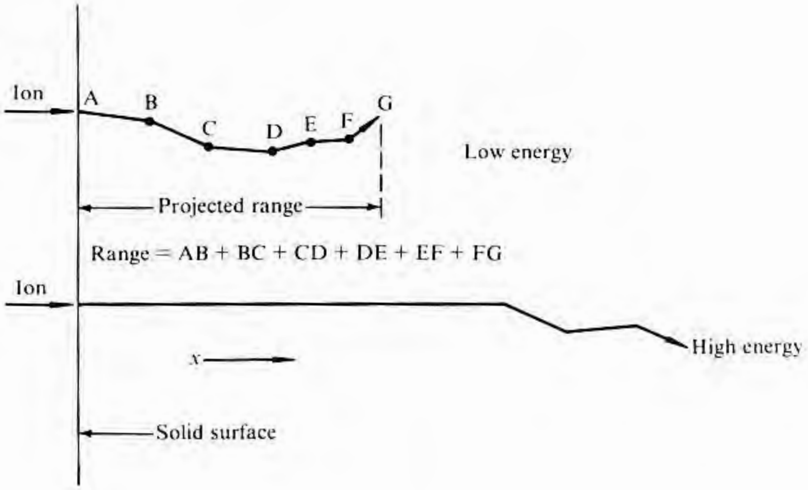
9.2.1 Range Calculations

The range can be expressed in terms of the distance of travel required for the energy of the incoming ion to be reduced from its initial value of E_0 to 0. The incremental energy loss $dE/d\zeta$, where ζ is a coordinate along the path of travel (and whose direction will change with each collision), is given by

$$\frac{dE}{d\zeta} = -N[S_n(E) + S_e(E)] \qquad 9.1$$

FIGURE 9.2

Paths of low- and high-energy ions. (The high-energy ion initially follows a straight path and loses energy by electron interactions. When it slows down sufficiently, nuclear collisions become most important.)



where N is the number of atoms per unit volume in the material being implanted, S_n is the nuclear stopping power, and S_e is the electronic stopping power. From Eq. 9.1,

$$\int d\zeta = -\frac{1}{N}\int \frac{dE}{S_n + S_e} \qquad 9.2$$

where the integral of $d\zeta$ is just the range R. Thus, if expressions for S_n and S_e are available, Eq. 9.2 can be integrated to give the range. To a first approximation (1, 2),

$$S_n = \frac{2.8 \times 10^{-15} M_1 Z_1 Z_2}{(M_1 + M_2)(Z_1^{2/3} + Z_2^{2/3})^{1/2}} \quad \text{eV-cm}^2 \qquad 9.3$$

where Z_1 is the atomic number of the ion, Z_2 is the atomic number of the target material, M_1 is the atomic mass of the ion, and M_2 is the atomic mass of the target. In this approximation, S_n is independent of the ion energy. In a similar approximation,

$$S_e = KE^{1/2} \quad \text{eV} - \text{cm}^2 \qquad 9.4$$

where K is proportional to $Z_1 + Z_2$ and is, for a silicon target, $0.2 \times 10^{-15}\sqrt{\text{eV}}$ cm^2. Experimentally, as shown in Fig. 9.3, S_e oscillates with Z (3, 4), at least for lower Z numbers. By assuming a somewhat different electronic screening function than was assumed in the original derivation, the shape of the curve of Fig. 9.3 can be reasonably well predicted theoretically (5).

What constitutes high-energy ions in relation to determining which loss mechanism dominates depends on the mass and charge

FIGURE 9.3

Electronic stopping power S_e versus atomic number for ions traveling with initial velocity of 1.5×10^8 cm/s in a [110] direction of silicon. (*Source:* From data in F.H. Eisen, *Can. J. Phys. 46*, p. 561, 1968, and James Comas and Robert G. Wilson, *J. Appl. Phys. 51*, p. 3697, 1980.)

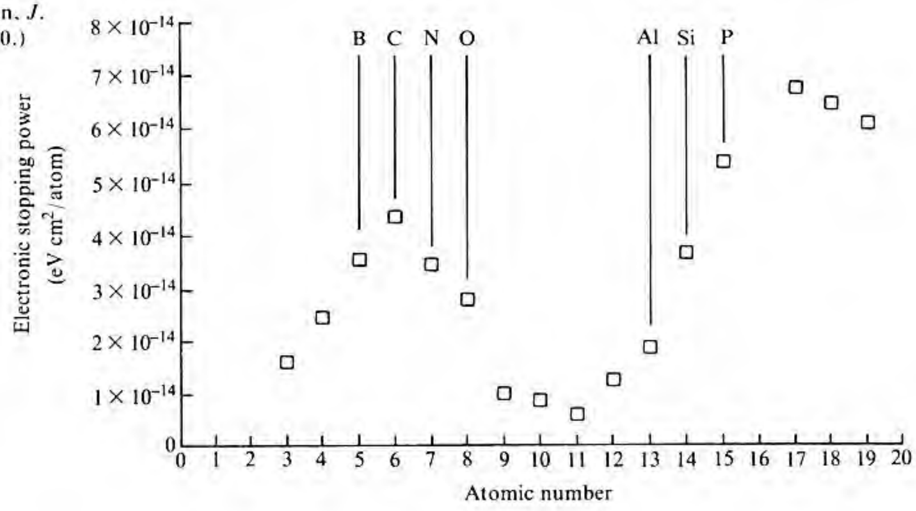
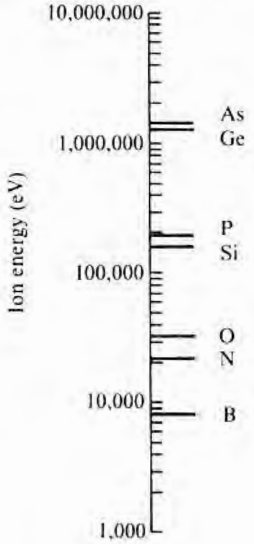


FIGURE 9.4

Approximate energy of ion being implanted into silicon for which nuclear and electronic stopping powers are equal.



of the ion. As an approximation, for energies of E_c, given by Eq. 9.5, losses from nuclear collisions equal those from electron interactions (1):

$$E_c = \frac{14Z_1Z_2M_1}{(M_1 + M_2)(Z_1^{2/3} + Z_2^{2/3})^{1/2}} \qquad 9.5$$

As the energy increases above E_c, the electronic contribution becomes dominant. Fig. 9.4 shows E_c for several materials used in silicon IC processing. Thus, except for boron, nitrogen, and perhaps oxygen, nuclear interactions will dominate since most implants use single-charge-state ions and energies of less than 200 keV.

In order to calculate the projected range R_p with reasonable accuracy, an integral much more complex than that of Eq. 9.2 must be used (1, 6).[1] In amorphous materials, the projected range can be roughly characterized by a mean depth and a standard deviation. These two parameters have been calculated for a wide variety of ions, targets, and energies and are available in book form (7, 8). Also, simplified approaches can be run on programmable calculators, and somewhat more sophisticated approaches can be used with relatively simple FORTRAN programs (9).

To a first approximation, the depth depends only on the implant ion mass and its energy. The standard deviation of the depth, however, depends on the ratio of the mass of the implant ion to the host atom. The larger the ratio, the smaller the standard deviation. As a reasonable approximation in the region where S_n dominates, the projected range is given by (1)

$$R_p = \frac{3M_1R}{3M_1 + M_2} \qquad 9.6$$

and is more accurate for $M_1 > M_2$, in which case the implanted ion has a higher atomic mass than the matrix. Calculated values (7) of R_p versus implant energy for selected impurities in silicon and gallium arsenide are given in Figs. 9.5 and 9.6. From Eq. 9.3, it would be expected that at low beam energies, the range would decrease with increasing atomic number, and that is what is shown in Fig. 9.5. The atomic numbers of most of the ions of interest for implanting in silicon lie between those of boron and antimony, so the ranges of interest are included between these two curves. At higher energies, where S_e is of greater significance, the range will then depend more on the value of S_e, which oscillates as Z increases. Thus, while

[1] This is often referred to as the LSS theory after the authors Linhard, Scharff, and Schiøtt.

FIGURE 9.5

Projected range versus implant energy for ions in silicon. (*Source:* From data in James F. Gibbons et al., *Projected Range Statistics*, 2d ed., Dowden, Hutchinson and Ross, Stroudsburg, Pa., 1975.)

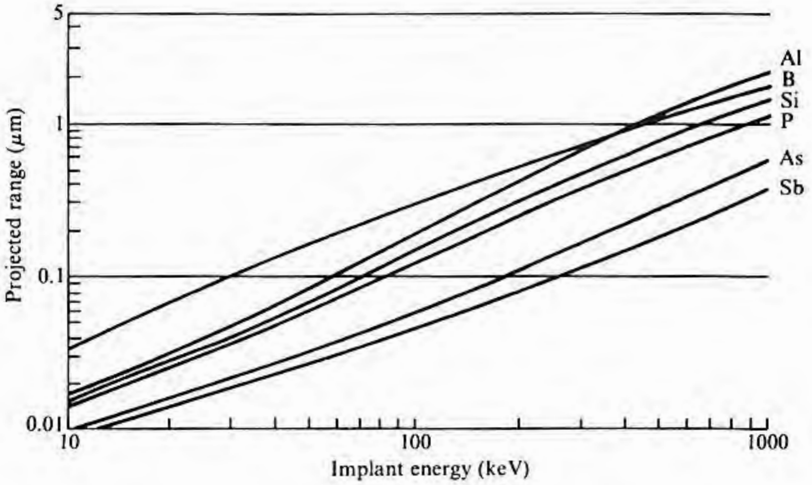
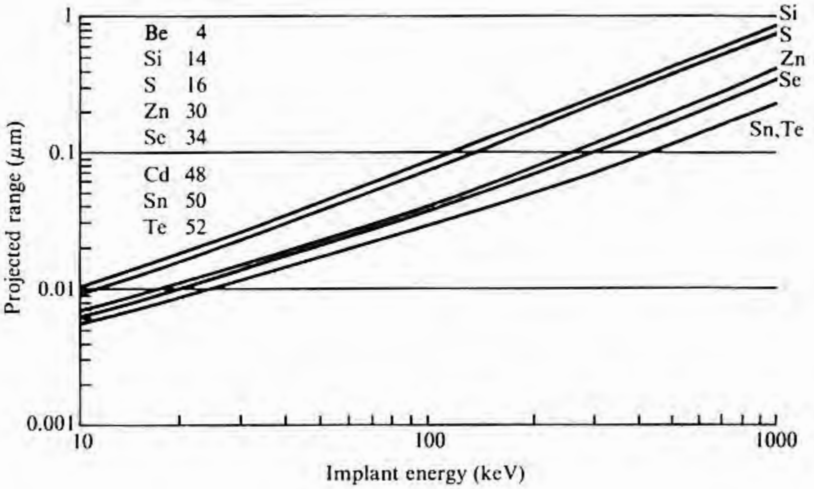


FIGURE 9.6

Projected range versus implant energy for ions in gallium arsenide. (*Source:* From data in James F. Gibbons et al., *Projected Range Statistics*, 2d ed., Dowden, Hutchinson and Ross, Stroudsburg, Pa., 1975.)



the ranges of aluminum, silicon, and phosphorus (atomic numbers of 13, 14, and 15) are very close together for an implant energy of 10 keV, because of the S_e variation shown in Fig. 9.3, a much wider separation occurs at 1 MeV. For reference, Table 9.1 lists for many of the elements the atomic number, mass, and relative abundance of the two major isotopes.

9.2.2 Range Dispersion

The projected range R_p is defined as $R_p = \Sigma r_p/N$ where r_p is the projected range of an individual ion and N is the total number of ions. If the ranges are distributed in a "normal" fashion about R_p,

TABLE 9.1

Atomic Numbers and Masses

No.	*Element*	*Mass(% Abundance)**	*No.*	*Element*	*Mass(% Abundance)*
1	Hydrogen	1(99.985)	35	Bromine	79(50.7), 81(49.3)
2	Helium	4(99.99986)	36	Krypton	84(57.0), 86(17.3)
3	Lithium	7(92.5), 6(7.5)	37	Rubidium	85(72.2), 87(27.8)
4	Beryllium	9(100)	38	Strontium	88(82.6), 86(9.9)
5	Boron	11(80), 10(20)	39	Yttrium	89(100)
6	Carbon	12(98.9), 13(1.1)	40	Zirconium	90(51.5), 94(17.3)
7	Nitrogen	14(99.6), 15(0.4)	41	Niobium	93(100)
8	Oxygen	16(99.76), 18(0.2)	42	Molybdenum	98(24.1), 96(16.7)
9	Fluorine	19(100)	43	—	—
10	Neon	20(90.5), 22(9.2)	44	Ruthenium	102(31.6), 104(18.7)
11	Sodium	23(100)	45	Rhodium	103(100)
12	Magnesium	24(79.0), 25(11.0)	46	Palladium	106(27.3), 108(26.5)
13	Aluminum	30(100)	47	Silver	107(51.8), 109(48.2)
14	Silicon	28(92.2), 29(4.67)	48	Cadmium	114(28.7), 112(24.1)
15	Phosphorus	31(100)	49	Indium	115(95.5), 113(4.5)
16	Sulfur	32(95), 34(4.2)	50	Tin	120(32.4), 118(24.3)
17	Chlorine	35(76), 37(24)	51	Antimony	121(57.3), 123(42.7)
18	Argon	40(99.0), 38(0.6)	52	Tellurium	130(33.8), 128(31.7)
19	Potassium	39(93.3), 41(6.7)	53	Iodine	127(100)
20	Calcium	40(96.94), 44(2.1)	54	Xenon	132(26.9), 129(26.4)
21	Scandium	45(100)	55	Cesium	133(100)
22	Titanium	48(73.8), 47(8.0)	56	Barium	138(71.7), 137(11.2)
23	Vanadium	51(99.75), 50(0.25)	57–71	Rare earths	—
24	Chromium	52(83.8), 53(9.5)	72	Hafnium	180(35.2), 178(27.1)
25	Manganese	55(100)	73	Tantalum	181(99.98)
26	Iron	56(91.7), 54(5.8)	74	Tungsten	184(30.7), 186(28.6)
27	Cobalt	59(100)	75	Rhenium	187(62.6), 185(37.4)
28	Nickel	58(68.3), 60(26.1)	76	Osmium	192(41.0), 190(26.4)
29	Copper	63(69.2), 65(30.8)	77	Iridium	193(62.7), 191(37.3)
30	Zinc	64(48.6), 66(27.9)	78	Platinum	195(33.8), 194(32.9)
31	Gallium	69(60.1), 71(39.9)	79	Gold	197(100)
32	Germanium	74(36.5), 72(27.4)	80	Mercury	202(29.6), 200(23.1)
33	Arsenic	75(100)	81	Thallium	205(70.5), 203(29.5)
34	Selenium	80(49.6), 78(23.5)	82	Lead	208(52.4), 206(24.1)
			83	Bismuth	209(100)

*The two most abundant isotopes are listed.

Source: Data from *Handbook of Chemistry and Physics*, 68th ed. (1987–1988), CRC Press, Boca Raton, Fla.

the distribution can be characterized in terms of the standard deviation σ where

$$\sigma = \frac{\sqrt{\Sigma R^2}}{N}$$

and $R = R_p - r_p$. The distribution of the number of ions $N(x)$ or $N(r_p)$ is given by

$$N(x) = \left(\frac{\phi}{\sigma\sqrt{2\pi}}\right)e^{-(x-R_p)^2/2\sigma^2} \qquad 9.8$$

where ϕ is the fluence per cm^2 (usually referred to as "dose"). $N(x)$ is in ions/cm^3. In implantation literature, the symbol ΔR_p is used instead of σ and is generally termed "ion straggle." Eq. 9.8 thus becomes

$$N(x) = \left(\frac{\phi}{\Delta R_p\sqrt{2\pi}}\right)e^{-(x-R_p)^2/2\Delta R_p^2} \qquad 9.9$$

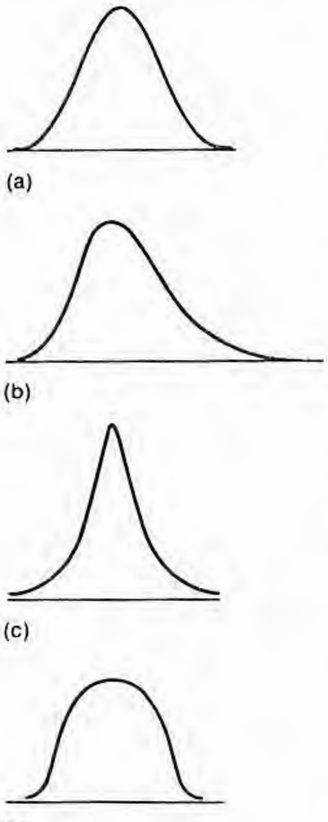
At the same time as the r_p's are calculated and an R_p is determined, a σ can also be calculated and is usually tabulated along with R_p. Unfortunately, there is nothing a priori that requires the distribution of r_p to actually be normal, and there are other curves that would meet the requirements of a given R_p and σ. Experimentally, it has been observed that the distributions do appear somewhat normal, and, for some time, Eq. 9.9 was a sufficient approximation.[2] However, more sophisticated uses and the widespread implementation of computer modeling have required a better description of the distribution. When a normal distribution as in Fig. 9.7a is skewed as in Fig. 9.7b, a measure of the skewness is given by the third moment Π_3. The curve might also have either a sharper or a blunter peak as in Figs. 9.7c or 9.7d. This distribution is referred to as kurtosic. The fourth moment[3] Π_4 is a measure of kurtosis. When Π_4 is normalized by dividing by σ^2, a value of 3 gives a normal distribution, values less than 3 indicate a curve with a flatter top (pla-

[2] When implanting is done into crystalline rather than amorphous solids, channeling, to be described in a later section, may occur. In such cases, major deviations from a normal distribution can occur. The present discussion pertains only to amorphous implanting.

[3] The first moment Π_1 is defined as $(\Sigma R)/N$ and is zero. The second moment Π_2 is given by $(\Sigma R^2)/N$ and is just σ^2. The third moment Π_3 is $(\Sigma R^3)/N$ and will be nonzero only when there is skew. The fourth moment Π_4 is $(\Sigma R^4)/N$ and is a measure of kurtosis. A normalized skewness γ_1 is sometimes defined by Π_3/σ; a normalized kurtosis β, by Π_4/σ. In implant literature, a first moment μ_1 is sometimes defined as R_p. (For more information, see a standard text on statistics.)

FIGURE 9.7

(a) Normal (Gaussian), (b) skewed, (c) leptokurtic, and (d) platykurtic distributions.


(a)

(b)

(c)

(d)

tykurtic), and values greater than 3 indicate a more peaked curve (leptokurtic).

The approach to finding an analytical expression for the implant profile has been to search for variations (usually algebraically complex) of Eq. 9.9 that will, with the tabulated inputs of R_p, σ, and Π_3, predict reasonably well the experimentally observed curves. One method of modifying Eq. 9.9 is by multiplying it by a polynomial. An early example of this approach, used long before the advent of ion implantation, was to use the first two terms of the Gram–Charlier series (10). Applied to Eq. 9.8, it gives

$$N(x) = \left(\frac{\phi}{\sigma\sqrt{2\pi}}\right)e^{-(x-R_p)^2/2\sigma^2}\left[1 - \frac{\alpha_3}{2}\left(\frac{x}{\sigma} - \frac{x^3}{3\sigma^3}\right)\right] \qquad 9.10$$

α_3 is the normalized skewness factor given by

$$\alpha_3^2 = \beta_1 = \frac{\Pi_3^2}{\Pi_2^3} \qquad 9.11$$

How the third moment and Eq. 9.10 can change a normal (Gaussian) profile is illustrated in Fig. 9.8. This particular equation gives a very poor fit with experimental implant data. However, an extension of Eq. 9.10 with more terms has been used, as has the joining of halves of two separate Gaussian curves with different σ's (11). To get a good fit of some ion profiles, the fourth moment must be also included, although it is often calculated from the third moment and hence is not an independent variable. The most successful equation is that of the Pearson IV distribution[4] (12, 13), but it is sometimes combined with an exponential function to further tailor the distribution tails. A FORTRAN program for profile calculations using Pearson IV is given in reference 13. SUPREM modeling of implant profiles is based on Pearson IV plus an exponential for each half of the distribution.

In general, the standard deviation σ decreases as the implant energy increases and as the atomic number of the implanted species increases. Fig. 9.9 shows the relative dispersion (σ/R_p) versus implant energy for several ions implanted into silicon. As Z increases, the nature of the curve changes from that of boron ($Z = 5$) to that of phosphorus ($Z = 15$) and then to that of As ($Z = 33$) and Sb ($Z = 51$). Typically, the lighter ions have a larger relative dispersion at low energies and a lower one at high energies than the heavy ions.

[4] Karl Pearson was a professor of applied mathematics at University College, London. He published tables of distribution function values that were determined uniquely by not just the second moment (normal curve), but by the second and third, and second, third, and fourth moments.

EXAMPLE □ Using the simple distribution equation (Eq. 9.9), determine the total implant flux ϕ required to give a peak concentration of 10^{18} atoms/cc if boron is implanted at 200 keV into silicon with negligible background doping.

Inspection of Eq. 9.9 shows that the peak $N(x)$ will occur when $x = R_p$ and that its value is given by $N(x) = (\phi/\Delta R_p\sqrt{2\pi})e^0$. ΔR_p can be found in a set of range tables, or it can be inferred from Figs. 9.5 and 9.9, which give R_p and the ratio σ/R_p (remember that $\sigma \equiv \Delta R_p$). From the figures, R_p = 0.53 μm and σ/R_p = 0.175 so that σ = 0.093 μm. Solving for ϕ gives

$$\begin{aligned}\phi &= (10^{18}\ \text{atoms/cm}^3)(2\pi)^{0.5}(0.09\ \mu\text{m} \times 10^{-4}\ \text{cm}/\mu\text{m}) \\ &= 2.3 \times 10^{14}\ \text{atoms/cm}^2\end{aligned}$$

□

The impact of the change of standard deviation on the implant profile is shown in Fig. 9.10 for the specific case of boron in silicon.

FIGURE 9.8

Example of Gaussian curve and one using a third moment to skew curve to left. (Gaussian, based on range and second moment of 150 keV boron implant in silicon; third moment, one third that calculated for boron.)

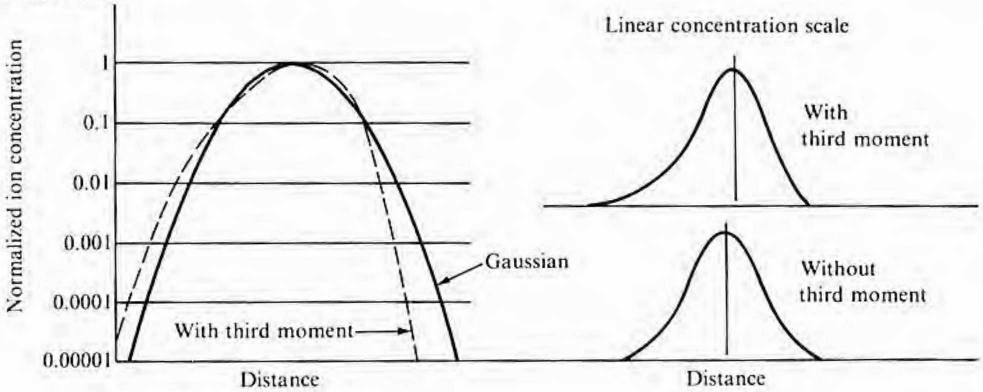
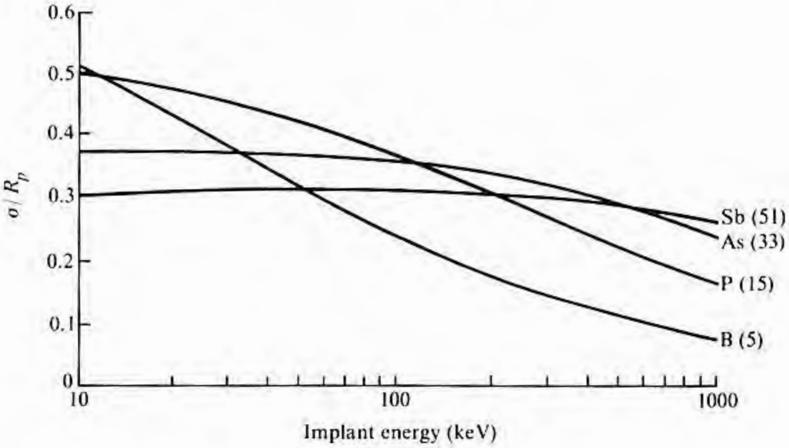


FIGURE 9.9

Ratio of projected range standard deviation to projected range. (*Source:* From calculated data in James F. Gibbons et al., *Projected Range Statistics*, 2d ed., Dowden, Hutchinson and Ross, Stroudsburg, Pa., 1975.)



The curves are plotted from the simple expression of Eq. 9.9 and make use of a high and a low implant energy to provide differing dispersions.

9.2.3 Lateral Range Dispersion

Because of the same random path travel that leads to ion straggle in the forward direction, there is also a scatter of ions perpendicular to the path of the incident beam. This number, like the standard deviation of the range, is also calculable and is listed in some range tables, such as the one in reference 7. It is referred to as lateral standard deviation $R_\perp$. $R_\perp$ is calculated based on cylindrical coordinates so that in Cartesian coordinates with the beam normal to the x axis, the lateral standard deviations σ_y and σ_z (also referred to as ΔY and ΔZ) are given by $\sigma_y = \sigma_z = R_\perp/\sqrt{2}$. $R_\perp$ is generally larger than σ and, for lighter atoms, may be twice σ at high implant energies. The values come closer together as the implant energy decreases, and as the ion mass increases, the two values are close together over the whole energy range. These trends are shown in Fig. 9.11.

FIGURE 9.10

Effect of implant voltage on implant depth and on spread of ions, plotted using Gaussian distribution.

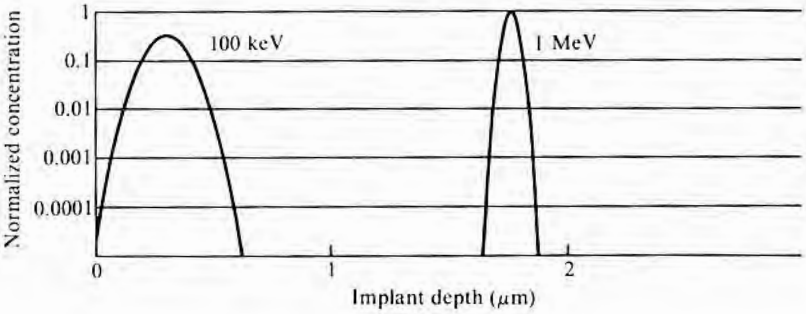
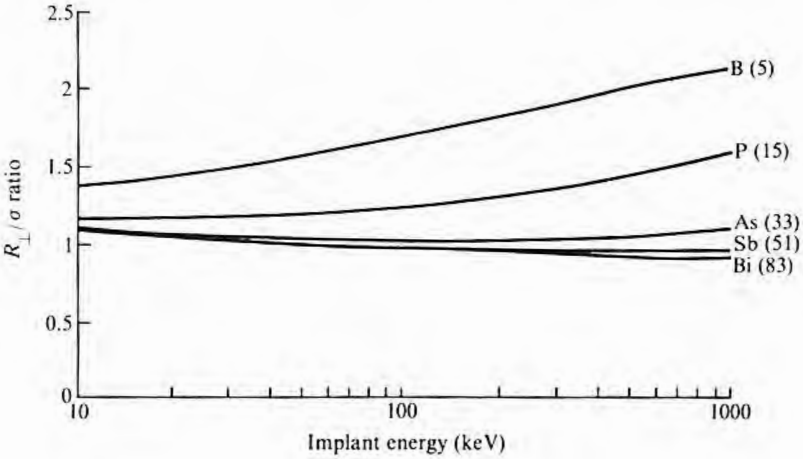


FIGURE 9.11

Ratio of $R_\perp/\sigma$ versus implant energy for ions of progressively higher atomic weight. (*Source:* From data in James F. Gibbons et al., *Projected Range Statistics*, 2d ed., Dowden, Hutchinson and Ross, Stroudsburg, Pa., 1975.)



The edge profile due to this scatter has been calculated for the specific case of implanting through a slit mask with the length much greater than the width and is given by (14)

$$N(x,y) = \left(\frac{\phi}{2\pi\sigma^2}\right)e^{(x-R_p)^2/2\sigma^2}\left[\left(\frac{1}{\sqrt{\pi}}\right) erf\left(\frac{x-a}{\sigma_y\sqrt{\pi}}\right)\right] \qquad 9.12$$

The erf expression does not arise from the solution to a diffusion equation but rather is an approximation when the slit length to width ratio is large. Eq. 9.12 is based on the assumption that the mask is straight walled and allows no transmission for $y > \pm a$. This gives impurity profiles like the one shown in Fig. 9.12. If the walls are tapered, then some ions will penetrate the mask and give the impression of an extralarge lateral spreading (15). Some assumptions lead to the conclusion that the lateral spreading can even be as much as that found during diffusion—that is, an amount comparable to R_p. While calculated R_p and σ values can and have been checked experimentally and found to agree very well, it is much more difficult to experimentally check Eq. 9.12 because of the masking difficulties. However, some data using implanted krypton indicate good agreement (16).

9.2.4 Channeling

The range and its standard deviation as discussed in the previous section assume that the implanted material is amorphous so that, regardless of the direction of travel of an ion, there will, on the average, be the same number of atoms per unit length of path. In single-crystal material, there are directions in which no nuclei will be encountered, and as long as the ion travels in one of those directions, its only stopping force will be due to electronic interactions (S_e) and the range will be considerably increased (channeled). The maximum channeling range for phosphorus implantation in <111> and <110> directions in single-crystal silicon, along with the cal-

FIGURE 9.12

Calculated ion isoconcentration lines for use of implanting 70 keV boron ions into silicon through an infinitely long, 1 μm wide opening.
(*Source:* Seijiro Furukawa et al., *Jap. J. Appl. Phys. 11*, p. 134, 1972.)

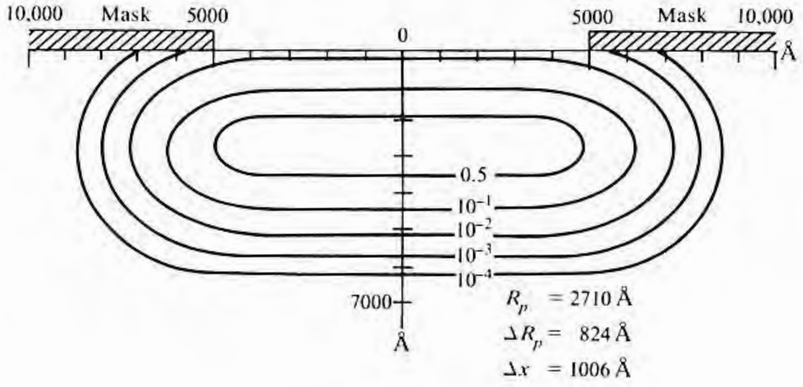
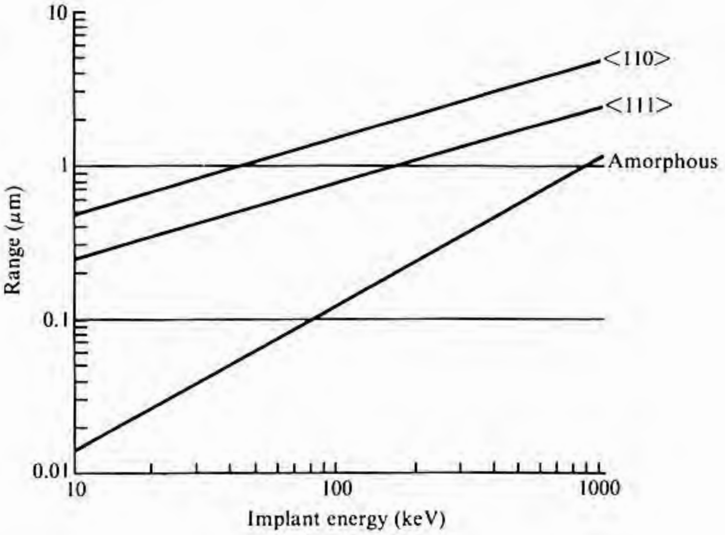


FIGURE 9.13

Maximum channeling range and amorphous range versus implant energy for phosphorus implanted in silicon.
(*Source:* From data in Goode et al., *Radiation Effects* 6, p. 237, 1970, and James F. Gibbons et al., *Projected Range Statistics, Semiconductors and Related Materials*, 2d ed., Dowden, Hutchinson, and Ross, Stroudsburg, Pa., 1975.)



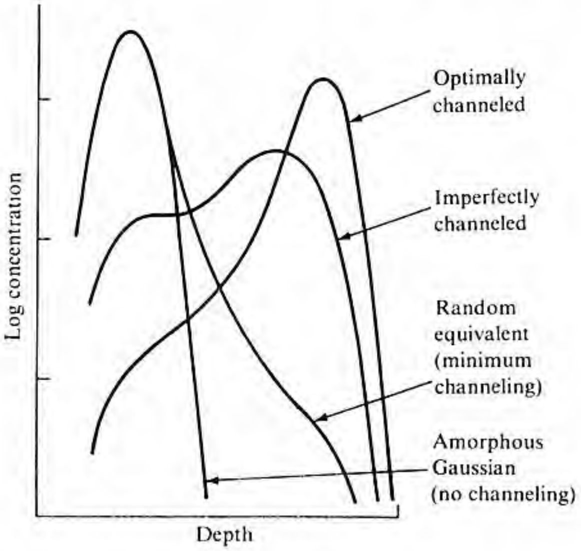
culated range in amorphous silicon, is shown in Fig. 9.13. This figure shows that the maximum range varies as the square root of implant energy. It also shows that since the range for amorphous materials has a different energy dependency at low energies, the separation between the two ranges decreases as energy increases. When the energy is high enough for S_e to be the major loss mechanism, the amorphous curve will become parallel to the single-crystal channel curves but will remain somewhat below them. It is for this reason that channeling becomes less important as implant energy increases.

As larger and larger fractions of the ions are channeled, the observed profile will change as shown in Fig. 9.14. For the case of the diamond and zinc blende structures, the most open directions are the <110> and <100> (see Appendix A), but many other directions offer some degree of freedom.

The ion beam direction does not have to be exactly aligned for channeling to occur. If an ion enters an open path at an angle of less than some critical angle ψ_c, which for most implant energies of interest varies as $1/E^{1/4}$ (17), it will be trapped and channeled along the path. However, when the direction of travel is not straight down the middle of the channel, the energy loss per unit distance of travel will be greater, and the range will be reduced. There is an increased likelihood of the ion's being deflected out of the channel (dechanneled). Even when the ion enters at an angle greater than ψ_c, channeling may still occur if subsequent nuclear scattering places it in a channeling direction. It will then continue to travel in that direction unless it becomes dechanneled.

FIGURE 9.14

Change in implant profile from amorphous to maximum channeling. (*Source:* Adapted from Robert G. Wilson et al., *NBS Special Pub. 400–49*, November 1978.)

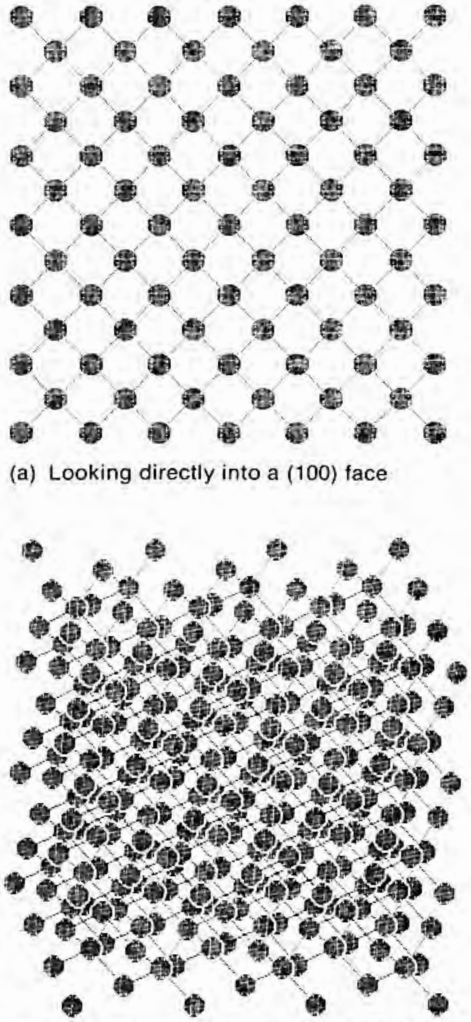


In most cases, one goal of implanting procedures is to keep the implant profile as near Gaussian as possible, which means minimizing or eliminating channeling. This can be done by preamorphizing the front surface of the wafer (18), by implanting through an amorphous layer such as thermal silicon oxide, or by changing the direction of the ion beam so that it does not enter in a direction that favors channeling. A qualitative picture of the effect of changing beam direction can be obtained from viewing a crystal model from different directions. Fig. 9.15a shows how the lattice appears when looking directly into a (100) face. A regular network of openings exists in which an ion might travel into the crystal in a <100> direction and not collide with a lattice atom. Contrast this view with the view of Fig. 9.15b, in which the crystal has been rotated 10° about the <001> zone axis and 20° about the <010> zone axis. In this case, the lattice atom positions appear to be nearly random, and no obvious paths exist down which an ion could travel without collision. Typically, misorientations of 7°–10° are used, but often little care is taken to ensure that the tilt is always about the optimum axis.

Theoretical modeling indicates that <100> and <110> axial channeling and (111) and (220) planar channeling are quite likely for boron implanted into silicon in a <100> direction (19). Planar channels are those in which the ions are constrained to move only between sets of parallel planes and not down given axial directions. Experimentally, over 30 different directions within 20° of the <100> direction (microchannels) have been observed that allow enhanced range (20). Based on both theoretical and experimental data (21), it

FIGURE 9.15

View of diamond lattice model showing position of atoms and bonds. (In part (b) some of the open spaces occur because the computer program producing the view used only a small number of atoms.)
(*Source:* Program courtesy of Dr. Anthony Stephens, Texas Instruments Incorporated.)



(a) Looking directly into a (100) face

(b) Looking into lattice in a direction chosen to impede motion of an ion (see text for direction)

appears that for (100) silicon wafers, a 7° tilt about an axis making an angle of 30° with the flat (a "thirty degree twist") should minimize channeling. Views of a crystal model (similar to that of Fig. 9.15b) also indicate a high degree of randomizing when that orientation is used (22). As the implant energy decreases, progressively more tilt is required. For example, if a 4° critical angle is calculated for arsenic implanted at 80 keV, then at 10 keV, the angle would have increased to 7° and, from a practical standpoint, the tilt angle should be about twice the critical angle in order to ensure that most ions are not channeled (23).

9.2.5 Implanting through Layers

Implanting through a surface layer occurs, for example, when ion implanting is being used for threshold adjust, ion mixing, or the shifting of Schottky barrier heights. These applications are discussed in section 9.4. In addition, when a layer of some material is used as a mask to localize the implant area, implanting through the layer will occur if it is not thick enough. The more commonly encountered layer materials are photoresist, silicon oxide, and silicon nitride. The range and moments for them have also been calculated (7, 8) and can be used in estimating a composite profile. In the case of CVD layers, the range may change with deposition conditions. The tabulations indicate substantial differences in resist range values as the formulation changes. For the specific cases of AZ111 (positive resist) and KTFR (negative resist), the range and ΔR_p in μm for 200 keV implants are as follows (7, 8):

	AZ111		*KTFR*	
	R_p	ΔR_p	R_p	ΔR_p
Boron	1.84	0.15	0.67	0.072
Phosphorus	0.85	0.13	0.40	0.097
Arsenic	0.44	0.073	0.17	0.038

As a very rough approximation, the implant profile can be calculated by considering the layer and substrate as one material. A better approximation for the case of the layer having a thickness d greater than the range R_{p1} in the layer is (13, 24)

$$N(x) = \left(\frac{\phi}{\Delta R_{p1}\sqrt{2\pi}}\right)e^{-(x-R_{p1})^2/2\Delta R_{p1}^2} \qquad \text{for } x < d \qquad 9.13a$$

$$N(x) = \left(\frac{\phi}{\Delta R_{p2}\sqrt{2\pi}}\right)e^{-[\{d+(R_{p1}-d)\Delta R_{p2}/\Delta R_{p1}-x\}^2/2\Delta R_{p2}^2]} \qquad 9.13b$$
$$\text{for } x > d$$

where R_{p2} is the range in the substrate. For the case in which the layer thickness is considerably less than the range R_{p1}, an approximation that can be used with higher moment distributions (such as Pearson IV) is (13)

$$N(x) = \left(\frac{\Delta R_{p2}}{\Delta R_{p1}}\right)N(x') \qquad \text{for } x < d \qquad 9.14a$$

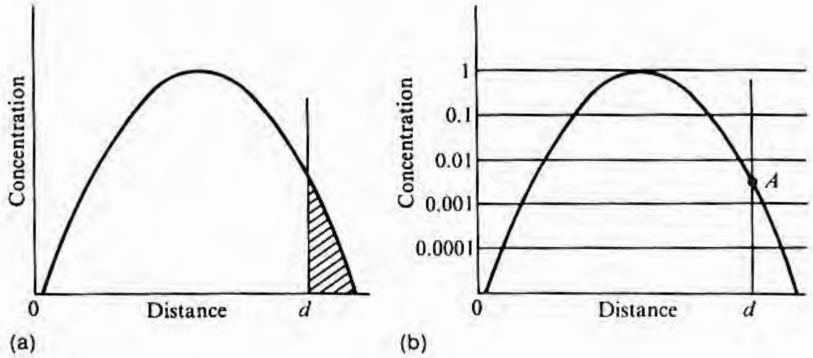
$$N(x) = N(x'') \qquad \text{for } x > d \qquad 9.14b$$

where $x' = (\Delta R_{p2}/\Delta R_{p1})x$ and $x'' = x - d(1 - \Delta R_{p2}/\Delta R_{p1})$.

In specifying masking, either the thickness of masking required to absorb a given fraction of the implanted dose or the thick-

FIGURE 9.16

Masking efficiency required as expressed either in terms of allowed flux leakage (crosshatched portion of part a), or in terms of maximum concentration in material being masked (approximately at *A* in part b; see text). (The mask–semiconductor interface is at *d*.)



ness required to ensure that $N(x)$ in the masked material is less than some value may be specified. These two choices are shown in Fig. 9.16, which has been plotted from Eq. 9.9 rather than from the better approximation of Eq. 9.13. In the example shown (using a range of 0.5 μm and a ΔR_p of 0.1 μm), $\phi_e/\phi = 0.0003$, while N/N_0 is only 0.0035. ϕ_e is the number of ions per cm^2 that escape the masking layer (crosshatched portion of Fig. 9.16a), ϕ is the total dose, N is the concentration at the substrate surface (at A in Fig. 9.16b), and N_0 is the maximum concentration. This example illustrates the need for an almost complete blocking of ions in order to prevent problems with high surface concentration. N can be calculated[5] from Eqs. 9.9, 9.13, or 9.14. ϕ_e is given by $\int_d^\infty N(x)dx$, and when Eq. 9.9 is used,

$$\frac{\phi_e}{\phi} = \int_d^\infty \left(\frac{1}{\Delta R_p\sqrt{2\pi}}\right) e^{-[(x-R_p)^2/2\Delta R_p^2]}dx \qquad 9.15$$

Since

$$\operatorname{erfc}(z) = \left(\frac{2}{\sqrt{\pi}}\right)\int_z^\infty e^{-z^2}dz$$

Eq. 9.15 becomes

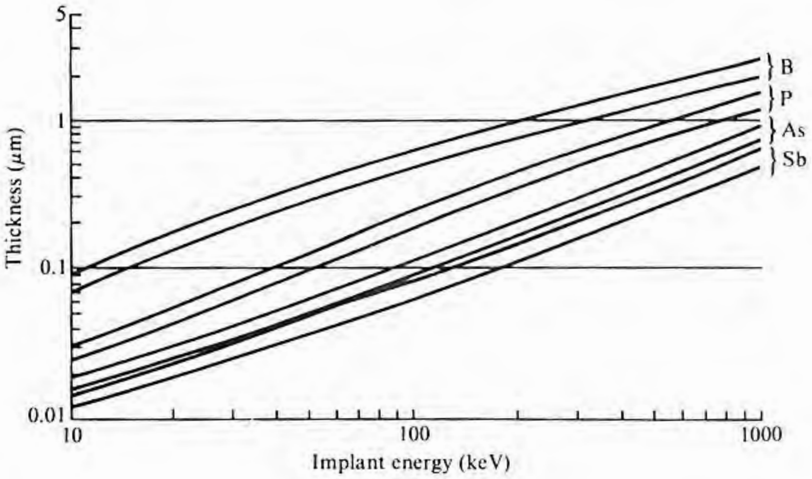
$$\frac{\phi_e}{\phi} = \frac{1}{2}\operatorname{erfc}\left(\frac{d-R_p}{\Delta R_p\sqrt{2}}\right) \qquad 9.16$$

Fig. 9.17 shows some required thicknesses, based on Eq. 9.16.

[5] It must be remembered that, like many of the other calculated values discussed in this book, substantial errors can occur in these numbers and experimental data must be taken to verify the predictions.

FIGURE 9.17

Calculated thickness of layer required to stop 99.999% of implanted ions. (For each pair, the top curve is for silicon dioxide and the bottom one is for silicon nitride.)
(*Source:* From Eq. 9.9 and James F. Gibbons et al., *Projected Range Statistics, Semiconductors and Related Materials*, 2d ed., Dowden, Hutchinson and Ross, Stroudsburg, Pa., 1975.)



EXAMPLE □ When a boron implant at 200 keV is used, what thickness of SiO_2 will be required to mask 99.999% of the implanted ions?

$\phi_s/\phi = 10^{-5}$ and, from the previous example, $R_p = 0.53$ μm and $\Delta R_p = 0.093$ μm. Substituting values into Eq. 9.16 gives

$$2 \times 10^{-5} = \operatorname{erfc}\left(\frac{d \times 10^{-4} - 0.53 \times 10^{-4}}{0.93 \times 1.414}\right) = \operatorname{erfc}(z)$$

where d is the desired thickness in μm. So $\operatorname{erfc}(z) = 2 \times 10^{-5} = 1 - \operatorname{erf}(z)$. Thus, $\operatorname{erf}(z) = 0.99998$ and, from the error function table in Chapter 8, $z_c = 3.02$. Since, by definition,

$$z = \frac{d \times 10^{-4} - 0.53 \times 10^{-4}}{0.93 \times 1.414}$$

$d = 0.93$ μm, which matches the data of Fig. 9.17. □

When Eq. 9.16 is used, the thickness will be overestimated if the masking material has a negative third moment. If the third moment is positive and/or if there is substantial tailing (fourth moment), the thickness will be underestimated. Also, during high-dose implants, the edges of photoresist may round (25) so that because of the effect discussed in section. 9.2.3, implanted geometries may be larger than anticipated. As can be seen from Fig. 9.17 and Table 9.2, high-energy implants require an inordinate mask thickness for all of the materials normally used in IC fabrication unless implants are restricted to the higher atomic weight materials such as arsenic and antimony. Of the materials listed, gold has the shortest range but would still need to be over a micron thick to mask boron at 1000 keV implant energy. Polyimide, even though it does not have great stop-

TABLE 9.2

Range in μm of 1000 keV Ions in Various Materials

	Ion			
Material	*B*	*P*	*As*	*Sb*
KTFR			0.849	
AZ111	5.889	3.718	2.200	1.548
Polyimide	Experimentally, appears comparable to KTFR.			
Si_3N_4	1.500	0.720	0.373	0.242
SiO_2	1.939	0.931	0.483	0.313
Al	1.802	1.025	0.518	0.332
Si	1.756	1.123	0.585	0.375
Ge	2.066	0.783	0.366	0.230
Mo			0.214	
Au	1.004	0.320	0.145	0.093
W	No data, but should be close to Au since density, atomic number, and atomic weight are similar.			

Sources: From data in reference 7, with information on KTFR from data in reference 8 and information on polyimide from T.O. Herndon et al., *Solid State Technology 179,* November 1984.

ping power, has potential because of its ability to have relatively straight-walled features etched in thick layers. Unfortunately, thick layers of organic materials sometimes out-gas enough during high-dose implants to affect implanter performance (26).

9.3 IMPLANT-INDUCED DEFECTS

As the ions being implanted move through a crystal lattice, they displace atoms in their path and thus cause a substantial amount of crystallographic damage. For most device applications, it is necessary to remove as much of the damage as possible. Further, the implanted ions generally do not come to rest in substitutional lattice sites. In order for them to exhibit the proper electrical properties, those implanted ions that would normally occupy substitutional sites must be induced to move to their proper positions. In the case of silicon (and germanium), all substitutional sites are equivalent. In the case of gallium arsenide (and other III–V compounds), there are two sites, corresponding to the normal locations of each constituent. Sometimes, even though the implanted ions move to substitutional sites, they move to the wrong ones and produce antistructure crystal defects. Thermal annealing is used both to repair the crystallographic damage produced by the passage of the implanted ions and to allow the implanted ions to move to their proper sites. Such annealing can be done slowly in a conventional diffusion tube or quickly by a rapid thermal anneal (RTA). The latter may be by scanned laser heating or by a quick exposure to high-intensity radiant heat.

9.3.1 Lattice Damage*

The kind of lattice damage initially produced depends somewhat on whether the ion is dissipating energy primarily by nuclear collisions or by inelastic electronic interactions. When nuclear collisions occur, atoms will be displaced from their substitutional sites and become interstitials, leaving vacancies behind. The energy imparted to atoms in the path is often enough to cause them to displace additional atoms so that a broad region of damage results. In some cases, each ion will leave a trail of amorphous material.

When there is enough overlapping of damage, an amorphous layer is produced in silicon. In the case of gallium arsenide, heavy lattice damage but no true amorphous regions are observed (28). A substantial difference exists between the annealing behavior of amorphous silicon regions and those that are heavily damaged but still crystalline. Ordinarily, the amorphous region is easier to regrow into damage-free material than is the region that has not quite been amorphized. Thus, the conditions for producing it have been studied extensively. For a given ion, the lower the implant temperature, the smaller the dose required for amorphizing. For a given temperature, the higher the atomic number of the ion, the smaller the dose. This behavior is shown in Fig. 9.18. The curves, which fit the experiment reasonably well, are based on a simple theory that assumes that the amorphous region formed by each ion is a right circular cylinder of length R_p and radius $r(T)$ that is temperature sensitive. The projection of the cross-sectional area of the amorphous cylinder onto the wafer surface is assumed to be A_i. The growth of the amorphized fractional area A_a is given by (27, 29)

$$\frac{dA_a}{d\phi} = A_i P_a = A_i(1 - A_a) \qquad 9.17$$

where P_a is the probability that an unamorphized area will be hit by the next ion. The solution of Eq. 9.17 is

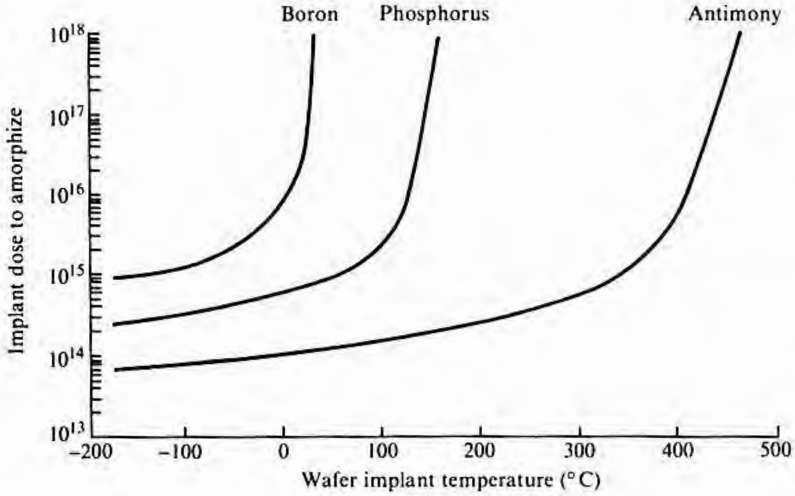
$$A_a = 1 - e^{-A_i\phi} \qquad 9.18$$

A_i is a function of temperature through the radius r of the damage cluster. It is assumed that the outlying damage will be repaired by defects diffusing away and thus reduce the final value of r by an amount $\Delta r(T)$. The amount of diffusion will increase as the implanted wafer temperature increases. As diffusion increases, r and consequently A_i will decrease and slow down the rate of amorphization. A critical implant dose ϕ^* can be defined as the amount required to amorphize a large fraction of the total area—for example,

*See reference 27.

FIGURE 9.18

Effect of wafer temperature during implant on dose required to produce an amorphous layer. (*Source:* From curves in James F. Gibbons, *Proc. IEEE 60*, p. 1062, 1972.)



90%. Combining this criterion with a variation of r with T of the form $r = r_0 - \Delta r(T) = r_0 - ae^{-b/kT}$ an appropriate relation between R_p and S_n, and Eq. 9.18 gives a ϕ^* of the form

$$\phi^* = \frac{A}{[1 - B(S_n^{-1/2})e^{-C/kT}]^{1/2}} \qquad 9.19$$

where A, B, and C are constants that depend on the implant species and energy.

As the total implant dose increases, the critical value will be reached first at the peak of the distribution (at a depth R_p), and a narrow amorphous band will form. Further implanting will cause the band to increase in width, and eventually it will reach the surface. On either side of the amorphous band, a thin band of relatively undamaged material is found, and next to it is a band containing heavy damage (30).

An amorphous layer can be detected by any of the usual methods such as X-ray and electron diffraction and Rutherford backscattering. In addition, the optical properties of amorphous silicon are different enough from those of single-crystal silicon that an amorphous layer will appear cloudy. Enough reflection even occurs at the amorphous–crystalline interface to sometimes produce observable interference color (31). In visually determining amorphization, it should be kept in mind that the onset of cloudiness does not necessarily mean that the whole layer is amorphous. It has been suggested that full amorphization requires nearly a 10× larger dose (32).

When light doses such as those required for threshold adjust are implanted, the lattice damage is easily removed by annealing. When very heavy doses are used and amorphization occurs, the amorphous layer can be epitaxially regrown to give good crystalline quality silicon. For intermediate dose ranges, it is sometimes difficult to remove the damage, and often a preamorphizing step using, for example, implanted silicon, will be used.

9.3.2 Annealing

Annealing requires temperatures of above ~600°C for silicon and above 800°C or 900°C for gallium arsenide. Silicon is quite stable at high temperatures, but gallium arsenide begins to dissociate above ~600°C and thus requires a cap layer to prevent arsenic loss (33). Layers of silicon nitride and CVD silicon oxide can be used, but because of differences in their expansivity and that of gallium arsenide, they sometimes crack. Annealing can also be done in enclosed compartments with an increased partial pressure of arsenic. By laying a slice of gallium arsenide directly on top of the wafer being annealed, much of the vaporization can be prevented without the inconvenience of either a closed chamber or an added CVD cap film. The partial pressure of arsenic over a molten layer of tin on a gallium arsenide slice is higher than it is over uncoated gallium arsenide, and it is thus possible to use such coated slices in close proximity to gallium arsenide wafers to further increase the arsenic pressure (34).

Two heating regimes are of interest for annealing. In one regime, the heating is slow enough that annealing is substantially isothermal. The two ways used to accomplish this isothermal annealing are by furnace annealing over a period of many minutes or by a rapid thermal anneal (35, 36) lasting for a few seconds. In either case, however, the time is considerably longer than the thermal time constant τ of the semiconductor[6] (37). In the other regime, the heating time is comparable to or less than τ and only a thin layer is heated.

[6] The thermal time constant τ is somewhat analogous to the electrical RC time constant and arises, for example, from the approximate solution of the equation for heat flow from each face of a slab (a wafer) of thickness d to the middle of the slab. If the temperature of the faces is suddenly increased by an amount T, the first-order approximation of the temperature difference ΔT between the faces and the middle of the slab is given by $\Delta T = Ae^{-t/\tau}$, where A is a constant and $\tau = d^2/\pi^2 D_T$. D_T is not the diffusivity discussed in Chapter 8, but rather the *thermal* diffusivity, which equals $k/\rho c$ where k is the thermal conductivity of the slab, ρ is its density, and c is its specific heat. In some literature, the π^2 is dropped and τ is defined as d^2/D_T.

EXAMPLE □ Calculate τ for a 75 mm diameter silicon wafer that is 20 mils thick.

As long as the diameter is much greater than the thickness, its value plays no part in the calculation. The thickness is 0.051 cm and is much less than the 7.5 cm wafer diameter. $D_T = k/\rho c$ and $\rho = 2.33$ grams/cc. Since k and c are functions of temperature, the temperature should have been specified. Assume 1000°C, in which case $k = 0.29$ W/(cm·K) and $c = 1.8$ cal/(gram·°C). Since these units are not compatible, express k in terms of cal/(s·cm·°C), or $k = .07$ cal/(s·cm·°C). D_T then equals 0.17 cm²/s. Using the expression $\tau = d^2/\pi^2 D_T$, $\tau = 1.5$ ms. Thus, a rapid anneal source lasting a few seconds will heat the wafer throughout. However, laser pulses lasting a microsecond or less will heat only a thin surface layer. □

In some cases, surface melting may be allowed so that regrowth is by liquid phase rather than solid phase epitaxy. Generally, the activation energy for dopant diffusion is less than for implant defect annealing, so when it is necessary to minimize diffusion, higher-temperature, shorter-time anneals are preferred. However, in the extreme case where annealing is done by rapid melting and cooling of a thin layer, the impurities will be redistributed throughout the melted region. To adequately remove defects when this sort of annealing is used, melting well past the peak impurity distribution and down into the inner band of defects may be necessary. This procedure, unfortunately, will allow the impurities to redistribute to a much greater depth than that of the original implant. Even when a nonmelting, rapid anneal cycle is used, more diffusion than predicted by the standard diffusion equations is sometimes observed and is apparently due to diffusion enhancement from point defects liberated during the anneal (38).

Depending on the anneal sequence, it is possible to grow large extended defects from the implant-induced point defects during the annealing process. In particular, high-concentration implants, followed by a high-temperature anneal coupled with simultaneous oxidation, can lead to a high density of stacking faults and dislocation loops. A low-temperature—for example, 800°C—initial oxidation for a few minutes will reduce defects observed after a follow-on high-temperature oxidation (as required for many drive-ins after implantation) by orders of magnitude (39).

The conditions for optimum activation of dopant atoms and defect impurity will depend on the implant species, implant energy, and dose. In the case of silicon, low-dose boron implants can be properly annealed in the 900°C range, but heavier ion implants will require temperatures of over 1000°C. For a rapid thermal anneal time of 2 s, a temperature of 1100°C is required to fully activate and remove damage from a 10^{16} ion dose of arsenic implanted at 80 keV

(35). For gallium arsenide, a furnace anneal of 15–20 minutes at 850°C–900°C is typically used. For a given set of implant conditions, the amount of implanted impurity activated often goes through a peak, giving rise to an "optimum" anneal point. Optimum rapid thermal anneals, using incoherent heat sources, last a few seconds with maximum temperatures ranging from 800°C to over 1000°C, depending on the implant dose and species. The fraction of dopant activated is in the range of from 50% to 90%.

9.4 APPLICATIONS

Fig. 9.19 shows typical implant dose ranges for the various applications to be considered. The required dose varies over at least 7 orders of magnitude, and to satisfactorily cover this range, medium-current machines with beam currents up to about 1 mA and high-current implanters with currents up to about 20 mA are typically used. However, new requirements for implanting buried layers of oxygen and nitrogen have led to high-current versions designed to operate near 100 mA. How the beam current relates to the dose requirements shown in Fig. 9.19 depends on the time allowed for implanting each wafer. For singly charged ions (the usual case), each

FIGURE 9.19

Typical implant dose ranges for various silicon IC applications.

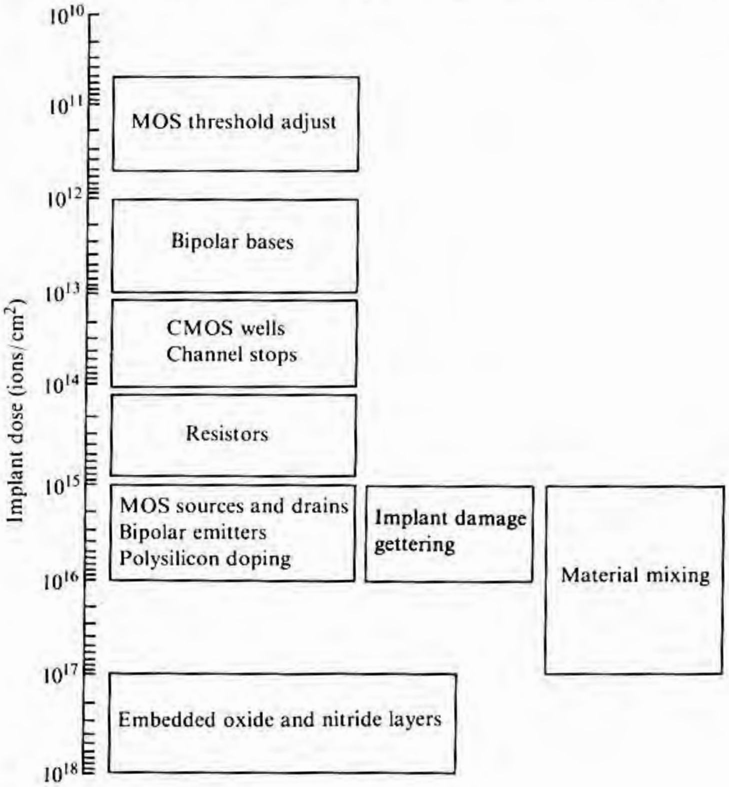
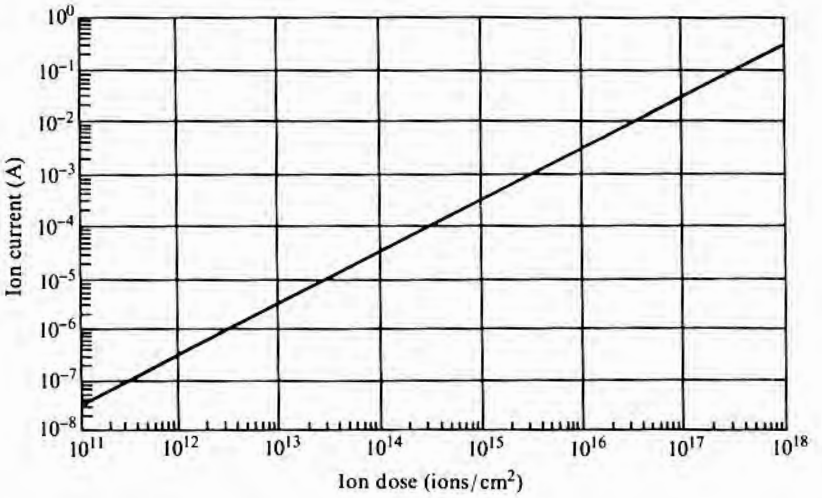


FIGURE 9.20

Ion current versus ion dose for implanting a single 125 mm diameter wafer in 1 minute with a singly charged ion.



ion has the charge q of one electron—that is, 1.6×10^{-19} coulombs (C). Thus, if ϕ ions per cm^2 are to be implanted, for a wafer of area A cm^2, $qA\phi$ coulombs are required. If the implant is to be accomplished in t seconds, $qA\phi/t$ coulombs per second are required. Since current in amperes is coulombs per second, the beam current I is

$$I \text{ in mA} = \frac{1.6 \times 10^{-16} A\phi}{t} \qquad 9.20$$

For the specific case of implanting a single 125 mm diameter wafer in 1 minute,[7] Fig. 9.20 gives a plot of dose versus beam current. This figure shows that for the low flux (dose) range, very little beam current is required, even when the implant time is reduced substantially from one minute per wafer. It also shows that for the high doses required for oxygen implants, even with 100 mA of beam current, throughput is quite low.

Another method of categorizing applications and machines is by low ion accelerating voltage[8] (~5–50 kV), medium voltage (~50–200 kV), and high voltage (300 kV to greater than 1 MV). The bulk of silicon applications fall in the medium-voltage range, but very shallow implants require low voltage. In order to substitute implanting for diffusion in gallium arsenide, high implant energy is required to get the necessary depths. High-energy beams also allow deeper

[7] This does not equate to implanting 60 wafers an hour since load time and perhaps some overscan to ensure that the wafer edges are completely covered must also be considered.

[8] Alternatively, by ion energy—for example, 50 keV.

amorphization in silicon, as well as the direct implantation of other structures such as deep retrograde CMOS wells.

9.4.1 MOS Threshold Adjust

Despite the fact that semiconductor device manufacturing applications for ion implanting have been discussed at least since 1952 (40), the first commercial application seems to have been that of MOS transistor threshold adjust begun by Mostek[9] in 1970. The threshold voltage V_t is given by

$$V_t = \phi_{MS} + \phi_B - \frac{Q_f}{C_{\text{ox}}} - \frac{Q_B}{C_{\text{ox}}} \qquad 9.21$$

where ϕ_{MS} is the metal–semiconductor work function,[10] ϕ_B is the potential barrier, Q_f is the fixed positive oxide charge, Q_B is a function of the substrate doping, and C_{ox} is a parallel-plate capacitance per unit area given by $C_{\text{ox}} = \varepsilon\varepsilon_0/w$ where w is the oxide thickness, ε is the dielectric constant of the oxide (~3.9), and ε_0 is the permittivity of free space (8.85×10^{-14} F/cm). If an additional charge Q_I is added just below the oxide–semiconductor surface by ion implantation (41–44), V_t will change by an amount

$$\Delta V_t = \frac{Q_I}{C_{\text{ox}}} \qquad 9.22$$

The implanted impurities that do not travel all the way through the oxide will not be ionized. Those going through the oxide and far enough into the semiconductor to stop beyond the edge of the space charge region will be in the neutral portion of the semiconductor and will also not be ionized. The rest will be ionized and are the part that causes a threshold shift. Eq. 9.22 is rigorously true only when the implanted charge is at the semiconductor–oxide interface. The farther the charge is from the interface, the less effective it is (45) so that all of the charge implanted beneath the oxide but still within the channel and space charge region cannot be equally weighted in determining a threshold shift.[11] When the peak of the distribution is between the oxide–semiconductor interface and the edge of the

[9] The earliest patent on threshold adjust by ion implantation appears to be held by someone other than any of the early authors of scientific papers on the subject (as is often the case). See David P. Robinson et al., "Method of Making Semiconductor Devices," U.S. Patent 3,653,978, March 4, 1972, claiming priority to March 11, 1968.

[10] Note that the symbol ϕ is used here to denote something other than implant dose. Both meanings conform to commonly used terminology.

[11] The range suggested in order to simultaneously minimize implant dose and provide the maximum localization of impurities near the silicon surface is somewhat greater than the oxide thickness (24).

space charge region, calculations of the shift based on the total charge and the location of its centroid can be made.[12] However, the implant dose actually needed is generally determined experimentally, with the initial trial based on Eq. 9.22.

For silicon, boron, which is a p-dopant and produces negative ions and a negative threshold shift in p-channel accumulation devices, is normally used. If an n-type dopant were used instead, it would increase the threshold of a p-channel transistor since its positive charge would add to the already positive Q_f.

EXAMPLE □ Estimate the implant dose required to reduce a p-channel threshold voltage by 1 V if the gate oxide is 400 Å thick.

Assume that the implant voltage is adjusted so that the peak of the distribution occurs at the oxide–silicon interface. Thus, half of the implant goes into the silicon. Further assume that 90% of the implanted ions in the silicon are electrically activated by the annealing process used. These assumptions allow 45% of the implanted ions to be used for threshold adjusting. Also assume that all of the charge in the silicon is effectively at the silicon–oxide interface so that Eq. 9.22 can be used. Thus,

$$\Delta V_t = 1 \text{ V} = \frac{Q_t}{C_{ox}}$$

$$Q_t \text{ (in C/cm}^2) = \frac{1 \text{ V} \times 3.9 \times (8.85 \times 10^{-14} \text{ F/cm})}{4 \times 10^{-6} \text{ cm}}$$

$$Q_t = 86.5 \times 10^{-9} \text{ C/cm}^2$$

The number of singly charged ions required for that amount of charge is Q_t/q, where q is the electronic charge (1.6×10^{-19} C per electron). Thus, $N_t = 5.4 \times 10^{11}$ ions, and since this represents 45% of the ions implanted, the total dose required is 1.2×10^{12} ions per cm^2. □

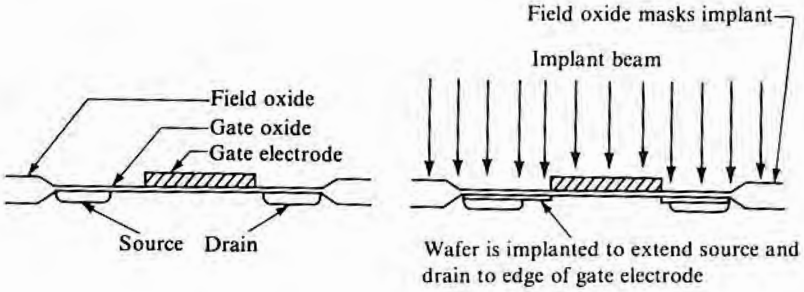
9.4.2 Self-Aligned MOS Gate

When the gate electrode appreciably overlaps source and drain, the extra capacitance degrades transistor performance. However, if the electrode does not reach to the source and drain, the uncovered portion of the channel cannot be turned on, and the transistor will not work. When the source and drain are defined in one step and the gate electrode in a later one, with the normal lithographic tolerances available, it is very difficult to exactly position the gate electrode so that no overlap or uncovered channel occur. To prevent this, the electrode can be deliberately made longer than the channel so that even with poor alignment no uncovered channel will occur. A better

[12] For details of these calculations, see, for example, S.M. Sze, *Physics of Semiconductor Physics*, 2d ed., John Wiley & Sons, New York, 1981.

FIGURE 9.21

MOS self-aligned gate process.


Field oxide
Gate oxide
Gate electrode
Source
Drain
Field oxide masks implant
Implant beam
Wafer is implanted to extend source and drain to edge of gate electrode

solution is the use of the self-aligned gate (SAG) process. In it, the source and drain regions are separated by more than the intended channel length so that the gate electrode positioning tolerance will allow it to be placed between them as shown in Fig. 9.21. The source and drain can then be extended up to the edge of the gate by using the gate electrode as the implant mask and implanting through the gate oxide that extends out from it (46). Alternatively, a somewhat heavier implant dose can be used, and the initial source–drain doping step can be eliminated. When polysilicon gates are used, the implant step can be used to dope the polysilicon as well.

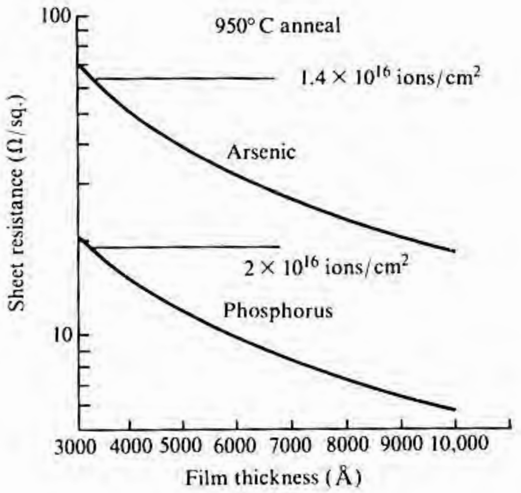
9.4.3 Polysilicon Doping

Since heavy doping of polycrystalline silicon during CVD is difficult, the usual procedure is to dope it later by either diffusion or ion implantation. The implant dose required depends on the dopant used, the thickness of the silicon layer, and the desired sheet resistance. To a lesser extent, there may be some dependency on the grain structure of the polysilicon and on the anneal temperature. If enough doping ions are implanted to meet or exceed the solubility limit for active doping, the sheet resistance R_s will decrease as the thickness is increased as shown in Fig. 9.22. While it would be expected that R_s would vary as $1/w$, where w is the thickness, experimentally it is observed (47) that $R_s \cong 1/(w - w_0)$ where w_0 is a constant dependent on the process and is in the range of 100–200 Å when the polysilicon is on an oxide. It is thought that w_0 represents a thin layer adjacent to the oxide that has a very low conductivity. If a fixed quantity of dopant is used and the polysilicon thickness is increased to the point that there is less dopant concentration than is required for electrical active saturation, the sheet resistance will stop decreasing, as shown by the straight lines in Fig. 9.22.

During thermal annealing after implanting, the dopant is activated and distributed uniformly through the thickness of the poly film by diffusion. In the case of impurities with high vapor pressures, such as arsenic, an oxide cap layer is required during annealing to prevent substantial loss of arsenic. Losses are higher for implants

FIGURE 9.22

Sheet resistance of arsenic- and phosphorus-implanted polysilicon. (The curved lines assume enough implant dose to saturate the polysilicon. The straight lines show the sheet resistance limit for the doses shown.) (*Source:* From data in N. Lifshitz, *J. Electrochem. Soc.* *130*, p. 2464, 1983.)



into polysilicon than for those into single crystal, presumably because of the increased diffusivity in polycrystalline material, which allows a more rapid arsenic diffusion to the surface (48). The anneal time required depends on implant species and the wafer temperature–time profile. As an example, with a 5×10^{15} arsenic implant into silicon, an anneal in a rapid anneal furnace set at 1200°C requires 10 s to reduce R_s to about 200 Ω/sq. and a further 15 s anneal to stabilize R_s at about 90 Ω/sq.

9.4.4 Alternative to Diffusion Predep

One of the most obvious uses of ion implantation is as an alternative to the diffusion predeposition step. However, ion implanting is usually more expensive than thermal diffusion, so unless there are clear-cut advantages in yield, throughput, or final device performance, there is no incentive to use it. Two advantages that have emerged are greater lateral doping uniformity over the wafer and lower-temperature processing. As discussed in Chapter 8, most predeps are done at a high temperature, and the solid solubility of the dopant is at that temperature, with diffusion time used to set the amount of dopant. By ion implanting, the total dopant added to the wafer can be controlled by the implant machine, and thus diffusion at high temperatures is not necessary. With the move to shallower diffusions, the lower temperatures are a distinct advantage.

9.4.5 Ion Beam Mixing and Ion Beam Damage

By irradiating an interface between two materials with a beam of ions before annealing, the interface can be smeared, and often the reaction between the components on either side of the interface will proceed at a lower temperature (49, 50). A major potential application is the reduced-temperature formation of various silicides useful as lead and contact materials.

An ion beam can also produce damage that may be used to advantage. As already discussed in Chapter 8, it can be used to provide gettering damage. Another use is amorphization, as discussed in this chapter. Finally, it can be used to increase the etch rate of, for example, silicon, silicon dioxide, and silicon nitride (51). Etch rate enhancements of from 2 to 5 can be obtained.

9.4.6 Formation of Buried Layers

Because of the high concentrations of ions that can be injected below the surface, it is possible, in principle, to form buried layers of a completely different composition. Examples are layers of silicon oxide and silicon nitride formed, respectively, by implanting oxygen and nitrogen into silicon and then annealing (52, 53). Such structures are possible alternatives to a thin silicon layer epitaxially deposited on single-crystal sapphire (SOS). By using SOS, devices built in the silicon layer are already partially electrically isolated because of the nonconductive sapphire substrate. Complete isolation is obtained by etching away the silicon layer as required between devices. Such total isolation allows very-high-speed, high-performance ICs to be built, but sapphire substrates are substantially more expensive than silicon slices.

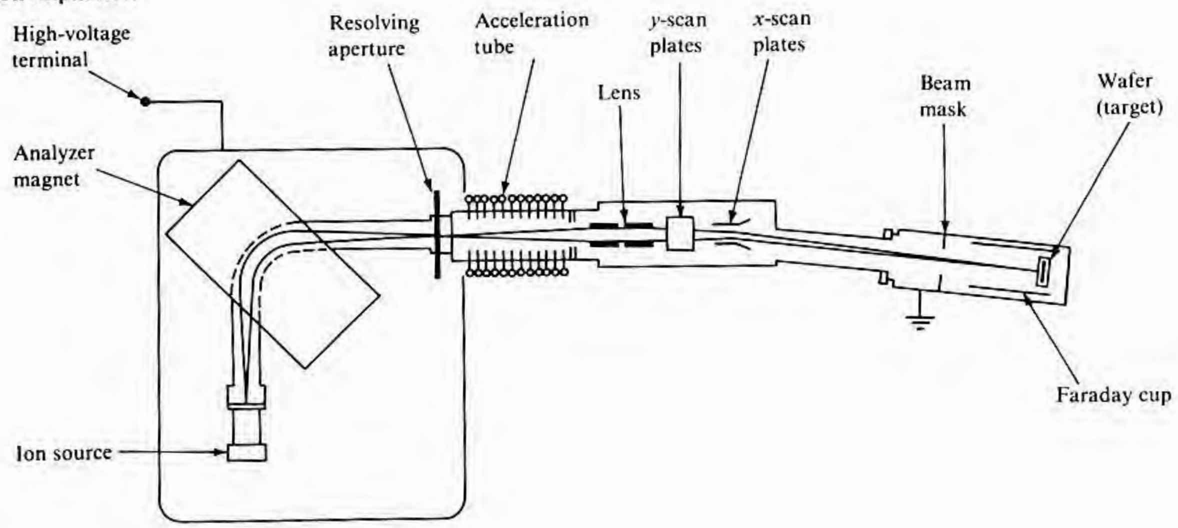
Silicon dioxide requires an implant dose of about 2×10^{18} atoms of oxygen per cm^2. Silicon nitride requires somewhat less ($\leq 7 \times 10^{17}$ atoms of nitrogen per cm^2). In either case, care must be taken to ensure that the amorphous silicon damage region that forms does not reach to the surface. If it does, no single-crystal silicon will be left above the implant region to seed regrowth during the annealing step. Without such single-crystal regrowth, only polycrystalline silicon will remain between the surface and the buried oxide or nitride. To minimize damage, implanting at temperatures around 500°C are used. Generally, the implant depth will not be enough to allow devices to be made in the thin single-crystal layer above the oxide or nitride, but an epitaxial layer can then be added to provide additional layer thickness (54).

9.5 ION IMPLANT EQUIPMENT

Fig. 9.23 is a simplified schematic that shows the major elements of a medium-current ion implanter. The ion source starts with an appropriate molecular species and converts it into ions. The ions are accelerated and then enter the mass analyzer for ion selection. The exit beam of desired implant ions is chosen based on the charge-to-mass ratio of the ions, and the analyzer is generally sensitive enough to discriminate against adjacent mass numbers. The ions are then given a final acceleration, after which the ion beam will be slightly

FIGURE 9.23

Schematic of a medium-current ion implanter.



electrostatically deflected to separate it from any neutral atoms that may have formed. The beam is then scanned over the wafer surface, either electrostatically, or mechanically, or by a combination of the two. In addition, an electron source may be near the wafers to "flood" the surface with electrons and prevent a charge buildup on insulating surfaces such as SiO_2 or silicon nitride (55). This kind of charge buildup can also occur on isolated conductors such as gate electrodes and can be severe enough to cause gate oxide failure because of electrical breakdown from the gate to the substrate through the oxide.

The ion beam is focused and deflected in the same manner as electron beams used in electron microscopy. Ion beam lenses, both convergent and divergent, can be made by appropriately shaping electron or magnetic fields. For example, as shown in Fig. 9.24, either a pair of closely spaced metal plates with holes in them or a pair of hollow metal cylinders placed end to end with a small gap between them can be used to focus ions. Acceleration of the ions is accomplished electrostatically while they travel down the high-voltage column. From a space standpoint, it is desirable to keep the column as short as possible, but the necessity of preventing electric arcing requires a minimum spacing between each electrode. Several electrodes are required in order to produce an electric field that will

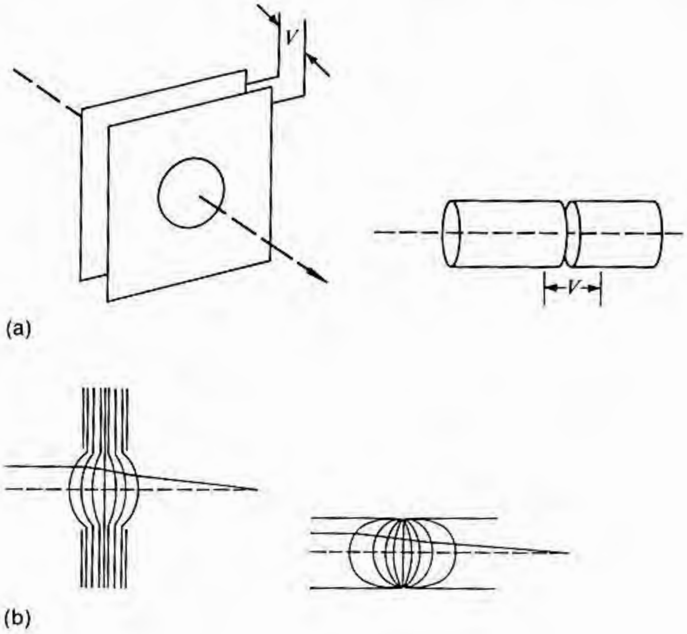


FIGURE 9.24

(a) Schematic of parallel-plate and hollow-tube electrostatic ion beam lenses. (b) Ion path through the two lenses.

not only provide ion acceleration but also keep the beam from substantially diverging while in the column.

There are two common ways to arrange the order in which the beam traverses the various elements just described. In both of them, however, the ions must have an initial accelerating voltage of at least a few kilovolts before entering the mass analyzer. In the preanalysis configuration (Fig. 9.23), most of the ion acceleration takes place after beam analysis. In the postanalysis configuration, the beam is accelerated to its final velocity and then analyzed. For very-low-voltage machines, preanalysis combined with a post-deaccelerating voltage may be necessary. A preanalysis machine can use a smaller magnet for analysis, but the analyzer section must operate at high voltage. With postanalysis, a larger magnet is required because of the higher ion velocity, but the analyzer can operate at ground potential (56, 57).

9.5.1 Ion Sources

The ion source supplies ions, usually singly positively charged, in enough quantity to provide beam currents of from 10 μA to 100 mA, depending on the rating of the implanter. Further, the source must be constructed so that the ions produced can be extracted and formed into a collimated beam. The species to be ionized, which may already be gaseous or be vaporized in or near the source, is confined in a chamber and ionized in a gaseous plasma by impact from electrons. The ions are then extracted through an opening in the chamber by the application of a negative voltage of several kilovolts between the ion confinement chamber and an extraction electrode. These two elements also form an electrostatic lens that helps focus the beam at the analyzer aperture. Additional lenses after the acceleration column are used to refocus the beam at the plane of the wafers.

In most cases, the electrons come from a hot filament (cathode), but occasionally, a cold cathode is used, and electrons are generated by secondary emission from positive ion bombardment. Also, a magnetic field is usually provided to cause spiraling of the electrons, thus increasing the length of path traveled before the electrons reach the anode and thereby improving ionization efficiency. Many different physical configurations have been used (58), with their relative advantages depending on such things as the physical form of the desired ion species, the amount of ions needed, and the ion generation rate stability required.

The Penning source is relatively simple and is often used for low beam currents. The anode is cylindrical and inside a chamber held at cathode potential. One end of the chamber is the cathode, and ions are extracted from a circular hole in it. The Freeman source is widely used in medium- and high-current machines. The chamber is at anode potential, and the ions are extracted from a slot in one wall. The cathode is a straight wire filament running parallel with and quite close to the slot. A magnetic field is directed along the length of the filament and slot. The Bernas source (based on the calutron source) also has the container at anode potential and extracts from a slot. However, the filament is located at one end of the container and near one end of the slot. As in the Freeman source, a magnetic field is parallel with the slot.

If the desired species is available as a gas, it can be fed directly into the source chamber. If it is available as a solid[13] with a reasonably high vapor pressure, a heated vaporizer immediately adjacent to the source chamber can be used to generate a source vapor. Since the source chamber pressure is $\sim 10^{-4}$ torr, it is feasible to have a sputtering capability built into the chamber. Thus, when the implan-

[13] Remember that elemental sources are not required since the implanter mass analyzer will separate out the desired atomic weight species from the rest of those present. Care must be taken, however, to ensure that a spurious complex ion of the same ionic weight as the desired ion is not formed—for example, $^{31}(CF)^+$ versus $^{31}P^+$.

tation of a material with a very low vapor pressure is required, a sputtering target of the desired material can be installed inside the source chamber, and sputtering can be used for vaporization. Table 9.3 lists some typical source materials.

9.5.2 Ion Analyzer

The ion analyzer works by injecting the ion beam into a magnetic field so that the beam is perpendicular to the field. In this case, the ions will travel in a circular path while remaining perpendicular to the field. The force exerted by the magnetic field is given by

$$f = nq(v \times \boldsymbol{B}) \tag{9.23a}$$

where q is the charge of an electron, n is the charge state of the ion, v is the velocity of the ion, and B is the magnetic field. For ion travel perpendicular to the field, the vector cross product of Eq. 9.23a can be expressed in scalar form as

$$f = nqvB \tag{9.23b}$$

This force, tending to make the ion travel in an inward spiral, is counteracted by a centrifugal force f_c given by

$$f_c = \frac{mv^2}{r} \tag{9.24}$$

TABLE 9.3

Possible Implant Source Materials

Implant Species	*Source*
Antimony	Sb_2O_3(s)*
Arsenic	AsF_3, AsH_3, GaAs(s)
Beryllium	Be(s), $BeCl_2$(s), BeF_2(s)
Boron	BCl_3, BF_3, B_2H_6
Cadmium	Cd(s), CdS(s)
Phosphorus	Red P(s), PCl_3, PF_3, PH_3, PF_5
Selenium	Se(s), CdSe(s), SeO_2(s)
Silicon	$SiCl_4$, SiF_4, SiH_4
Sulfur	S(s), SO_2, CdS(s), H_2S
Tellurium	Te(s), CdSe(s)
Tin	Sn(s), $SnCl_2$(s)
Zinc	$ZnCl_2$(s)

*Solid material. Unmarked entries are gaseous.

Note: Chlorine and fluorine compounds may cause source deterioration. Hydrogen may find its way into the rest of the vacuum system and prevent cryopumping.

Source: From information in reference 58 and in A. Axmann, "Ionizable Materials To Produce Ions for Implantation," *Solid State Technology*, p. 47, February 1975.

where m is the mass of the ion and r is the radius of the path. These two forces will be equal so that

$$nqB = \frac{mv}{r} \qquad 9.25$$

The velocity of the ion is given by

$$\frac{mv^2}{2} = nqV \qquad 9.26$$

where V is the voltage through which the ion was accelerated before reaching the analyzer. Note that the use of a doubly charged ion is equivalent to doubling the acceleration voltage. Substituting Eq. 9.26 into Eq. 9.25 and using a consistent set of units give

$$r = \frac{1}{B}\sqrt{\frac{2mV}{nq}} = \frac{144}{B}\sqrt{\frac{MV}{n}} \qquad 9.27$$

where r is in centimeters, M is the ion mass in atomic mass units (as given in Table 9.1), B is in gauss units, and V is in volts.

EXAMPLE □ If the apertures in the analyzer are set for a 10 cm radius and the ions are extracted from the source with 30 kV, what magnetic field is required to pass singly charged arsenic?

From Table 9.1, $M_{As} = 75$. Substituting this value of M, an r of 10, and an n of 1 (singly charged) into Eq. 9.27 gives

$$B = \frac{144}{10}\sqrt{75 \times 3 \times 10^4} = 21.6 \text{ kgauss}$$

Under these conditions, how far from the exit slit would germanium ions hit (how effective will the system be in separating the two ions)?

Assume the mass 74 Ge isotope. In this case, modifying Eq. 9.27 gives

$$\Delta r = \frac{144}{21.6 \times 10^3}[(75 - 74) \times 3 \times 10^4]^{0.5} = 1.2 \text{ cm}$$

To a first approximation, the separation of the two beams will equal Δr. Thus, if the slits were 5 mm wide, the two beams would just be separated. □

9.5.3 Wafer Handling

Wafer handling includes transporting the wafer into and out of the chamber, controlling the temperature of the wafer during implanting, and moving the wafer as necessary during implanting to ensure implant uniformity. In addition, provision must be made for adjusting the angle of implant in order to minimize channeling. Wafer scanning is usually done by a combination of a rotating multiple-slice wafer holder and radial beam deflection. The wafers can generally be automatically transferred from multiple-wafer cassette holders to the wafer holder. In some designs, the wafers enter the implant chamber individually and are then placed on the holder. In other designs, a holder is loaded with wafers in normal room atmosphere and then is transferred to the implant chamber. This design has the advantage of easily allowing dedicated wafer holders in order to minimize cross-contamination (see the following section). In either case, great care must be taken to minimize particulate contamination during the load/transfer operation. Even particles that do not initially land on wafers may be blown onto them later during venting.

9.5.4 Wafer Contamination

Wafer contamination arises from hydrocarbons that are adsorbed on the wafer and polymerized or carbonized by the beam, from material sputtered from fixturing by the ion beam redepositing on the wafer, and from an impure ion beam. Organic-based pump oils used in an improperly trapped vacuum system are a possible source of hydrocarbons. The use of perfluorinated polyether pump oil will minimize this problem (59). Cryopumping is an alternative, but only if the ion source feedstock does not produce appreciable hydrogen, as occurs, for example, when phosphine is used (60).

Some ion beam sputtering will occur wherever the ion beam strikes, and when the striking surface is other than the wafer itself, the sputtered material can be a source of wafer contamination (60, 61). Because some overtravel of the ion beam is necessary to ensure that the whole wafer surface is implanted, the beam will strike the metal fixturing a portion of the time. Therefore, metals such as stainless steel are usually avoided, and aluminum is used instead. In addition, dopant atoms are deposited on the fixturing by the implant beam and then subsequently sputtered away. Thus, cross-contamination from run to run is possible. However, by using dedicated fixturing for each dopant species, this effect can be eliminated. Dopant buildup can also be reduced by periodically making dummy cleanup runs. If implanting is always done through a layer that will be removed before heat treatment, then the layer will act as a shield for sputter contamination.

To determine the purity of the beam, the degree to which ions can be separated must be considered. In principle, ions with a single mass number (atomic weight) may be selected. However, most elements have several isotopes, and for those with higher atomic num-

TABLE 9.4

Atomic Weights for Natural Isotopes of Tin, Antimony, and Tellurium

Tin	*Antimony*	*Tellurium*
112(1.0)*		
114(0.7)		
115(0.4)		
116(14.7)		
117(7.3)		
118(24.3)		
119(8.6)		
120(32.4)		**120**(0.1)
	121(57.3)	
122(4.6)		**122**(2.5)
	123(42.7)	**123**(0.9)
124(5.6)		**124**(4.6)
		125(7.0)
		126(18.7)
		130(31.7)

*Numbers in parentheses give the percentage abundance of the isotope.

Source: Data from *Handbook of Chemistry and Physics*, 68th ed. (1987–1988), CRC Press, Boca Raton, Fla.

bers, some may overlap. For example, Table 9.4 shows the naturally occurring isotopes of tin, antimony, and tellurium. Thus, if a source were contaminated, beam contamination could also occur. The species ionized can also affect the implanted ion. For example, if molecular nitrogen were ionized to a +1 charge ($^{14}N_2^+$), it could be implanted with the same instrument setting as that for implanting $^{28}Si^+$

9.5.5 Implant Uniformity Measurements

The total number of ions implanted is tracked during implantation by integrating the beam current, but the uniformity over the wafer depends on the scan uniformity and is not readily amenable to in situ monitoring. One common way of measuring uniformity is by profiling the sheet resistance of an implanted layer after it has been annealed for activation (62). The results of a 1985 round robin test involving various semiconductor manufacturers indicated that sheet resistance values measured in the various facilities varied by no more than about 1% (63). It also showed that the actual sheet resistance, based on the measured implant dose, varied by several percent from facility to facility, thereby highlighting the fact that substantial difficulty can be encountered in accurately and reproducibly measuring dose. Another technique, applicable to low-value implants where the resulting sheet resistance is very high and difficult to measure, is to measure the threshold voltage of MOS transistors that have had their threshold shifted by the implant. Capacitance–voltage measurements on an array of MOS capacitors can also be used (64).

Rather than measuring the electrical properties of ions implanted in the semiconductor wafer, changes in the properties of the silicon itself can be used. The implant damage causes changes in optical properties and in thermal properties, both of which can be sensed. The change in optical transmissivity of silicon deposited on sapphire test wafers can be correlated with ion implant dose (65). In principle, ellipsometry and Raman spectroscopy (66) are also applicable, and both can be used to make observations from the top side of a wafer.

When thermal waves are introduced in the wafer by laser beam heating, ion-implant-induced damage can be detected (67, 68). The signal level can be related to implant dose and can be used to produce uniformity contours similar to those obtained from spreading resistance measurements. Such a technique is advantageous in that no annealing to electrically activate the impurities is required.

The change in optical transmissivity of some ion-sensitive coating other than silicon on a transparent test wafer can also be measured. Materials that have been used are photoresist (65) and nylon impregnated with a radiation-sensitive dye (69).

CHAPTER SAFETY 9

Since ion implanters use high voltages, care against electrocution is a major consideration. Hence, safety interlocks should never be circumvented unless required during machine servicing, and then only by qualified repair and maintenance personnel. Also, because of the high operating voltage, X rays are generated; therefore, shields must not be indiscriminately removed. The major source of X rays is electrons that become trapped in the ion accelerator column and are accelerated in the opposite direction so that they strike various metal parts in the vicinity of the source (or the analyzer in post-acceleration machines). As part of the initial installation of an implanter, a check of X-ray radiation should be required. However, even after that check, X-ray dosimetry badges should be worn by all personnel working near ion implant machines.

Highly toxic ion source gases are often used, and provision should be made to continuously monitor for leaks. The supply is usually limited to small lecture bottles, but even so, a leak could be very serious. (Refer to Chapter 4 for additional discussion on the handling of toxic gases.)

CHAPTER 9 KEY IDEAS

□ Ion implantation introduces impurities into the wafer by accelerating ions to a high velocity and directing them toward the wafer surface.

□ The depth distribution of implanted impurities in an amorphous material is roughly Gaussian, with the center of the distribution at a depth (range) determined by the implant energy.

□ The range is generally less than 1 μm, but with light ions and implanters operating at several MeV, it can be a few μm.

□ Implantation can be done at room temperature.

□ Implantation produces crystallographic damage. Very high doses cause enough damage to convert crystalline material to amorphous material. To remove damage and electrically activate the impurities, an anneal between 500°C and 900°C is generally required.

□ During implantation into crystalline materials, if the implant direction is along a major crystallographic direction, ion channeling may occur and substantially increase the depth of some of the ions.

□ Ion implantation can be substituted for shallow diffusions in most cases and can also be used for MOS transistor threshold adjusting by implanting through the gate oxide.

CHAPTER 9 PROBLEMS

1. If a 400 keV implanter is available and an n-type dopant is to be implanted into silicon, which dopant will give the most range? What will be the range if the ion used is doubly charged? (Assume no channeling.)
2. What will the straggle be when implanting arsenic at 300 keV into silicon? Assuming no moment other than the second, plot the impurity distribution.
3. If a p-type silicon wafer of 1 Ω-cm resistivity is to have localized low-resistivity n-regions (10^{19} atoms/cc peak concentration) implanted, how thick should a resist masking layer be if the implant is phosphorus done at 200 keV? List the assumptions made.
4. If a silicon wafer is held at 100°C, estimate the dose of implanted silicon required for amorphization? What is a major advantage of using silicon as the implant species?
5. A 150 mm wafer is to be implanted in 1 minute with a dose of 10^{13} ions/cm^2 of a doubly charged ion. What beam current will be required?
6. An n-channel silicon MOS transistor needs to have its V_t adjusted downward by 2 V. The oxide is 500 Å thick. What is a suitable ion for implanting, and what will be the approximate dose required for V_t correction?
7. If a 3000 Å layer of SiO_2 is to be formed inside a silicon wafer, what is the minimum number of oxygen ions that must be implanted? List the assumptions made.
8. If the magnetic field used for deflection is the same in each case, how much greater will the radius of curvature in an analyzer be when 1000 keV rather than 100 keV ions are used?

CHAPTER 9 REFERENCES

1. J. Linhard, M. Scharff, and H. Schiøtt, "Range Concepts and Heavy Ion Ranges," *Mat. Fys. Medd. Dan. Vid. Sclsk. 33*, pp. 1–39, 1963.
2. James F. Gibbons, "Ion Implantation in Semiconductors—Part I, Range Distribution Theory and Experiments," *Proc. IEEE 56*, pp. 295–319, 1968.
3. F.H. Eisen, "Channeling of Medium Mass Ions through Silicon," *Can. J. Phys. 46*, pp. 561–572, 1968.
4. James Comas and Robert G. Wilson, "Channeling and Random Equivalent Depth Distributions in 150 keV Li, Be, and B Implanted in Si," *J. Appl. Phys. 51*, pp. 3697–3701, 1980.
5. I.M. Cheshire et al., "The Z_1 Dependence of Electronic Stopping," *Physics Letters 27A*, pp. 304–305, 1968.
6. J.P. Biersack and J.F. Ziegler, "The Stopping and Range of Ions in Solids," pp. 122–156, in H. Ryssel and H. Glawischnig, eds., *Ion Implantation Techniques*, Springer-Verlag, New York, 1982.
7. James F. Gibbons, William S. Johnson, and Steven W. Mylroie, *Projected Range Statistics, Semiconductors and Related Materials*, 2d ed., Dowden, Hutchinson and Ross, Stroudsburg, Pa., 1975.
8. B. Smith, *Ion Implantation Range Data for Silicon and Germanium Device Technologies*, Research Studies, Forest Grove, Oreg., 1977.
9. J.P. Biersack and J.F. Ziegler, "The Calculation of Ion Ranges in Solids with Analytic Solutions," in H. Ryssel and H. Glawischnig, eds., *Ion Implantation Techniques*, Springer-Verlag, New York, 1982.
10. W.A. Shewhart, *Economic Control of Quality of Manufactured Product*, D. Van Nostrand Co., New York, 1931.
11. James F. Gibbons and Steven F. Mylroie, "Estimation of Impurity Profiles in Ion Implanted Amorphous Targets Using Joined Half-Gaussian Distributions," *Appl. Phys. Lett. 22*, pp. 568–569, 1973.
12. W.K. Hofker, *Philips Research Rpts., Suppl. 8*, 1975.
13. Heiner Ryssel, "Range Distributions," pp. 177–205, in H. Ryssel and H. Glawischnig, eds., *Ion Implantation Techniques*, Springer-Verlag, New York, 1982.
14. Seijiro Furukawa et al., "Theoretical Considerations on Lateral Spread of Implanted Layers," *Jap. J. Appl. Phys. 11*, pp. 134–142, 1972.
15. H. Runge, "Distribution of Implanted Ions under Arbitrary Shaped Mask Edges," *Phys. Stat. Sol. (a) 39*, pp. 595–599, 1977.
16. Seijiro Furukawa and Hideka Matsumura, "Backscattering Study on Lateral Spread of Implanted Ions," *Appl. Phys. Lett. 22*, pp. 97–98, 1973.
17. James W. Mayer et al., *Ion Implantation in Semiconductors*, Academic Press, New York, 1970, p. 40.
18. I. Suni et al., "Effect of Preamorphization Depth on Channeling Tails in B^+ and As^+ Implanted Silicon," in B.R. Appleton et al., eds., *Ion Beam Processes in Advanced Electronic Materials and Device Technology*, Vol. 45, Materials Research Society, 1985.
19. Karen W. Brannon and R.F. Lever, "Computational Investigation of Channeling of Boron in Silicon," *Electrochem. Soc. Ext. Abst. 86–2*, pp. 813–814, October 1986.
20. J.F. Ziegler, "The Channeling of Ions near the <100> Axis," in B.R. Appleton et al., eds., *Ion Beam Processes in Advanced Electronic Materials and Device Technology*, Vol. 45, Materials Research Society, 1985.
21. Masayasu Miyake et al., "Incidence Angle Dependence of Planar Channeling in Boron Ion Implantation in Silicon," *J. Electrochem. Soc. 130*, pp. 716–719, 1983.
22. Norman L. Turner et al., "Effects of Planar Channeling Using Modern Ion Implantation Equipment," *Solid State Technology*, pp. 163–172, February 1985.
23. Robert G. Wilson et al., "Angular Sensitivity of Controlled Implanted Doping Profiles," *NBS Special Pub. 400–49*, November 1978.
24. H. Ishiwara et al., "Projected Range Distribution of Implanted Ions in a Double-Layer Sub-

strate," pp. 423–428, in S. Namba, ed., *Ion Implantation*, Plenum Press, New York, 1975.

25. T.O. Herndon et al., "Polyimide for High Resolution Ion Implantation Masking," *Solid State Technology*, pp. 179–183, November 1984.
26. D. Roche, "Outgassing of Photoresist during Ion Implantation," *Proc. Materials Research Society Symposium 45*, pp. 203–210, 1985.
27. James F. Gibbons, "Ion Implantation in Semiconductors—Part II, Damage Production and Annealing," *Proc. IEEE 60*, pp. 1062–1096, 1972.
28. Sorab K. Ghandhi, *VLSI Fabrication Principles*, John Wiley & Sons, New York, p. 324, 1983.
29. F.F. Morehead, Jr., and B.L. Crowder, "A Model for the Formation of Amorphous Si by Ion Bombardment," pp. 25–30, in L. Chadderton and F. Eisen, eds., *Proc. 1st International Conference on Ion Implantation*, Gordon and Breach, New York, 1971.
30. J. Narayan and O.W. Holland, "Characteristics of Ion-Implantation Damage and Annealing Phenomena in Semiconductors," *J. Electrochem. Soc. 131*, pp. 2651–2662, 1984.
31. D.G. Beanland and D.J. Chivers, "Color-Band Generation during High Dose Ion Implantation of Silicon Wafers," *J. Electrochem. Soc. 125*, pp. 1331–1338, 1978.
32. James W. Mayer, Lennart Eriksson, and John A. Davies, *Ion Implantation in Semiconductors*, Academic Press, New York, p. 111, 1970.
33. F. Eisen et al., "Implantation into GaAs," pp. 117–144, in James V. DiLorenzo and Deen D. Khandelwal, eds., *GaAs FET Principles and Technology*, Artech House, Inc., Dedham, Mass., 1982.
34. C.A. Armiento et al., "Capless Rapid Thermal Annealing of GaAs Implanted with Si Using an Enhanced Overpressure Proximity Method," *J. Electrochem. Soc. 134*, pp. 2010–2016, 1987.
35. R. Kwor et al., "Effect of Furnace Preanneal and Rapid Thermal Anneal on Arsenic-Implanted Silicon," *J. Electrochem. Soc. 132*, pp. 1201–1206, 1985.
36. S.K. Tiku and W.M. Duncan, "Self-Compensation in Rapid Thermal Annealed Silicon-Implanted Gallium Arsenide," *J. Electrochem. Soc. 132*, pp. 2237–2239, 1985.
37. T.O. Sedgwick, "Short Time Annealing," *J. Electrochem. Soc. 130*, pp. 484–492, 1983.
38. K. Cho et al., "Transient Enhanced Diffusion during Rapid Thermal Annealing of Boron Implanted Silicon," *Appl. Phys. Lett. 47*, pp. 1321–1323, 1985.
39. D. Hagmann et al., "A Method To Impede the Formation of Crystal Defects after High Dose Arsenic Implants," *J. Electrochem. Soc. 133*, pp. 2597–2600, 1986.
40. R.S. Ohl, "Properties of Ionic Bombarded Silicon," *Bell Syst. Tech. J. 31*, pp. 104–122, 1952.
41. K.G. Aubuchon, "The Use of Ion-Implantation To Set the Threshold Voltage of MOS Transistors," pp. 575–593, in *Proc. International Conference on Properties and Use of MIS Structures*, Grenoble, France, 1969.
42. J. MacDougall, K. Manchester, and R. Palmer, "Ion Implantation Offers a Bagfull of Benefits for MOS," *Electronics 43*, pp. 86–90, June 1970.
43. M.R. MacPherson, "The Adjustment of MOS Transistor Threshold Voltage by Ion Implantation," *Appl. Phys. Lett. 18*, pp. 502–504, 1971.
44. R.B. Palmer et al., "The Effect of Oxide Thickness on Threshold Voltage of Boron Ion Implanted MOSFET," *J. Electrochem. Soc. 120*, pp. 999–1001, 1973.
45. S.D. Brotherton and P. Burton, "The Influence of Non-Uniformly Doped Substrates on MOS CV Curves," *Solid-State Electronics 13*, pp. 1591–1595, 1970.
46. Robert W. Bower et al., "MOS Field Effect Transistors Formed by Gate Masked Ion Implantation," *IEEE Trans. on Electron Dev. ED-15*, pp. 757–761, 1968.
47. N. Lifshitz, "Solubility of Implanted Dopants in Polysilicon: Phosphorus and Arsenic," *J. Electrochem. Soc. 130*, pp. 2464–2467, 1983.
48. S.R. Wilson et al., "Properties of Ion-Implanted Polycrystalline Si Layers Subjected to Rapid Thermal Anneal," *J. Electrochem. Soc. 132*, pp. 922–929, 1985.
49. B.Y. Tsaur et al., "Ion-Beam Induced Metastable Pt_2Si_3 Phase: I. Formation, Structure, and Properties," *J. Appl. Phys. 51*, pp. 5326–5333, 1980.
50. T.W. Orent et al., "Effects of Ion Implantation on the Thermal Growth of Pt and NiPt Silicides," *J. Electrochem. Soc. 130*, pp. 687–691, 1983.

51. A.H. van Ommen et al., "Etch Rate Modification of Si_3N_4 Layers by Ion Bombardment and Annealing," *J. Electrochem. Soc. 133*, pp. 2140–2147 and included references, 1986.
52. C.G. Tuppen and G.J. Davies, "An AES Investigation into the Phase Distribution of Ion-Implanted Oxygen in Silicon N-Channel Devices," *J. Electrochem. Soc. 131*, pp. 1423–1427 and included references, 1984.
53. L. Nesbit et al., "Microstructure of Silicon Implanted with High Doses of Nitrogen and Oxygen," *J. Electrochem. Soc. 133*, pp. 1186–1190 and included references, 1986.
54. W.W. Lloyd and R. Dexter, "Ion Implanted and Conventional Epitaxy to Produce Dielectrically Isolated Silicon Layers," U.S. Patent 3,855,009, December 17, 1974.
55. N. White et al., "Wafer Charging and Beam Interactions in Ion Implantation," *Solid State Technology*, pp. 151–158, February 1985.
56. Hans Glawischnig, "Ion Implantation System Concepts, pp. 3–21, in H. Ryssel and H. Glawischnig, eds., *Ion Implantation Techniques*, Springer-Verlag, New York, 1982.
57. Pieter Burggraaf, "Ion Implanters: Major 1986 Trends," *Semiconductor International*, pp. 78–89, April 1986.
58. D. Aitken, "Ion Sources," pp. 23–71, in H. Ryssel and H. Glawischnig, eds., *Ion Implantation Techniques*, Springer-Verlag, New York, 1982.
59. M.Y. Tsai et al., "Study of Surface Contamination Produced during High Dose Ion Implantation," *J. Electrochem. Soc. 126*, pp. 98–102, 1979.
60. G. Ryding, "Evolution and Performance of the Nova NV–10 Predep™ Implanter," pp. 319–342, in H. Ryssel and H. Glawischnig, eds., *Ion Implantation Techniques*, Springer-Verlag, New York, 1982.
61. L.A. Larson and M.I. Current, "Metallic Impurities and Dopant Cross-Contamination Effects in Ion Implanted Surfaces," *Proc. Materials Research Society Symposium 45*, pp. 381–388, 1985.
62. D.S. Perloff et al., "Dose Accuracy and Doping Uniformity of Ion Implantation Equipment," *Solid State Technology*, pp. 112–120, February 1981.
63. M.I. Current and W.A. Keenan, "A Performance Survey of Production Ion Implanters," *Solid State Technology*, pp. 139–146, February 1985.
64. R.O. Demming and W.A. Keenan, "Low Dose Ion Implant Monitoring," *Solid State Technology*, pp. 163–167, September 1985.
65. J.R. Golin et al., "Advanced Methods of Ion Implant Monitoring Using Optical Dosimetry," *Solid State Technology*, pp. 155–163, June 1985.
66. A.C. deWilton et al., "Raman Spectroscopy for Nondestructive Depth Profile Studies of Ion Implantation in Silicon," *J. Electrochem. Soc. 133*, pp. 988–995, 1986.
67. W. Lee Smith et al., "Ion Implant Monitoring with Thermal Wave Technology," *Appl. Phys. Lett. 47*, pp. 584–586, 1985.
68. W.L. Smith et al., "Ion Implant Monitoring with Thermal Wave Technology," *Solid State Technology*, pp. 85–92, January 1986.
69. Kranti V. Anand and Myron Cagin, "Fluence (Dose) Monitoring of Energetic H^+, B^+, N^+, P^+, and As^+ Ions Using Ionization in a Radiachromic Film," *J. Electrochem. Soc. 132*, pp. 1206–1208, 1985.

CHAPTER 10

Ohmic Contacts, Schottky Barriers, and Interconnects

10.1 INTRODUCTION

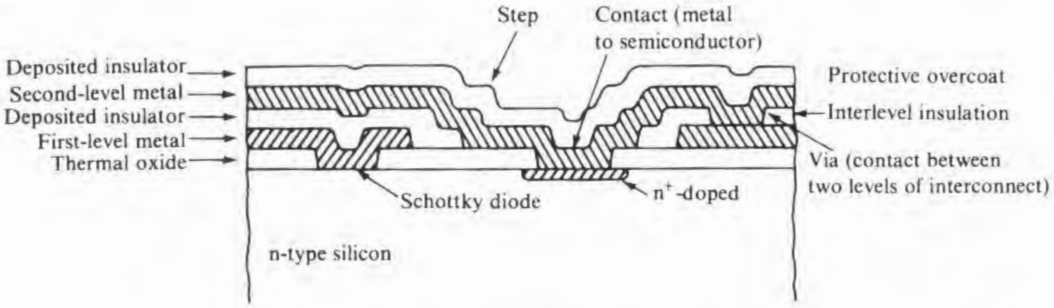
Before a circuit element formed in silicon or gallium arsenide can perform a useful function, it generally must be electrically connected with other elements on the same chip and most certainly must be connected with circuitry not on the same chip. An integrated circuit is by definition a number of electrically interconnected circuit elements on the same chip. Some of the interconnections are done in the silicon itself, but most are done by means of thin conductive stripes running across the top surface of the wafer. The connection to off-chip circuitry is included in packaging technology and will not be considered here.

Fig. 10.1 is a cross section of a small portion of a silicon wafer, showing a contact, a Schottky diode, and two levels of interconnections. The usual definition of a contact is metal contacting the semiconductor and providing a low-resistance ohmic electrical connection. What constitutes a low-resistance connection depends on the particular circuit it is used in, but generally it is a few microohms per square centimeter ($\mu\Omega/cm^2$) of contact area. When two levels of conductors contact each other at a via (hole in the separating insulator), a contact resistance will also be present. Its magnitude will primarily depend on how clean the first surface was before the second layer was applied. A Schottky barrier metal is one that contacts a semiconductor surface and forms one element of a Schottky diode. The interconnects are the electrical conductors connecting various contacts and Schottky diodes together so that they perform a useful function. The conductors must also be of low resistance, and, again, what constitutes low resistance depends on where they are used. In general, however, the resistance will be in the range of 25–100 Ω per centimeter of length.

In order to reliably fabricate a structure such as that of Fig. 10.1, interactions between the layers must be considered. In many cases, some of the functional layers shown in the figure will be comprised of layers of two or more different materials because no single

FIGURE 10.1

Cross section of a silicon IC interconnect system with two levels of metallization (not to scale).



universal contact and interconnect material exists that will work properly in all applications. As soon as more than one material is involved, interactions between materials must be considered, and the likelihood of deleterious interactions grows. Table 10.1 lists properties that must be considered in such a multiple-material interconnect system. The first three electrical properties are dictated by the required IC performance. All the other properties are related to reliability of the chip.

Aluminum was used almost from the beginning of silicon device fabrication for low-resistivity contacts,[1] and was extended to the

TABLE 10.1

Material Properties of Importance to an Interconnect System

Electrical	*Physical*	*Chemical*	*Purity*
Resistivity	Adhesion to SiO_2 or other insulators	Resistance to oxidation	Level of sodium
Contact resistance	Adhesion to adjacent conductive layers	Resistance to corrosion	Level of radioactive impurities
Schottky barrier height	Diffusion barrier efficacy	Reactivity with adjacent interconnect materials	
Amount of electromigration	Thermal expansivity		
	Stress induced during deposition		
	Surface morphology		

[1] Early grown junction silicon transistors (the kind first commercially available) used either electroless nickel plating or baked-on silver paste. (For a discussion of electroless plating on silicon, see, for example, Miles V. Sullivan and John H. Eigler, *J. Electrochem. Soc. 104*, p. 226, 1957.)

interconnect material as well when planar ICs were introduced. Aluminum has high electrical conductivity, makes good contact to silicon, is easy to deposit by evaporation or sputtering, and unlike some metals such as gold, adheres well to silicon dioxide. Unfortunately, some of its other properties, such as having only a modest resistance to electromigration, a propensity to grow spikes and puncture interlevel oxides, and a silicon alloy depth highly dependent on the silicon surface cleanliness, have prevented its complete applicability.

10.2 OHMIC CONTACTS

The total resistance of a contact is comprised of two components in series. One is the bulk spreading resistance r_s of the semiconductor beneath the contact, and the other is the resistance r_c of the metal–semiconductor interface. Thus, the total resistance R_T of a contact is given by $r_s + r_c$. By convention, "contact resistance" means only r_c. The spreading resistance is always ohmic,[2] but the contact resistance often is not.

10.2.1 Spreading Resistance

For flat, circular contacts on a semi-infinite body of uniform resistivity ρ (see Fig. 10.2), the spreading resistance r_s is given by

$$r_s = \frac{\rho}{4a} \qquad 10.1$$

where a is the radius of the contact.

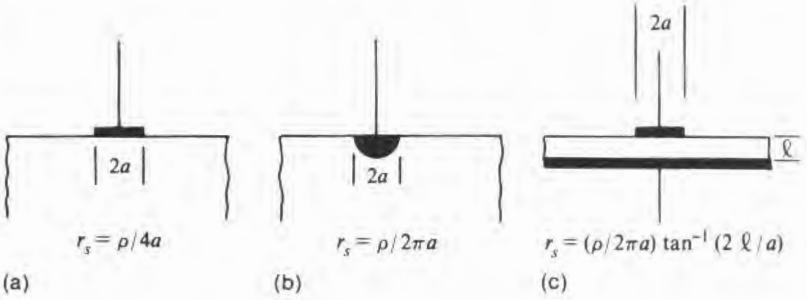


FIGURE 10.2

Geometries used in spreading resistance calculations.

[2] An ohmic contact is one in which the I–V curve is linear and symmetric about the origin.

If the contact is an imbedded hemisphere, then

$$r_s = \frac{\rho}{2\pi a} \qquad 10.2$$

For the more general case of the flat contact on a layer of finite thickness ℓ and resistivity ρ, Eq. 10.1 becomes (1)

$$r_s = \frac{\rho}{2\pi a}\tan^{-1}\frac{2\ell}{a} \qquad 10.3$$

For the case of $\ell << a$,

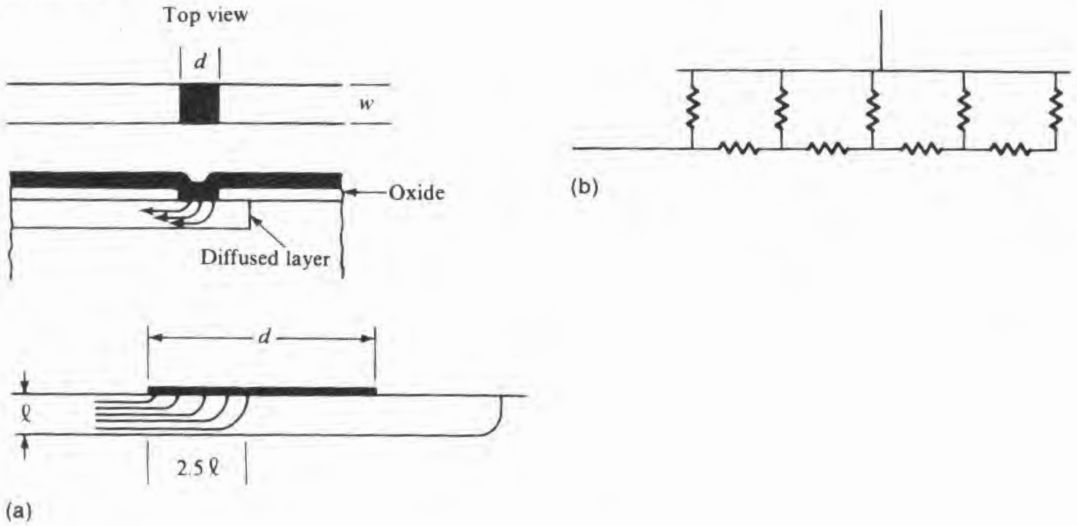
$$r_s = \frac{\rho\ell}{\pi a^2} \qquad 10.4$$

or the familiar $R = \rho\ell/A$ expression.

These cases are not exactly applicable to most IC geometries. In an IC, contact is generally made to a thin semiconductor layer isolated from the rest of the wafer by a pn junction. Thus, as shown in Fig. 10.3, the current must often turn and flow laterally in the thin section. This current will generally then become crowded near the leading edge of the contact (Fig. 10.3a). In such cases, increasing the length d of the contact past some value d' will not noticeably decrease the resistance. For the case in which the contact resistance r_c is negligible compared to the resistance of the layer (2), the dis-

FIGURE 10.3

(a) Current flow from a contact to a diffused lead when no contact resistance is present. (The layer thickness ℓ is given as $2\sqrt{Dt}$ and represents the distance not to the isolating junction but to the point where the doping concentration has decreased by a factor of ~1000.) (b) Transmission line equivalent circuit when contact resistance is present.



tance d' can be approximated by $5\sqrt{Dt}$, where Dt is the diffusivity–time product of the diffusion used to produce the diffused layer being contacted. In most cases, $2\sqrt{Dt}$ is a little less than the distance from the surface to the isolating junction, so d' is about 2.5 times the layer thickness (ℓ).

A derivation for the resistance, based on the assumption that r_c is not negligible, gives rise to a transmission line equivalent circuit as shown in Fig. 10.3b. For the case in which the contact covers the full width w of a thin layer, as was shown in Fig. 10.3a, the total resistance of the contact can be approximated by

$$R_T = \frac{1}{w}\sqrt{R_s A r_c}\,\coth\left(d\sqrt{\frac{R_s}{Ar_c}}\right) \qquad 10.5$$

where R_s is the sheet resistance of the layer and A is the area (wd) (3, 4). (For more expressions obtained by making various assumptions, see reference 4.)

10.2.2 Contact Resistance

Contact resistance arises from the fact that a metal in contact with a semiconductor surface forms a Schottky diode current barrier. The I–V relationship, like that of a pn junction, is, in the ideal case, given by

$$I = I_o(e^{qV/kT} - 1) \qquad 10.6$$

where I_o is a constant (the diode saturation current), q is the electronic charge, and k is Boltzmann's constant. The curve is nonlinear, but a diode differential resistance for any voltage can be defined as dV/dI. To be useful, an ohmic contact must have a low voltage drop and be usable for voltage swings in each direction. Hence, dV/dI evaluated at zero voltage is appropriate. If current density J is used instead of current, the ratio dV/dJ is referred to as the specific contact resistance R_c with units of Ω-cm^2. The contact resistance then becomes

$$r_c = \frac{R_c}{A} \qquad 10.7$$

where A is the contact area.

Fig. 10.4a shows a diode I–V trace for a swing of several volts, while Fig. 10.4b is an expanded scale covering less than ± 100 mV and shows that near the origin the contact is ohmic. $J(V)$ has the same form as $I(V)$ in Eq. 10.6 so that

$$R_c = \frac{dV}{dJ} = \frac{kT}{qJ_o} \quad \text{at } V = 0 \qquad 10.8$$

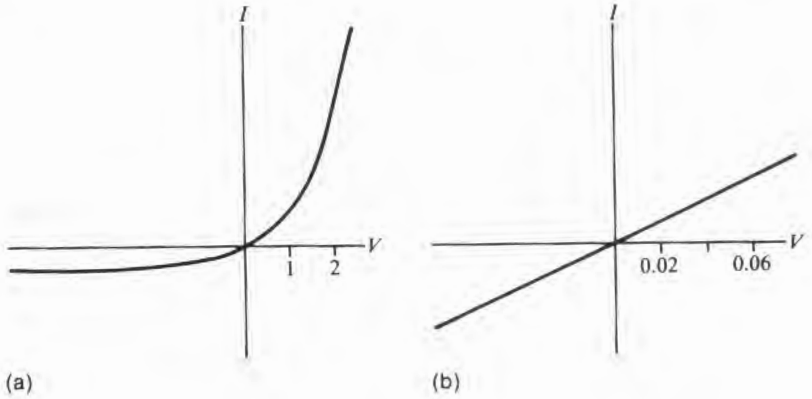


FIGURE 10.4

(a) Forward and reverse I–V curve of an ideal diode. (b) Middle portion of same curve with both scales expanded $50\times$. (The curve is quite linear for $-3kT/q < V < 3kT/q$.)

Current flow across a metal–semiconductor interface is primarily by thermionic emission when the doping level is in the 10^{17} atoms/cc and less range and J is given by

$$J = A^{**}T^2 e^{-q\phi_B/kT} \times (e^{qV/kT} - 1) \qquad 10.9$$

where A^{**} is the Richardson constant and ϕ_B is the Schottky barrier height in volts. The specific contact resistance near zero voltage is given by

$$R_c = \frac{dV}{dJ} = \left(\frac{k}{qA^{**}T}\right) e^{q\phi_B/kT} \qquad 10.10$$

Thus, the larger the barrier height, the larger R_c. Table 10.2 gives barrier heights for several metals on silicon and gallium arsenide. Table 10.3 gives values of R_c as calculated for silicon from Eq. 10.10 (5) and shows a range of 10 orders of magnitude as ϕ_B goes from 0.25 V to 0.85 V. As Table 10.2 shows, heights for p-type material are generally much lower than those for n-type, and it is therefore easier to make acceptable contacts to p-type material. It is clear from these tables that finding a metal for n-type silicon with a low enough barrier height to give an R_c in the useful range of $1\text{–}2 \times 10^{-6}$ $\Omega\text{-cm}^2$ is unlikely, although the lanthanides are close. From a practical standpoint, however, the R_c values can be much lower than those predicted from Eq. 10.10. Without considerable care in manufacturing, the diode saturation current I_o will be several orders of magnitude larger than theoretical and will reduce R_c. In the early days of the transistor, satisfactory contacts to etched 2 Ω-cm silicon were made with baked-on silver paste.

By increasing the doping level of the semiconductor, the space charge region width can be reduced to a thickness that allows sub-

TABLE 10.2

Metal–Semiconductor Schottky Barrier Heights

	Barrier Height in Volts			
Metal	*n-Si*	*p-Si*	*n-GaAs*	*p-GaAs*
Ag			0.93	0.44
Al	0.72	0.58	0.8	0.63
Au	0.8	0.35	0.95	0.48
Cr	0.61			
La	0.4	0.7		
Mo	0.68	0.42		
Pt	0.9		0.94	0.48
Ti	0.50	0.61		
W	0.67	0.45	0.77	
Y	0.4	0.7		

Note: A substantial spread in the values reported in the literature occurs, depending on the method of measurement, the surface treatment given the semiconductor before the metal was applied, and the specific heat treatment given the contact. There is also a small variation due to the semiconductor doping level.

TABLE 10.3

Calculated R_c for an n-Type Silicon Schottky Diode near $V = 0$

ϕ_B *(V)*	R_c *(Ω-cm²)*
0.85	4.7×10^5
0.70	1.4×10^3
0.55	4.3
0.40	1.3×10^{-2}
0.25	4×10^{-5}

Source: From data in L.P. Lepselter and J.M. Andrews, "Ohmic Contact to Silicon," in Bertram Schwartz, ed., *Ohmic Contacts to Semiconductors*, Electrochemical Society, New York, 1969.

stantial electron tunneling (as opposed to the electrons surmounting the barrier by thermal emission). This reduction in thickness provides for a much increased current flow and drop in specific contact resistance.

When the doping level is greater than about 10^{19} atoms/cc (6), rather than $R_c \propto e^{q\phi_B/kT}$,

$$R_c \propto e^{B\phi_B/\sqrt{N_D}} \qquad 10.11$$

$B = (4\pi/h)\sqrt{\varepsilon m^*}$ where h is Planck's constant, ε is the dielectric permittivity, and m^* is the effective mass of the charge carrier. The trend of R_c versus doping for silicon and gallium arsenide over both the thermionic emission and the tunneling regimes is shown in Figs. 10.5 and 10.6.

Experimentally, n-type silicon follows the theoretical prediction rather well. In the case of n-type gallium arsenide with alloyed germanium contacts, the data scatter is substantial, and the contact resistance decreases over a wide range of initial doping levels approximately as $1/N_D$ (7). For p-type silicon with alloyed aluminum contacts, a similar behavior is observed (8). The trends are shown in Fig. 10.7. The failure to show an $e^{q\phi_B/kT}$ dependence at low N_D values is presumably because the contact systems dope the material, regardless of its starting resistivity, to the point where tunneling dominates. For the case of gallium arsenide, it has been postulated that the $1/N_D$ rather than the $1/\sqrt{N_D}$ variation is the outgrowth of an uneven and spiked contact–GaAs interface (6). In such cases, the

FIGURE 10.5

Theoretical specific contact resistance as a function of silicon donor doping level. (*Source:* From data in C.Y. Chang et al., *Solid-State Electronics 14*, p. 541, 1971.)

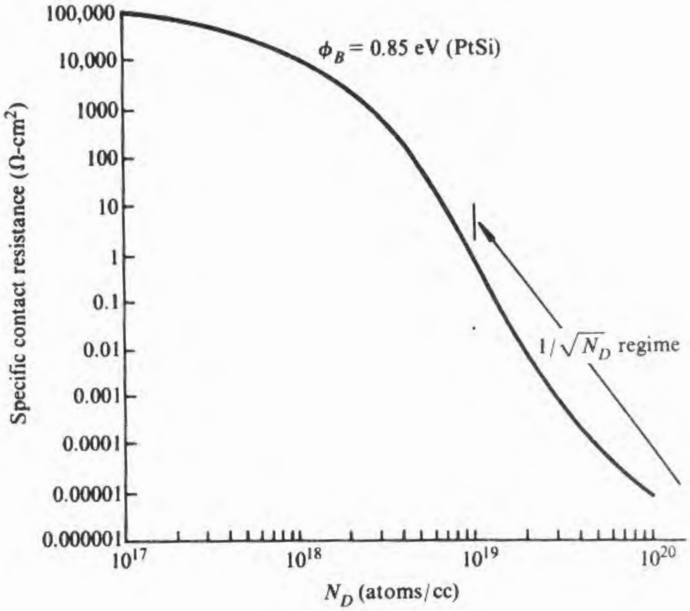
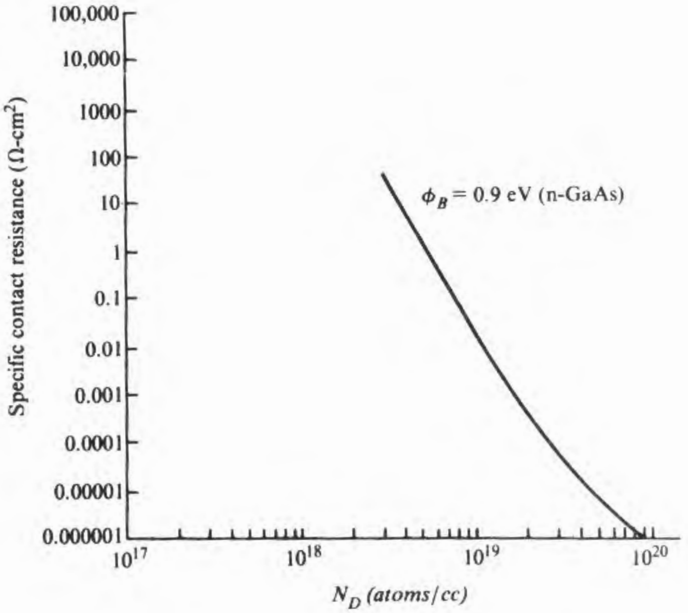


FIGURE 10.6

Theoretical specific contact resistance as a function of gallium arsenide donor doping level. (*Source:* From data in Gary Robinson, CEI, September 1986.)



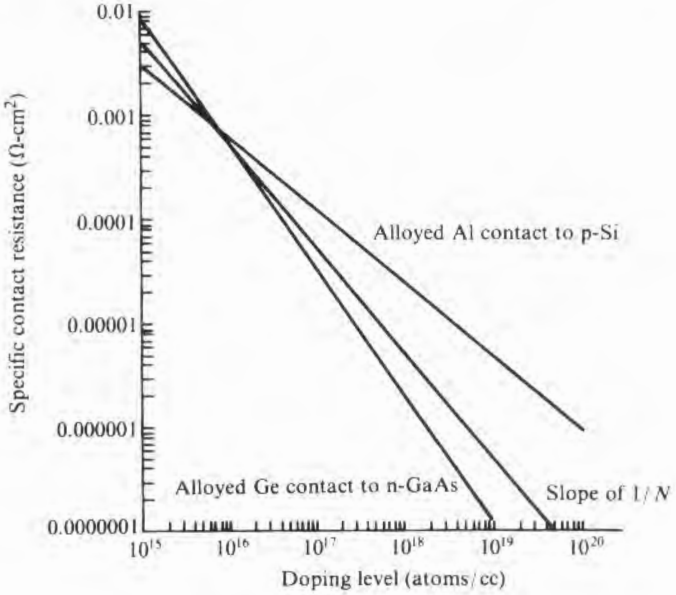


FIGURE 10.7

Experimentally observed trend of specific contact resistance versus doping level for alloyed contacts to silicon and gallium arsenide. (*Source:* Data from H.H. Berger, *J. Electrochem. Soc. 119*, p. 507, 1972, and N. Braslau, *J. Vac. Sci. Technol. 19*, p. 803, 1981.)

contact resistance would then be determined by the series of point contacts, which according to Eq. 10.2 would be proportional to ρ, which in turn is proportional to $1/N_D$.

Table 10.3 showed how R_c varies with barrier height when thermal emission dominates. However, since heavy doping and tunneling are actually used, it is more helpful to examine the effect of ϕ_B in the region where R_c varies as $1/\sqrt{N_D}$. For silicon, as shown in Fig. 10.8, R_c still decreases with decreasing barrier height, but the effect is much less pronounced.

An alternate approach to making ohmic contacts to a wide bandgap material such as gallium arsenide has been proposed in which a thin layer of a second, low bandgap material is added to the surface of the wafer (9). If the two doping levels are properly adjusted, the heterojunction formed should offer little resistance, and because the second material has a low bandgap, making ohmic contacts to it is relatively easy.

The theoretical specific contact resistance curves just shown are for n-type material and cannot be used for p-type material because of the effective mass difference between holes and electrons. m^* enters into the thermionic emission expression (Eq. 10.10) through the Richardson equation and will have only a multiplicative effect on R_c. However, it enters into the tunneling expression (Eq. 10.11) exponentially and hence will affect it much more.

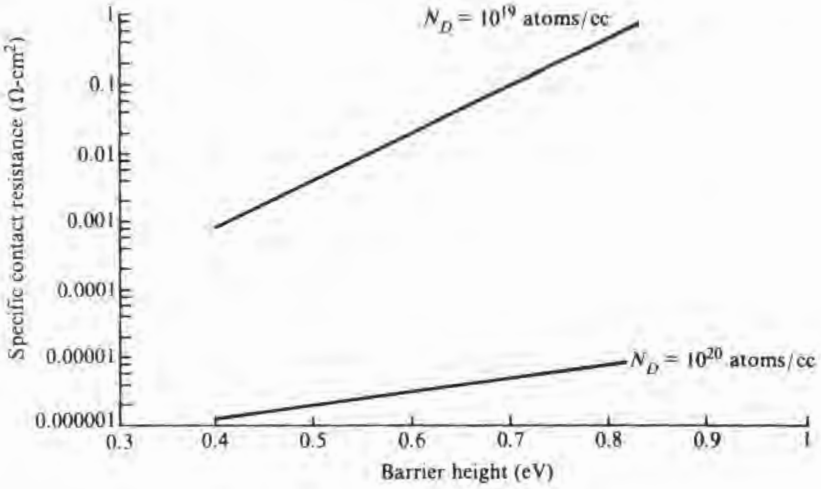


FIGURE 10.8

Theoretical trend of specific contact resistance as a function of Schottky barrier height for two silicon doping levels. (*Source*: Based on data in C.Y. Chang et al., *Solid-State Electronics 14*, p. 541, 1971.)

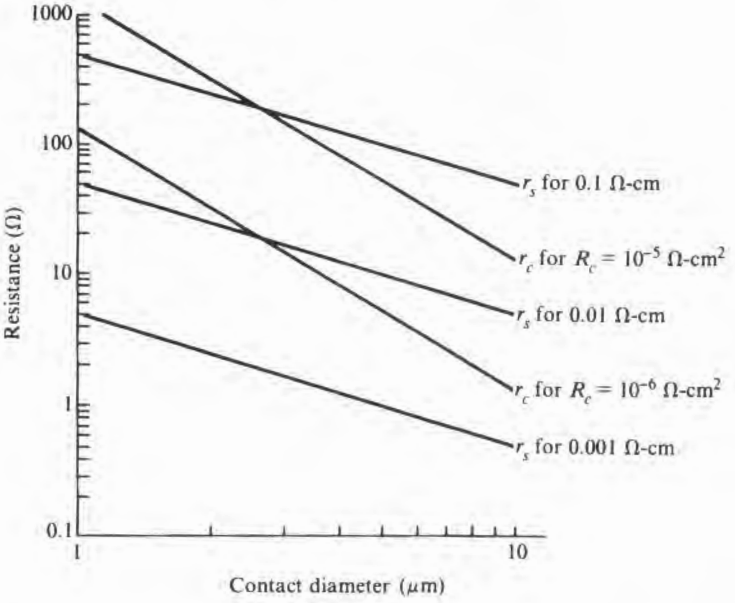
10.2.3 Total Contact Resistance

The value of total contact resistance R_T required in an integrated circuit varies with the specific device but is generally in the 10–100 Ω range. The R_c value that can be tolerated varies from 10^{-6} Ω-cm^2 to about 10^{-5} Ω-cm^2. Of concern as devices continue to decrease in size is how contact resistance must scale in order to not degrade the device. Bipolar behavior seems to dictate that the current density of emitters remain near 400 A/cm^2 regardless of emitter area. Since allowable circuit voltage drops will remain constant, independent of current and geometry, as the current decreases, the resistance can increase commensurately. Thus, the move to smaller bipolar geometries should not put additional restraints on contact resistance. In MOS circuits, however, reduced geometries may lead to reduced voltages as well as currents. In such cases, a constant resistance value during size reduction may be required, and that must come from a reduction of the contact resistance r_c.

To see how R_c and bulk resistivity ρ interact, consider Fig. 10.9. It has curves of contact resistance versus contact area for representative values of R_c and curves of spreading resistance versus area for different silicon doping levels. These curves show that for the total resistance to be in the <100 Ω range for the small-diameter contacts, the silicon doping must be 0.01 Ω-cm or less. The heavy doping is required for r_s as well as for R_c reduction. Current technology primarily uses locally diffused regions (p$^+$ or n$^+$) under the contact to secure the high doping necessary. However, the doping that occurs during regrowth from a metal-alloyed region can also be used. For example, regrowth from aluminum will give p-type aluminum-doped silicon, as will regrowth from gold to which small

FIGURE 10.9

Spreading resistance of circular contacts for various resistivities of material and contact resistance for two values of specific contact resistance.



amounts of gallium have been added. Similarly, gold with small amounts of antimony will regrow an antimony-doped n-region. Sandblasting has also been used to provide a damaged surface layer whose electrical properties are similar to those of heavily doped surfaces.

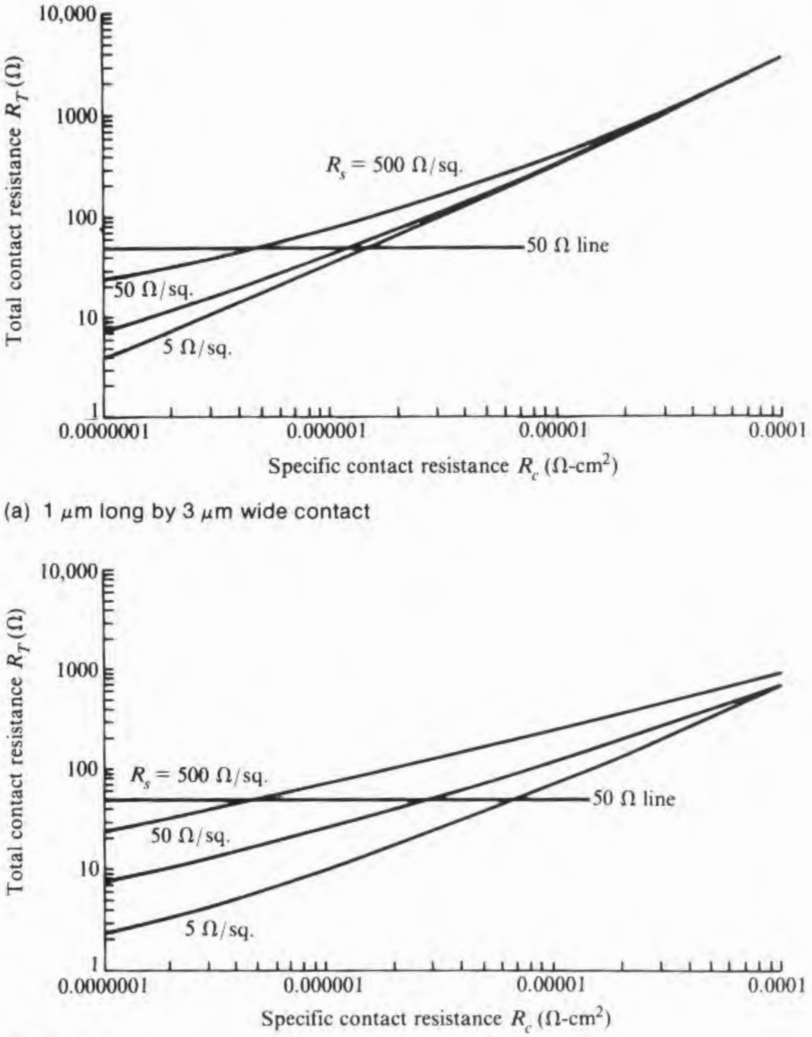
Perhaps more directly applicable to the understanding of these relations as they pertain to integrated circuits is a reconsideration of Fig. 10.3, which showed contact to a thin isolated diffused layer. The total contact resistance R_T for that geometry can be approximated by Eq. 10.5. Using that same equation, R_T versus specific contact resistance is plotted for representative sheet resistances of 5, 50, and 500 Ω/sq. and shown in Fig. 10.10. For example, with a contact 1 μm long by 3 μm wide, a contact resistance of less than 50 Ω can, for any of the three sheet resistance values shown, only be achieved if R_c is less than about 10^{-6} Ω-cm^2. However, with a 5 μm contact, 50 Ω can be obtained with a 5 Ω/sq. sheet and an R_c of just a little less than 10^{-5} Ω-cm^2. For this reason, as contacts have shrunk, more emphasis has been placed on reducing specific contact resistance.

10.2.4 The Aluminum–Silicon Ohmic Contact

Silicon has an appreciable solubility at elevated temperatures in several of the possible contact materials. Aluminum, which has been used since the inception of the planar silicon IC for contacts and metallization, is an example. Its ability to dissolve thin layers of

FIGURE 10.10

Specific contact resistance versus total contact resistance for various sheet resistance values.



(a) 1 μm long by 3 μm wide contact

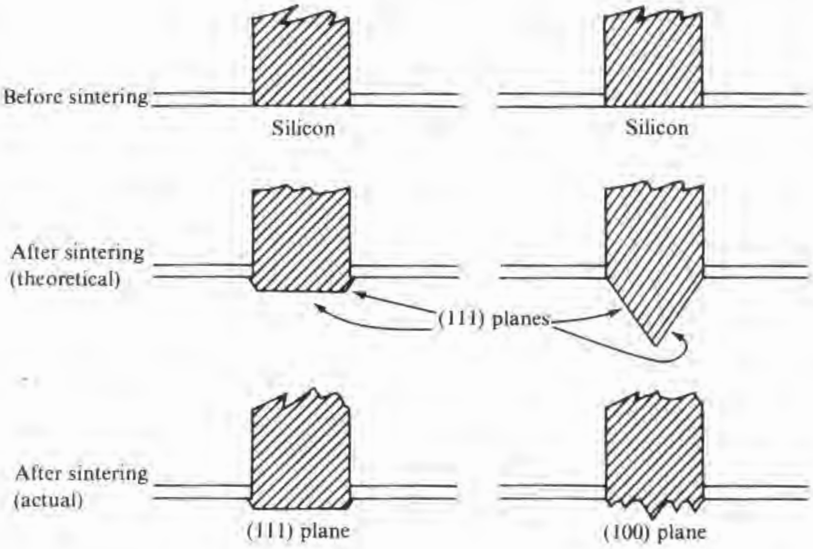
(b) 5 μm long by 3 μm wide contact

SiO_2 helps ensure good physical contact to the silicon even if surface cleaning is not complete. However, enough silicon will dissolve in the aluminum during the contacting operation to form pits in the silicon surface. The pits are filled with aluminum, and the phenomenon is often referred to as aluminum spiking. Fig. 10.11 shows the way the dissolution (alloying) takes place. This pitting is seen well below the aluminum–silicon eutectic temperature of 577°C. (See Appendix B for a discussion of solubility, eutectic temperature, and phase diagrams.) Some early work postulated that the eutectic temperature was depressed because of stress and that melting actually occurred. However, the high diffusivity of silicon in aluminum can account for the observed effect.

Like aqueous etching, the etching of silicon by metals is crystallographic orientation dependent, and as in aqueous etching, the slow etching planes are the {111}'s (10). Because of the slow dissolving (111)'s, it is possible to get very flat alloy fronts when alloying into (111) surfaces (11). Thus, germanium and silicon alloy transistors were all made in (111) oriented wafers. The equilibrium form of patterns formed by the aluminum–silicon interaction will appear as flat-bottomed, triangular pits in (111) and square-topped, pyramidal pits in (100) material. Occasionally, hexagonal shapes will be seen on (111) material because at early stages the ($\bar{1}11$), ($1\bar{1}1$), and ($11\bar{1}$) planes as well as the ($\bar{1}\bar{1}1$), ($1\bar{1}\bar{1}$), and ($11\bar{1}$) planes limit the reaction.[3] In cross section, they will appear as in Fig. 10.11. Since no (111) plane is parallel to a (100) surface to act as an etch stop, the aluminum–silicon interface of contacts on (100) wafers will not be smooth. Of more importance, because the thin layer of interfacial oxide that must first be dissolved will not fail uniformly, different regions will begin reacting at different times. The first ones can allow enough silicon into the aluminum to saturate much of the contact

FIGURE 10.11

Metal front before and after sintering (alloying). (The actual depth depends on sinter time, temperature, and amount of metal present.)



[3] It is arguable that the equilibrium shapes formed by solid-state diffusion might be different from those occurring during liquid dissolution, but experimentally, they are observed to be the same.

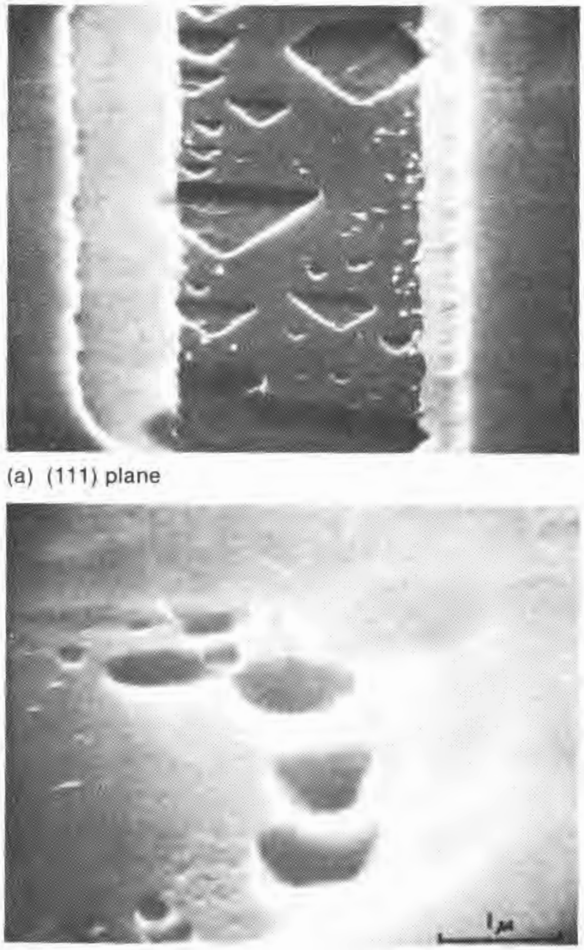
and thus limit dissolution over the remainder of the contact area. The unfortunate result is that the first pits will be much deeper than the rest and, in shallow junction devices, may extend through a junction and cause shorting. Fig. 10.12 shows scanning electron microscope photographs of (111) and (100) silicon surfaces after the aluminum contacts have been removed. The excessive pitting of the (100) surface is clearly visible.

A planar transistor or an integrated circuit has rather thin metallization, and the sintering temperature is intentionally held below the metal–silicon eutectic. Nevertheless, the solubility of silicon in aluminum is high enough for appreciable dissolution to occur (12–18). The total amount of silicon dissolved can be calculated from the

FIGURE 10.12

SEM views of (111) and (100) silicon surfaces showing pits remaining after removing aluminum metallization from contact windows.

(*Source:* Photographs courtesy of Dr. P.B. Ghate, Texas Instruments Incorporated.)

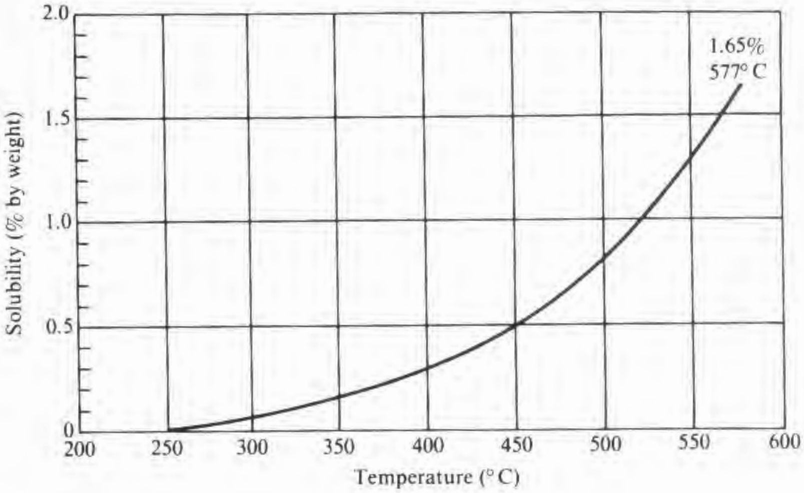


(a) (111) plane

(b) (100) plane

FIGURE 10.13

Solubility of silicon in aluminum. (*Source:* From data in Max Hansen, *Constitution of Binary Alloys*, McGraw-Hill Book Co., New York, 1958.)

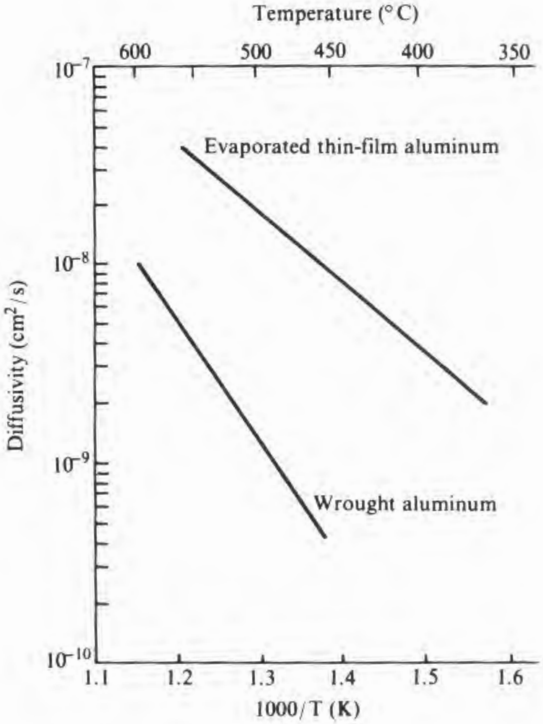


solid solubility of silicon in aluminum and the amount of aluminum available. Fig. 10.13 gives solubility data up to 577°C. Determining the amount of solvent aluminum is not completely straightforward, however. In principle, given enough time for the silicon to diffuse along its length, the entire aluminum lead system could become saturated. In practice, the time at high temperature is not enough for this to happen. Substantial diffusion distances are possible, however, because of the high silicon diffusivity (Fig. 10.14). The effect of leads acting as silicon sinks can clearly be seen by noting that when a contact has a lead extending out from it in one direction only, more dissolution generally occurs on that side of the contact. Since the removal of the silicon does not produce a void under the aluminum, it must be assumed that a simultaneous motion of the aluminum occurs (aluminum self-diffusion) that keeps voids from forming. In order to use the data of Fig. 10.14, it must be further assumed that it is the silicon diffusion, and not that of aluminum, that limits motion of silicon in the aluminum.

EXAMPLE □ If an aluminum lead the same width as the contact (10 μm) and 1 μm thick extends out from the contact to infinity in one direction only, how much silicon will diffuse into it in 20 minutes at 450°C? Assume that the aluminum over the contact is already saturated with aluminum and that it will remain saturated by diffusion from the wafer.

FIGURE 10.14

Diffusivity of silicon in aluminum. (The upper curve is representive of IC metallization; the lower curve for wrought aluminum is intended only as a trend line. The spread of reported data is substantial.) (*Source:* Upper curve data from J.O. McCaldin and H. Sankur, *Appl. Phys. Lett. 19,* p. 524, 1971; lower curve based on old data tabulated in McCaldin and Sankur.)



From Fig. 10.13, at 450°C, the solubility limit of silicon in aluminum is ~0.5% by weight, or 0.012 gram/cc. Let that equal N_0. $N(x) = N_0 \operatorname{erfc}(x/2\sqrt{Dt})$. (For details on diffusion, see Chapter 8.) The total amount M of silicon that diffused in is given by $\int_0^\infty N(x)dx$, or

$$M = \int_0^\infty \operatorname{erfc}\left(\frac{x}{2\sqrt{Dt}}\right) = 2N_0\sqrt{\frac{Dt}{\pi}}$$

From Fig. 10.14, at 450°C, $D \cong 10^{-8}$ cm^2/s. Substituting these numbers into the equation above gives $M = 4.8 \times 10^{-6}$ grams/cm^2. The cross section of the lead is 10^{-7} cm^2, so 4.8×10^{-13} grams of silicon diffused down the lead. Since silicon has a density of 2.3 grams/cc, this represents a volume of $\sim 2 \times 10^{-13}$ cc. If the contact were square, its area would be 10^{-6} cm^2; if the silicon were uniformly removed from beneath it, the surface would move down 2×10^{-7} cm. While this number may seem very small, if all the silicon came from one 1 μm square area, as might happen in a (100) wafer, the pit would be 0.7 μm deep. □

Experimentally,[4] it has been reported that sintering at 300°C produces discernible pitting with some depths to 0.2 μm. At 350°C, pits of 0.75 μm have been observed, and at 450°C, pits of 2 μm (19). Individual behavior will, of course, depend on the amount and placement of the aluminum over and around the contact window.

In order to prevent such pitting, the aluminum may be deposited saturated with silicon so that it is unable to absorb any more (12, 13, 20), or it may be deposited over a thin sacrificial layer of polycrystalline silicon (21). When this problem was first recognized, there was considerable reticence to the use of silicon-doped aluminum because of the difficulty of deposition. A number of other alternatives, such as the deposition of a thin layer of aluminum followed by sintering and then a subsequent thick aluminum deposition, were also tried (18). With the later development of improved sputtering equipment came the routine deposition of aluminum–silicon of controlled composition.[5] However, oxide radiation damage may be induced that requires additional thermal annealing for removal.

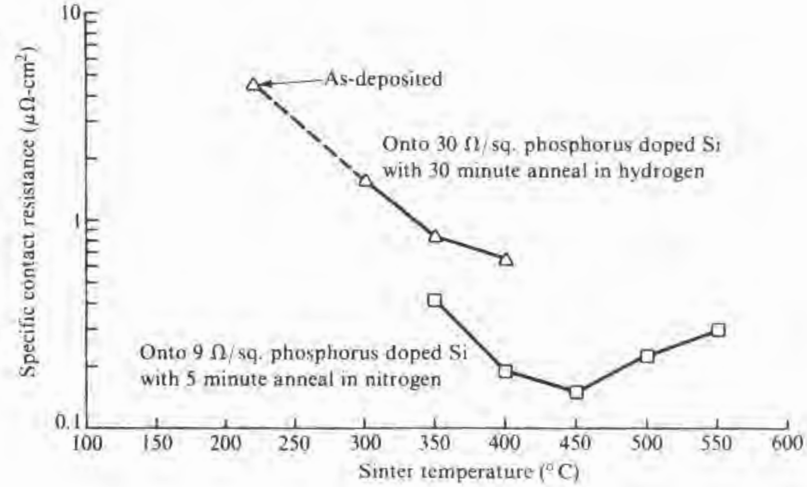


FIGURE 10.15

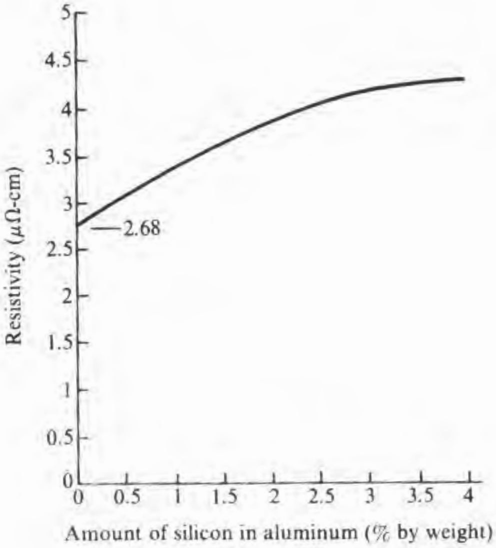
Specific contact resistance versus sinter temperature of aluminum on heavily doped phosphorus-diffused silicon. (The two sets of data were obtained from two different test patterns that may not give comparable results.) (*Source:* Upper curve data from R.A. Levy et al., *J. Electrochem. Soc. 132*, p. 159, 1985; lower curve calculated from data in H.M. Naguib and L.H. Hobbs, *J. Electrochem. Soc. 124*, p. 573, 1977.)

[4] Some idea of the penetration depth can be obtained from cross sections, but it is difficult to find a section that stops just at the bottom of a pit. Referenced data were obtained by replicating the contact surface after etching away the aluminum and examining the replica with a SEM.

[5] The addition of silicon requires some modification of the procedures used to bond connecting wires to the chip bonding pads. Because of the silicon precipitates, somewhat different etching techniques are needed to prevent silicon from being left on the wafer surface after lead pattern definition and to prevent jagged edge definition.

FIGURE 10.16

Aluminum resistivity at room temperature versus amount of silicon present. (*Source:* From data in L.W. Kempf, "Properties of Aluminum–Silicon Alloys," in Taylor Lyman, ed., *Metals Handbook*, American Society for Metals, 1948.)



To assess the minimum temperature that can be used for sintering, and thus to reduce pitting, several studies of specific contact resistance versus sinter temperature have been made (19, 22, 23). Fig. 10.15 shows the general behavior for n-type silicon and illustrates the need for sinter temperatures in the 400°C–450°C range. The reason for the upturn near 450°C for n-type silicon is not due to an increase in the bulk resistance of aluminum because of increased silicon uptake since, as shown in Fig. 10.16, the effect of silicon is considerably less than that of the sintering. The increase, not observed in p-type, apparently occurs because enough aluminum-doped silicon (p-type) regrows in the contact window to affect the contact resistance.

10.2.5 The Silicide–Silicon Ohmic Contact

One of the requirements of contact materials is that they withstand heating to a few hundred degrees while in contact with the silicon. Often, such heating is required for breaking through the residual surface oxide so that good electrical contact can be made. In addition, annealing near 500°C is sometimes required to adjust the SiO_2–Si interface states, and in multilevel metal systems, the interlevel insulator may require a 300°C–500°C deposition temperature. The result is that many potential metal–silicon contacts cannot exist because the silicon and the metal react to form a silicide. However, the silicides generally have very high electrical conductivity and in themselves make very dependable silicon contacts.[6] Silicides for contacts are usually made by depositing a thin layer of the metal over the entire wafer, heating the wafer to a high enough temperature for the silicon and metal to react in the contact window areas, and then etching away the unreacted metal on top of the oxide. Subsequently, additional metallization is added that makes contact to the silicide. This multiple-step operation prevents voids due to the volume change that occurs when a silicide is formed.

Most of the metals form silicides, as shown in Table 10.4, and in many cases, form more than one. Examples are PtSi, Pt_2Si, Pt_3Si, Pt_5Si_2, Pt_7Si_3, and $Pt_{12}Si_5$. For some of the compositions, high- and low-temperature crystallographic modifications may also occur. Although the list of silicides is formidable, when they are grouped according to their position in the periodic chart, some commonality of

TABLE 10.4

Metals Reported to Form Silicides

IA	*IB**	*IIA*	*IIB**	*IIIA*	*IIIB**	*IVB†*	*VB†*	*VIB†*	*VIIB*	*VIII‡*		
	Li			None reported								
	Na	Mg										
Cu	K		Ca		Sc	Ti	V	Cr	Mn	Fe	Co	Ni
	Rb		Sr		Y	Zr	Nb	Mo		Ru	Rh	Pd
	Cs		Ba		Rare earths§	Hf	Ta	W	Re	Os	Ir	Pt
						Th‖		U‖		Np‖	Pu‖	

*Group IB contains the alkali metals; group IIB contains the alkali earth metals; and group IIIB contains rare earth metals, also referred to as the lanthanides.

†The first three rows of groups IVB, VB, and VIB are sometimes referred to as transition and sometimes as refractory metals.

‡The third column of group VIII is sometimes referred to as near-noble elements. (The noble elements are gold, silver, mercury, and copper, so named because they are found free in nature.)

§La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu.

‖The elements thorium, uranium, neptunium, plutonium, and the others from atomic No. 89–103 are the actinides.

Sources: From compilations in Harold P. Klug and Robert C. Brasted, "The Elements and Compounds of Group IVA," in M. Cannon Sneed and Robert C. Brasted, eds., *Comprehensive Inorganic Chemistry*, Vol. 7, D. Van Nostrand Co., New York, 1958; in Bertil Aronsson, Torsten Lundstrom, and Stig Rundqvist, *Borides, Silicides, and Phosphides*, Meuthen & Co. Ltd., London, 1965; and in Marc-A. Nicolet and S.S. Lau, "Formation and Characterization of Transition-Metal Silicides," in Norman G. Einspruch and Graydon B. Larrabee, eds., *VLSI Electronics Microstructure Science 6*, Academic Press, New York, 1983. (For an extensive listing of the properties of silicides, see tables in the third reference just given.)

[6] Some silicides also make good Schottky diodes when contacting high-resistivity silicon (see section 10.3 on Schottky diodes). Thus, if windows are opened for both the Schottky diodes and the ohmic contacts, both structures can be made at the same time. Because of their low resistivity and high-temperature stability, silicides are also sometimes used as lead conductors in parallel with the higher-resistivity polycrystalline silicon.

the silicides and their siliciding properties can be obtained (24). These properties can be used to initially screen the silicides for potential semiconductor use. The reasons for elimination vary from reactivity with water to radioactivity.

Groups I and II silicides are generally considered inappropriate for semiconductor use. Cu, group IA, is a very fast diffusing lifetime killer in silicon. The group IB silicides that have been studied all spontaneously ignite in air. Most of the group II silicides slowly decompose from moisture in the air (25). The group IIIB rare earth metals form disilicides at very low temperatures and have n-Si barrier heights in the 0.3–0.4 eV range. The metals themselves oxidize rather easily and hence, from a manufacturing standpoint, are rather intractable (24). The actinide series (thorium, uranium, neptunium, and plutonium) are radioactive, and their use would build in a source of radiation damage. Thus, about half the known silicides do not appear to be candidates for semiconductor metallization at this time.

The transition (refractory) metals (groups IVB, VB, and VIB) require temperatures near 600°C for silicide formation. Silicides of the form $MoSi_2$ are generally the final form and generally have the lowest resistivity. In most cases, the species diffusing during silicide formation is silicon. When the silicide is formed from a co-deposition of metal and silicon, if there is a deficiency of metal, silicon will be rejected. If there is an excess of metal, it will be corrected by silicon diffusing in from the contact window. If no free silicon is available and the silicide is contacting silicon dioxide, the oxide is sometimes reduced.

TABLE 10.5

Silicide Barrier Heights on n-Type Silicon

Material	ϕ_B *(V)*
$CoSi_2$	0.65
$CrSi_2$	0.57
$HfSi_2$	0.55
$MoSi_2$	0.55
$NbSi_2$	0.62
$NiSi_2$	0.7
Pd_2Si	0.75
PtSi	0.84
$TaSi_2$	0.6
$TiSi_2$	0.6
VSi_2	0.55
WSi_2	0.65
$ZrSi_2$	0.55

Source: From data in Marc-A. Nicolet and S.S. Lau, "Formation and Characterization of Transition-Metal Silicides," in Norman G. Einspruch and Graydon B. Larrabee, eds., *VLSI Electronics Microstructure Science 6*, Academic Press, New York, 1983.

The near-noble elements Pt and Pd form metal-rich silicides or monosilicide at temperatures near 400°C. When produced from a Pt or Pd layer on Si, the siliciding is by metal rather than silicon diffusion. PtSi is widely used, both for contacts and for Schottky diodes. It was the first silicide commercially used for silicon contacts and was chosen because of its metallurgical stability, corrosion resistance, ability to be formed in the solid phase, and a visual appearance distinctly different from either Pt or Si (26). The latter property allowed optical examination of the contact areas after siliciding. Palladium has been used occasionally as an alternative since it can be deposited by evaporation and thus not produce radiation damage.

The rest of the group VIII elements, as well as those of group VIIB, are apparently potentially usable. Except for those with very low Schottky barrier heights, the silicides, like aluminum, make good ohmic contact only when contacting highly doped silicon. Contact resistance values have been measured for a few of the materials, but their behavior can be judged from the barrier height (see Table 10.5) and Fig. 10.8. It should be remembered, however, that just because a silicide is relatively stable and has a low resistivity does not necessarily mean that it will make a satisfactory contact material. Properties such as thermal expansivity, adhesion, and internal stresses must also be considered. When used only as a contact, the silicide can be very thin, and stresses may cause no problem. When contacts are combined with leads, much thicker layers are required. Then, if the silicide is made in situ, where lateral growth is confined, as in contact windows, stresses may be high enough to cause the silicide to break away from the silicon (27).

10.2.6 Alloyed GaAs Ohmic Contacts

As was shown in Fig. 10.6, the doping level of n-type GaAs required for good ohmic contact is in excess of 10^{19} atoms/cc. Since the making of n^+-regions by conventional diffusion is difficult and generally avoided if possible, the n^+-silicon contact has no commercial GaAs equivalent. Instead, alloyed contact metallization that includes a dopant source is generally used. Since such contacts present problems of metal balling and nonuniform dissolution of the GaAs, however, nonalloyed contacts continue to be examined.

One of the earliest contact metals was tin, but it has significant motion at device operating temperature when it is under the influence of high electric fields. As a result, it can be transported through a device and produce conducting channels (28). Silver, which does not dope, mixed with germanium for n-type and zinc for p-type contacts has also been used (29). Currently, gold with germanium for doping is used almost universally (7, 30). A eutectic mixture of germanium and gold (88% Au, 12% Ge) is now the most commonly used contact (31). The eutectic provides a liquid alloy source at the low temperature of 356°C. Germanium can be either a p- or an n-dopant but apparently in this case resides on a gallium site and dopes the GaAs n-type. Even with this doping, however, the specific contact resistance declines with increasing doping as was shown in Fig. 10.7 so that n^+-doping is still desirable. The surface melting and regrowth can leave a very rough surface, but the addition of an overlaying nickel layer will minimize that roughness. The ternary combination (Au–Ge–Ni) interacts with the GaAs and provides for much deeper penetration of the germanium than expected. Substantial lattice disruption also occurs, and there are some reports of a thin, high-resistivity layer underlying the contact (32).

10.2.7 Nonalloyed GaAs Ohmic Contacts

Three different nonalloyed contact systems have been reported. One uses ion implantation to introduce the heavy doping in the contact windows and then rapid thermal annealing to activate the dopant (33–35). The second forms a germanide on the GaAs surface and then depends on germanium diffusing out of it and into the GaAs for the high-concentration doping (36, 37). The third uses molecular beam epitaxy either to deposit highly doped GaAs in the windows

(38–40) or to deposit a layer of material with a lower bandgap (heterojunction) (17, 41).

10.2.8 Measurement of Contact Resistance

One of the earliest methods used for measuring semiconductor contact resistance was to use an epitaxial structure as shown in Fig. 10.17a and measure R_T for several contact sizes (29). For any given contact,

$$R_T = r_c + r_s + r_N \qquad 10.12$$

where r_N is the resistance of the n^+-layer and lower contact and is assumed to be approximately the same for all contact sizes. The spreading resistance r_s of the contact due to the epitaxial layer can be calculated from Eq. 10.3. If $R_T - r_s$ is then plotted versus 1/area A, the slope of the curve is R_c (or Ar_c), and the y axis intercept is r_N.

The Kelvin bridge is applicable to metal–metal contact resistance (as in a via) and to metal–diffused-layer contact resistance. Three versions are shown in Fig. 10.17b (42, 43). The first is appropriate for metal-to-metal measurements, and the two layers may be separated by an oxide except for the desired contact area. The other two depend on a diffused pattern and pn junction isolation to restrict current flow in the bottom half of the test pattern. They further assume that the current density is uniform over the whole contact area, in which case r_c is given directly by the expression $\Delta V/I$, where $\Delta V = |V_1 - V_2|$.

For contacts to a diffused stripe, such as a resistor, as discussed in section 10.2.1, a transmission line model is more appropriate, and Eq. 10.5 must be solved (8). If $d\sqrt{R_s/R_c} > 2$, then the coth term becomes 1, and

$$R_c = \frac{w^2 R_T^2}{R_s} \qquad 10.13$$

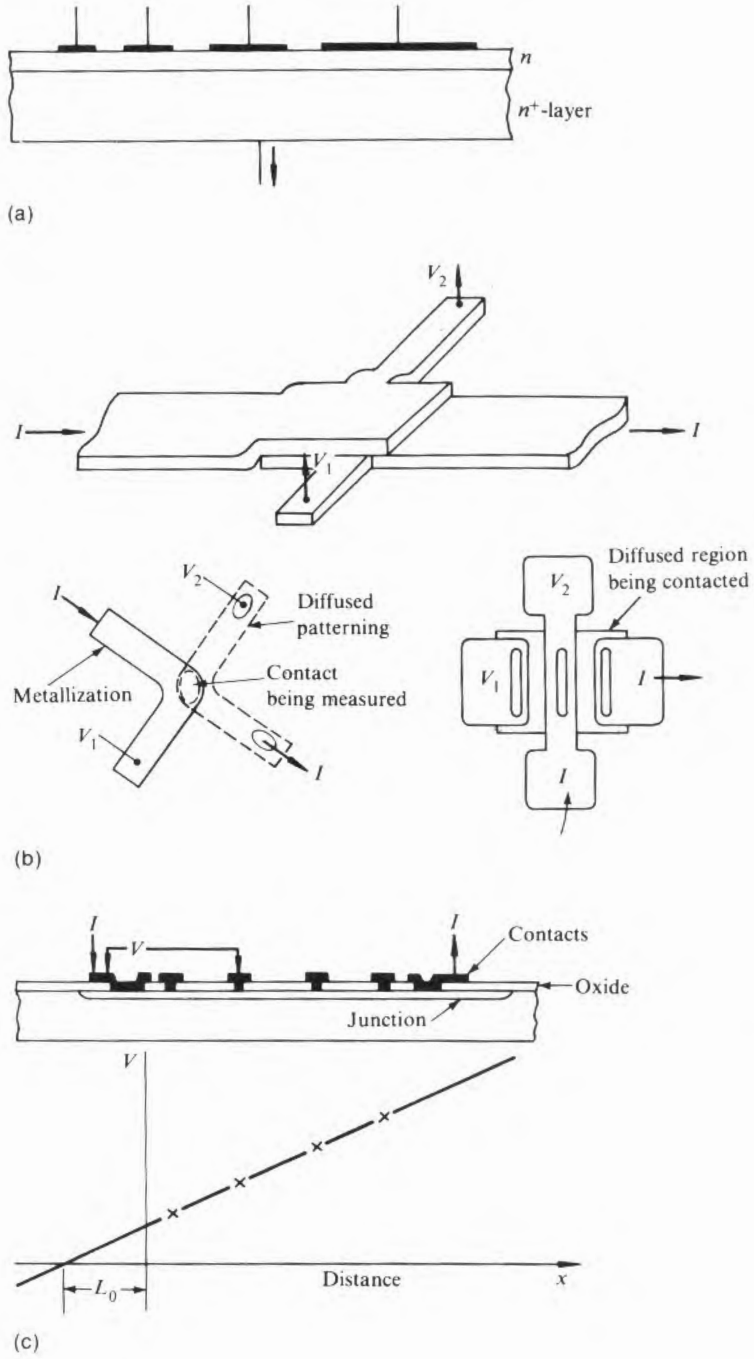
where R_c, the specific contact resistance, has been substituted for Ar_c. $d\sqrt{R_s/R_c} > 2$ is not a very stringent requirement since typically R_s is greater than 10 and R_c is less than 10^{-5}. Thus, a $d > 20$ μm will suffice. An alternate analysis, which eliminates the requirement for knowing the contact area, is shown in Fig. 10.17c (44). A series of contacts are made, and the voltage along the resistor is measured and plotted as shown. The curve is then extrapolated back to L_0, and R_c is given by

$$R_c = R_s L_0^2 \qquad 10.14$$

Other schemes as well have been used on occasion. For example, the behavior of a series of concentric contacts to a thin conducting layer has been interpreted in terms of the transmission line

FIGURE 10.17

Contact resistance measuring methods.



model (45). The voltage distribution under one of the current contacts of a transmission line structure, combined with the sheet resistance, can also be used to calculate R_c (46). In this case, an additional very thin diffusion is required to bring out test points that can be contacted. (For a discussion of errors associated with various methods, see reference 8.)

10.3 SCHOTTKY DIODES

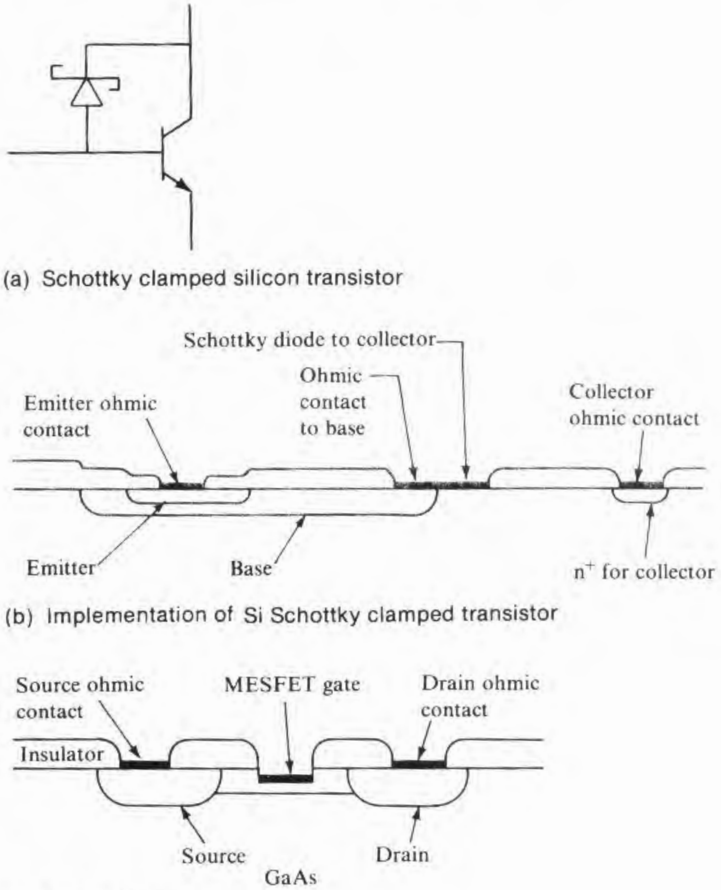
10.3.1 Rectifying Schottky Barriers

In previous sections, Schottky barriers were discussed in terms of a metal or metal-like material contacting a highly doped semiconductor so that the *I–V* behavior was linear. When the semiconductor is n-type and the doping level is less than perhaps 10^{17} atoms/cc, the *I–V* characteristics have the same general characteristics as those of a pn junction. There are, however, several basic differences. The two that make it a desirable circuit element in bipolar circuits are the fact that substantially all current is majority carrier current and the fact that its forward voltage drop is less than that of a pn junction. Thus, when polarity is changed, there is no long delay while minority carriers are recombining. A Schottky diode can be put in parallel with the collector–base of a bipolar switching transistor and act as a short to keep the collector–base junction from being driven into saturation. This allows the transistor to switch much more rapidly and is the basis for the high-speed Schottky TTL ICs. Fig. 10.18a shows how the diode is placed in the circuit. Fig. 10.18b shows how it is integrated on the chip and demonstrates that some integration can be done without resorting to additional leads. The Schottky barrier makes ohmic contact to the base but, where it overlaps the collector, forms a rectifying junction. In MOS circuitry, Schottky barriers are used as gates in GaAs MESFET transistors as shown in Fig. 10.18c.

In the fabrication of Schottky diodes, the same parameters as those applicable to pn junction diodes are important—namely, the reverse breakdown voltage, the reverse leakage well below breakdown, and the forward current. (For a discussion of both pn junction and Schottky diode diagnostics, see Chapter 11.) The breakdown voltage, as in pn junctions, is determined by avalanching in the semiconductor. The field at the edge of the Schottky metallization will be higher, however, and breakdown will appear to be premature. In addition, space charge effects at the barrier periphery can cause substantial excess current, both in the forward and the reverse directions (47). In particular, an accumulated surface will cause excess forward current, and because of the properties of silicon thermal oxide (see Chapter 3), oxidized n-type silicon surfaces show accumulation. To solve these problems, either pn junction guard rings (48) or MOS field plates (49) can be used. These structures are

FIGURE 10.18

Schottky diode applications to integrated circuits.



(a) Schottky clamped silicon transistor

(b) Implementation of Si Schottky clamped transistor

(c) GaAs MESFET transistor

shown in Fig. 10.19. The pn junction guard ring has the disadvantage of requiring additional fabrication steps and increasing the diode area, while the field plates are often not as effective as guard rings, particularly if the oxide is quite thick.

The theoretical Schottky forward current density J is given by

$$J = A^{**}T^2e^{-q\Phi_B/kT}e^{qV/nkT} \qquad 10.15$$

where n is slightly greater than 1 because of barrier lowering (50). Experimentally, depending on the cleanliness of the surface and how well edge effects are minimized by guard rings or field plates, n may range from about 1.03 to as much as 1.5. Fig. 10.20 shows experimental I–V plots for PtSi and Cr metallization on silicon. The

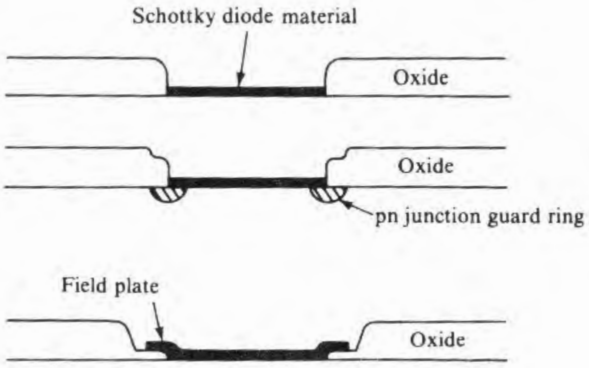


FIGURE 10.19

Schottky diode configurations showing both protected and unprotected perimeters.

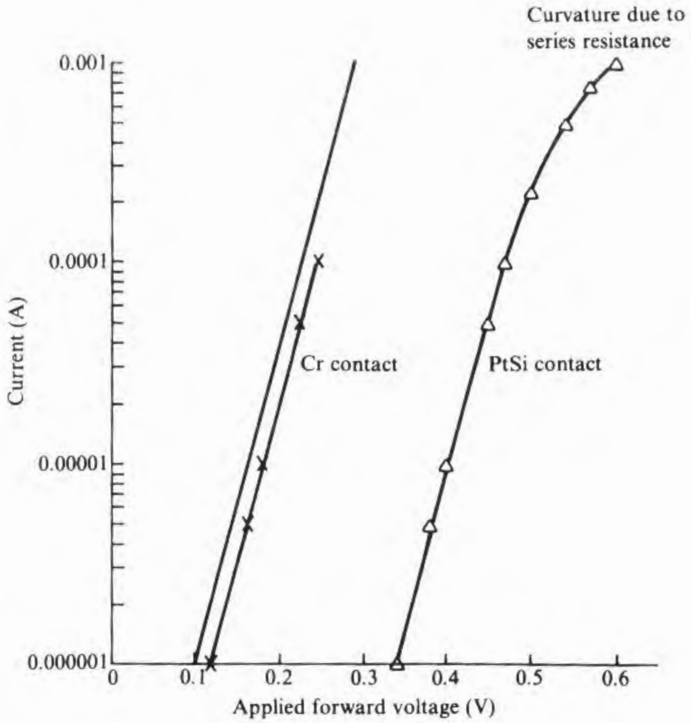


FIGURE 10.20

Silicon Schottky diode forward I–V plots for PtSi and Cr contacts. (The third line is for a contact with a barrier height 0.03 V lower than that of Cr.)

areas are the same, and, in each case, the n value is about 1.04.[7] The figure also shows the effect of changing barrier height on the current for a given voltage. Based on Eq. 10.15, the barrier height difference

[7] The n value can be determined by multiplying the voltage change required to change the current by 2 orders of magnitude by 8.38.

between the two contact materials is 0.021 V, although Tables 10.2 and 10.5 indicate that the difference between Cr and PtSi is 0.023 V. To show the sensitivity of forward current to ϕ_B, an additional curve based on Eq. 10.15 and separated from the Cr curve by 0.03 V is plotted.

10.3.2 Choice of Metal for Schottky Barriers

Despite the many elements and silicides discussed earlier that form Schottky barriers, only a few are routinely used. Aluminum, because of its widespread use as a contact and lead material for silicon ICs, has been extensively studied. Unfortunately, it exhibits substantial changes in barrier height (from a low of ~0.6 V to a high of ~0.8 V) depending on heat treatments after deposition. These changes depend on the thickness of interfacial oxide that might be present and on the effect of recrystallized aluminum-doped silicon under the barrier (51). Thus, aluminum is not used. When stability and compatibility with the rest of the metallization system are considered, only PtSi and Pd_2Si are in common use. In the case of GaAs, a somewhat different problem exists in that gallium will diffuse out of the GaAs and into some of the possible materials. Al, Mo, W, and Ti are the materials most often used. Occasionally, a circuit application will require Schottky diodes with two different "turn-on" voltages.[8] This requirement can most easily be accomplished by using two different materials with differing barrier heights (as shown, for example, in Fig. 10.20). Pd_2Si and PtSi have been used, but it is also possible to change the doping directly under the barrier by ion implantation and affect the barrier height by several tenths of a volt (52).

10.4 LEADS (INTERCONNECTS)

There are several possible choices for interconnect leads, as shown in Fig. 10.21, where the complexity and cost increase from top to bottom. The least expensive and easiest-to-use metal is aluminum. When the geometries are reduced to the point where more than one level of metal is required, doped aluminum is used to minimize hillock growth. Hillock minimization is necessary to keep the aluminum hillocks from puncturing the interlevel insulation. Further, as leads become narrower, current density increases so that doping is required to minimize electromigration.[9] The same forces driving

[8] Even though the forward I–V curves of pn and Schottky diodes appear linear on a semilog plot, they turn up rather abruptly somewhere below 1 V when plotted linearly (the familiar forward junction curve). For a given current density, the "turn-on" voltage is less for Schottky than for pn junction diodes, and for Schottkys, is less, the smaller the barrier height.

[9] Electromigration will be discussed in section 10.8.1.

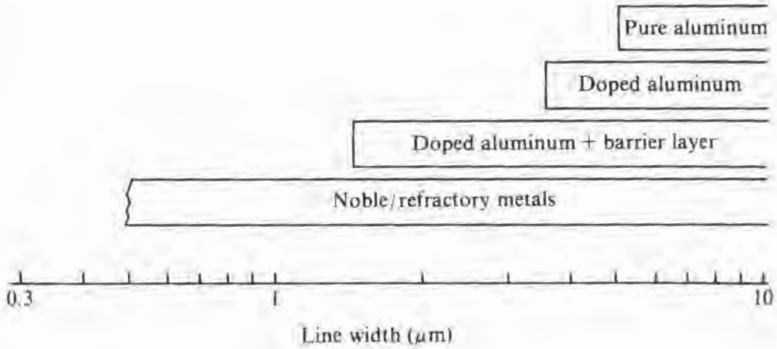


FIGURE 10.21

Applicability of various metal combinations to silicon IC device interconnections.

closer spacing and narrower leads also drive toward shallower junctions, which makes aluminum contact spiking more of a problem. Hence, the smaller geometries require a barrier layer between aluminum and silicon. Finally, some requirements are so demanding that aluminum is no longer applicable, and various noble/refractory metal combinations must be used. The various properties of potential lead materials is examined next.

10.4.1 Lead Resistance

When IC chips are very small and geometries are large, lead resistance is not a serious problem. However, as geometries shrink and operating frequency and chip size increase, the series resistance, coupled with lead-distributed capacitance, can cause undesirable pulse delays. New processing flows such as those using self-aligned gates and multilevel metal have required lead materials that can stand higher subsequent processing temperatures than aluminum and that often have electrical conductivities much lower than that of aluminum. Table 10.6 lists pertinent properties of several materials used in silicon contact/interconnect systems. Fig. 10.22 shows the range of lead lengths encountered in a particular VLSI chip (53). While the majority are less than 1 mm in length, a substantial number are considerably longer. One empirical rule is that the longest lead length is approximately half the square root of the chip area (54, 55). The impact of the available range of resistivities can be seen in Table 10.7, where line resistance and voltage drop for a representative set of conditions (including a lead length of 1 mm) are tabulated. From this table, it is clear that if the lead is of significant length and a few milliamperes of current are to be carried, neither polysilicon nor thin diffused layers will be satisfactory[10] because of their high resistance.

[10] In many parts of a circuit, a few millivolts drop in a lead is intolerable.

TABLE 10.6

Properties of Contact/ Interconnect Metals

	*Resistivity** *(10^{-6} Ω-cm)*			
Material	*Bulk Value*	*Thin-Film Value*	*Melting Point (°C)*	*Adherence to SiO_2†*
Aluminum	2.6	2.7–3	660	3
Copper	1.7		1083	2
Gold	2.4	3.4–5	1063	1
Molybdenum	5.8		2625	3
Palladium	11		1555	1
Platinum	10.5		1755	1
Polysilicon		1000	1420	3
Diffused silicon		1000	1420	NA
Silver	1.6		960	—
Tantalum	15.5		2850	—
Titanium	47.8		1800	3
Ti/W pseudo alloy				3
Tungsten	5.5	11–16	3370	2

*Thin-film values of resistivity are generally somewhat higher than those of bulk material because of additional crystallographic defects introduced during deposition. For a further discussion of this effect, see section 10.10.

†On a scale of 1 (poor), 2 (medium), and 3 (good).

FIGURE 10.22

Distribution of signal path lengths for a combination of the first and second levels of metallization of a VLSI circuit. (*Source:* Adapted from P.B. Ghate, *Proc. 20th IEEE Reliability Physics*, 1982.)

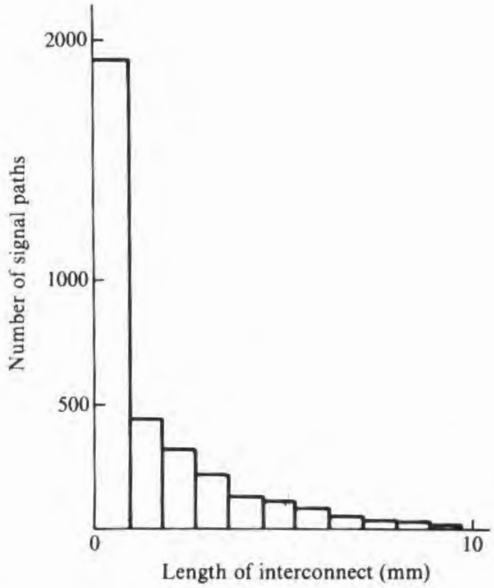


TABLE 10.7

Resistance of an IC Lead

*Material**	*Resistance* *(Ω)*	*Voltage drop (1 mA)* *(mV)*
Aluminum	10	10
Molybdenum	30	30
Titanium	150	150
$MoSi_2$†	300	300
Polysilicon	3000	3000
Diffused silicon	3000	3000

*3 μm wide, 1 μm thick, and 1 mm long.
†The resistivity of various silicides is given in Table 10.8.

The combination of lead resistance and its associated capacitance will cause a pulse delay time that increases as the lead material resistivity increases (55, 56). The delay time will remain constant as the lead width is increased since the resistance decreases and the capacitance increases in step. Thus, the delay time cannot be decreased by making the leads wider. Making them thicker will help since in that case the resistance is decreased without increasing the capacitance, but then lead definition becomes more difficult. When the lead attaches to a separate capacitor, as, for example, a MOS gate, the gate capacitance and the lead resistance may determine the delay time. Then, increasing the lead width to lower its resistance will help until the lead becomes wide enough for its capacitance to dominate the gate capacitance. At that point, the delay time will begin to increase. This behavior is shown in Fig. 10.23, which illustrates that even a 10 μΩ-cm resistivity may not be adequate to allow top performance of very fine geometry circuits (57).

For short runs, such as crossunders to provide paths under conventional leads over oxide, diffused layers have been used since the first ICs were made and work well for most purposes. However, as operating frequencies increase, air-isolated bridges such as are now used in GaAs, may be required for silicon. Because of its ability to withstand subsequent high-temperature operations, polysilicon is commonly used as a combination MOS gate electrode material and partial first-level interconnect (58). In this application, current flow is small, but the high resistance can cause pulse rise-time and propagation delay problems. The problem of high resistance causing delay was recognized soon after the use of polysilicon was introduced and led to the early proposal for a refractory metal lead system (56). This approach has not been well accepted, but processes that add a layer of metal silicide to the top of the polysilicon are now being used. The silicides, as shown in Table 10.8, have between 10 and 100 times the conductivity of polysilicon and thus offer considerable

FIGURE 10.23

Delay time versus gate/lead width for various electrical conductivities. (*Source:* N. Yamamoto et al., *J. Electrochem. Soc. 133*, p. 401, 1986. Reprinted by permission of the publisher, The Electrochemical Society, Inc.)

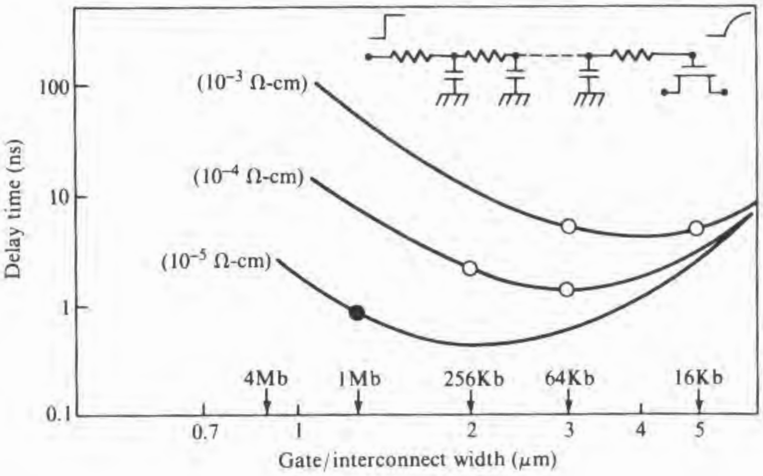


TABLE 10.8

Silicide Resistivity at Room Temperature

Material	*Co-Sputter (μΩ-cm)*	*Metal–Polysilicon Reaction (μΩ-cm)*	*Reference*
$CoSi_2$	25	17–20	(1)
$HfSi_2$		45–50	(1)
$MoSi_2$	100		(1)
$NbSi_2$	70		(2)
$NiSi_2$	50–60	50	(1)
$PdSi_2$		30–35	(1)
PtSi		28–35	(1)
$TaSi_2$	50–55	35–45	(1)
$TiSi_2$	25	13–16	(1)
WSi_2	40*–70		(1, 3)
$ZrSi_2$		35–40	(1)

*Slightly lower values are reported for CVD co-deposition. See, for example, S. Sachdev and R. Castellano, *Semiconductor International*, May 1985.

1. S.P. Murarka, *J. Vac. Sci. Technol. 17*, p. 775, 1980.
2. T.P. Chow et al., *J. Electrochem. Soc. 133*, p. 175, 1986.
3. J. Kato et al., *J. Electrochem. Soc. 133*, p. 794, 1986.

improvement. The layer can be formed by depositing a layer of metal on the surface of polysilicon and then alloying it and the top portion of the polysilicon together. Alternatively, a silicide can be deposited by sputtering from a silicide target, by co-sputtering from silicon and metal targets, or by CVD. Of the materials listed, titanium silicide affords a good compromise of low resistivity, high adherence, ease of forming, and the ability to make acceptable

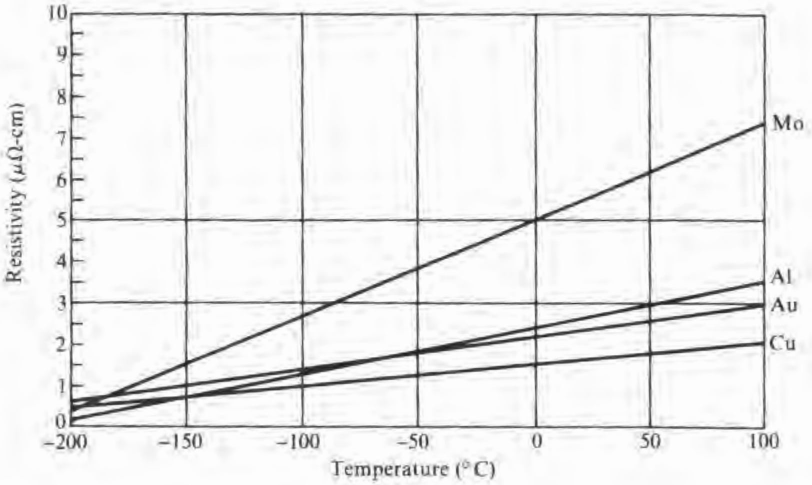


FIGURE 10.24

Effect of temperature on electrical resistivity of several metals. (*Source:* From data in Alexander Goldsmith et al., *Handbook of Thermophysical Properties of Solid Materials*, Vol. 1, Pergamon Press, New York, 1961.)

Schottky diodes so that one material can be used for both leads and Schottky contacts. Its conductivity is, however, 5 to 10 times less than that of the metals commonly used for leads. As can be seen from Fig. 10.23, and as has been projected, even the metals of Table 10.6 may become inadequate for very fine-line geometries. One method of further reducing the lead resistance is by cooling the leads (and the semiconductor device as well) (59). Fig. 10.24 shows how several potential lead materials behave at low temperatures. By going to liquid nitrogen temperature, the resistance could be reduced by a factor of ~5. Before such a step is taken, however, the device design must be optimized for this temperature. If higher-temperature superconducting materials become a reality, if they can remain superconducting at the high current densities at which ICs operate, and if they can be made compatible with IC processing, they would be a very appealing alternative (60). Another high-speed interconnect system, but one that is more specialized, is the use of light pipes and optical coupling directly on the chip (61, 62).

10.4.2 Adhesion to SiO_2

Most IC leads run over silicon dioxide, and thus it is very important that adhesion be good. Table 10.6 categorized the adhesive properties of the more common interconnects as good, medium, and poor. Aluminum and polysilicon adhere very well on clean oxide, and when a lead is torn away, it will usually take away the oxide and some silicon beneath. A case of peeling leads usually occurs because the oxide surface was not well cleaned but may occur because of a faulty deposition. In the case of Ti:W, adhesion is dependent on such deposition parameters as sputter target potential, system pressure, and system leaks (63). In the case of the medium adherers, it is possible to have peeling because of excessive shear developed at the interface from internal film stresses. In such cases, either a change in deposition conditions or a post-deposition anneal may solve the problem. It may be helpful to monitor the stress using the same procedure described in Chapter 3 for determining stress in oxide films. When the leads are multicomponent, adhesion may depend on the exact composition. For example, silicon-rich tungsten disilicide appears to adhere better than tungsten-rich disilicide. Films such as gold with poor adhesive qualities should not be used directly. They can, however, be used on top of a thin layer of some other metal with less desirable electrical properties. An example is the use of a very thin buffer layer of titanium, followed by a thick gold layer. In this particular example, yet another layer is required to prevent interaction between the gold and the titanium, and it is usually platinum.

10.4.3 Diffusion Barrier

Diffusion barriers may be used to prevent undesirable impurities from reaching some part of the circuit. An example is the shielding of a MOS gate oxide from impurities diffusing through the gate after a source–drain ion implantation. More often, however, diffusion barriers are thin layers inserted between two other layers of multilayer leads or between the main body of the lead and the semiconductor to prevent interdiffusion and chemical reaction (63, 64). A common application is between aluminum leads and silicon to prevent contact pitting. Another is the use of a layer between PtSi and aluminum to prevent the aluminum–PtSi reaction from changing the Schottky barrier height. Under some circumstances, layers may serve the dual role of improving adhesion over oxide and preventing interactions in the contact area.

Barriers have been categorized based on the manner in which they work as passive or sacrificial (65). Passive layers are quite inert to the layer on each side and thus will, in principle, always keep the two layers separated. Unfortunately, there may be substantial diffusion of one or both components through the passive barrier. It should be noted that high-temperature diffusion coefficient data extrapolated to low temperatures will usually substantially understate the low-temperature value. Sometimes, it is possible to add a component to the barrier to reduce diffusion. This combination is referred to as a stuffed barrier. An example of a stuffed barrier is thin-film TiN that has been sputtered in the presence of oxygen. Without the oxygen, substantial diffusion of silicon through the layer occurs when the layer is used as a barrier between silicon and aluminum. The diffusion apparently takes place along grain boundaries, where the inclusion of oxygen will inhibit diffusion (66). Sacrificial barriers will react, but at a slow enough rate that they are still useful.

Almost all barriers in common use are either sacrificial or else allow noticeable diffusion. However, some of the transition metal nitrides, borides, and carbides may be truly passive. As a class, they have high electrical conductivity and are stable in the temperature range of interest (67, 68). For example, TiN can be sputter deposited with a resistivity of about 20 μΩ-cm (69) and will make good ohmic contact to low-resistivity silicon. When the behavior of the barrier is considered, the effect of additives to the films being separated should not be neglected. For example, it is common to add a few percent copper to aluminum to reduce electromigration, and the copper's presence retards the reaction between aluminum and titanium (70). Table 10.9 shows a few of the many combinations of leads and barriers now in use. It is to be emphasized that as new metallization systems are introduced, they must be carefully examined for latent failures caused by interdiffusion and unexpected compound formation.

Not listed in the table is a barrier for use between aluminum and gold, the lack of which produced the most famous case of premature failure in the history of the semiconductor industry. Gold wires are commonly used to connect from gold-plated package leads to aluminum bonding pads. At the upper end of the temperature range over which silicon transistors will operate, aluminum and gold slowly form a series of colored intermetallics (the "purple plague") that will ultimately cause failure. By somewhat restricting the temperature range (also necessary for plastic packages), the problem can be contained. For cases where higher temperatures are required (as for components for some military applications), an all-aluminum system using aluminum leads, aluminum wires, and aluminum-coated package connections is sometimes used. As an alternative, several gold metallization systems have been developed. Among these are the beam lead method (26) and a gold lead metallization that can be used directly with gold wire bonding.

TABLE 10.9

Examples of Diffusion Barrier Applications

Barrier Material	*Films Separated*	*Reason*
PtSi	Si and Al	Prevents aluminum contact spiking.
Ti:W	PtSi and Al	Prevents Al–PtSi reaction.
Pt	Ti and Au	Prevents Ti–Au reaction.
W	Au–Ge–GaAs and Au	Prevents Ga diffusing into Au.
Si (poly)	Gate oxide and poly n^+	Prevents dopant from reaching gate.

10.4.4 Interconnect Material Contamination

Care must be taken to ensure that the interconnect materials used are free of sodium contamination. High-purity materials processed especially for the semiconductor industry should pose no problem, but alternative materials being investigated may not have been refined to semiconductor-grade standards. Radioactive impurity levels, although generally quite low, are a potential problem in the case of high-density MOS memories, where soft errors can occur because of carrier generation from random radioactive particles.

10.4.5 Step Coverage

An important part of lead technology, and one that has little to do with lead composition, is step coverage. Leads are generally about 1 μm thick, and steps range from about 0.25 μm to greater than 1 μm high. The step profiles range from gentle, relatively easy-to-cover slopes (Fig. 10.25a) to virtually impossible-to-cover overhangs (Fig. 10.25d). (For a discussion of the causes of these various profiles, none of which are metal related, see Chapter 6.) In addition to etching procedures, some other process-related factors may affect the step profile. If a contact is being etched in a relatively thin oxide window nested inside a thicker oxide as shown in Fig. 10.26a, misalignment of the mask before etching can cause an unexpectedly high and abrupt step as shown in Fig. 10.26b. If steps are in CVD oxides, the oxide can be doped with phosphorus, arsenic, or boron to reduce the melting point and allow the oxide to be reflowed after etching, thus rounding the corners and allowing better coverage. For the specific case of leads covering holes in oxides leading down to a contact area, selective metal deposition can be used to deposit conductive plugs and fill the holes as shown in Fig. 10.27. In the following step, metal for the leads can be deposited without worrying about the step down into the contact opening.

In order to produce a uniform (conformal) coating over steps, the incident atoms should arrive over a wide range of directions. However, most metallization sources allow arrivals from a very limited solid angle so that step coverage is determined by geometric shadowing. Fig. 10.28a shows a cross section of a lead that traverses a relatively modest slope and yet that has substantial thinning over the step. As the step approaches 90°, the profile becomes like that of Fig. 10.28b. The deep groove at the bottom causes a serious reliability problem and illustrates the effect of a profile changing from that of Fig. 10.25a to that of Fig. 10.25b. The low step of Fig. 10.25a,

FIGURE 10.25

Oxide step profiles encountered in IC manufacturing.

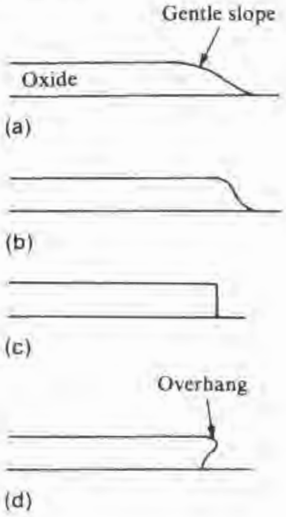


FIGURE 10.26

Effect of contact window misregistration on oxide step height.

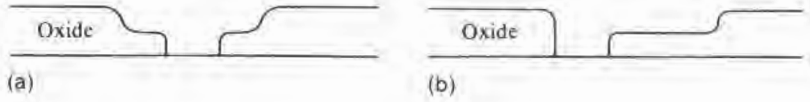


FIGURE 10.27

Use of selective metal deposition to eliminate a step.

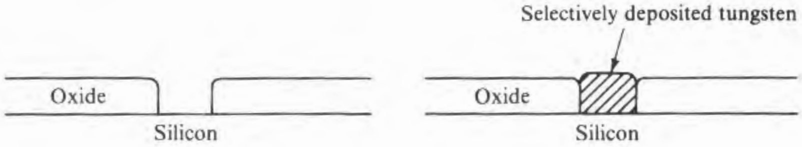
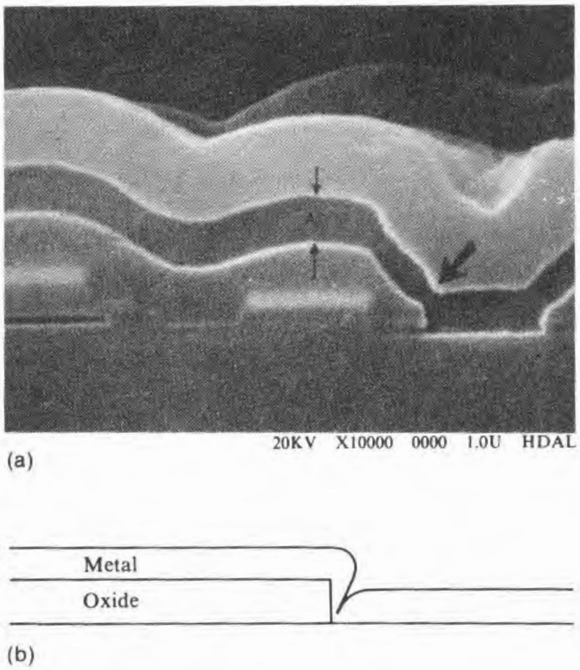


FIGURE 10.28

(a) Metal and CVD oxide step coverage. (*Source:* Photograph courtesy of Dr. P.B. Ghate, Texas Instruments Incorporated.) (b) Metal coverage to be expected if the step were vertical.



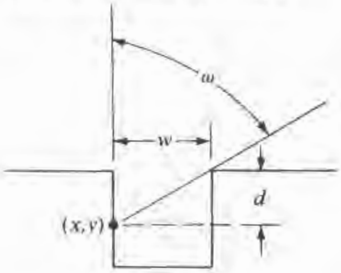
combined with a thick metallization, can be acceptably covered. However, the high step of Fig. 10.25b leads to a metal coverage profile like that of Fig. 10.28b.

Calculations of the coverage to be expected from an evaporation or sputter source have been made (71, 72). Assumptions usually made are as follows:

1. In evaporation, the mean free path distance is greater than the source-to-wafer distance.
2. The source-to-wafer distance is large compared to the step height.
3. The arriving atomic sticking coefficient is 1, and there is no surface mobility.

FIGURE 10.29

Maximum spread of arrival angles to a point (x,y) on a vertical wall of a hole or trench.



4. The layer grows at a rate proportional to cos ω/r, where ω is the angle between the incoming vapor stream and the surface being coated and r is the distance from the source to the surface.

If the source-to-wafer distance is considered constant and the solution is restricted to two dimensions, expressions of the form given by Eq. 10.16 result (71):

$$\Delta x = C(\cos \omega_1 - \cos \omega_2) \qquad 10.16a$$

$$\Delta y = C(\sin \omega_2 - \sin \omega_1) \qquad 10.16b$$

where Δx and Δy are the changes in the coordinates of a point (x,y) on the surface, C is a constant that includes the deposition rate and the time span over which Δx and Δy are calculated, and ω_1 and ω_2 are the angular limits of ω and will change as x and y change. To determine a final profile after a substantial thickness has been deposited, it is necessary to do many calculations, advancing points on the surface each time by a small amount of Δx and Δy. The limits of ω_1 and ω_2 can change with each new calculation since an additional thickness may block off some of the incoming beam. Further, the limits can be different in different regions of the surface. For example, for points on a broad horizontal surface, the angle will be limited only by the source. However, as shown in Fig. 10.29, no matter what the source angle, the incident angle along a side wall is limited. For the particular case of vertical walls of the well of the figure, $\omega = \tan^{-1}(w/d)$ where w is the width of the opening and d is the distance down to the point (x,y) (73).

Use of this kind of model gives predictions as shown in Fig. 10.30 and demonstrates how voids can be produced in holes such as could be encountered in small-geometry vias and contacts (73, 74). Fig. 10.31 shows the manner in which the thickness of a lead over an isolated step changes as the step angle varies (64). Computer programs are available for calculating step coverage, and SAMPLE is a very popular one (72).

Evaporation sources are very small. Thus, if the slices are stationary, there is only one angle of arrival, and shadowing would be completely unacceptable. To extend the angles of arrival, planetary slice holders are used that move the slices so that the source appears to be nearly hemispherical. Sputter sources are generally large flat plates, and the pressures are higher so that the mean free paths are shorter. However, under normal sputtering conditions, little difference exists between evaporation and sputter coverage (64). However, it has been reported that if the conditions can be changed so

FIGURE 10.30

Calculated step coverage for various ratios of film thickness to step height. (A total arrival angle on the flat surface of 1π was assumed.) (*Source:* A.C. Adams, *Solid State Technology*, p. 135, April 1983. Reprinted with permission of *Solid State Technology*.)

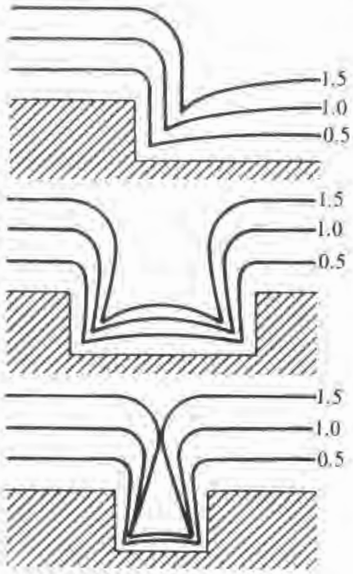
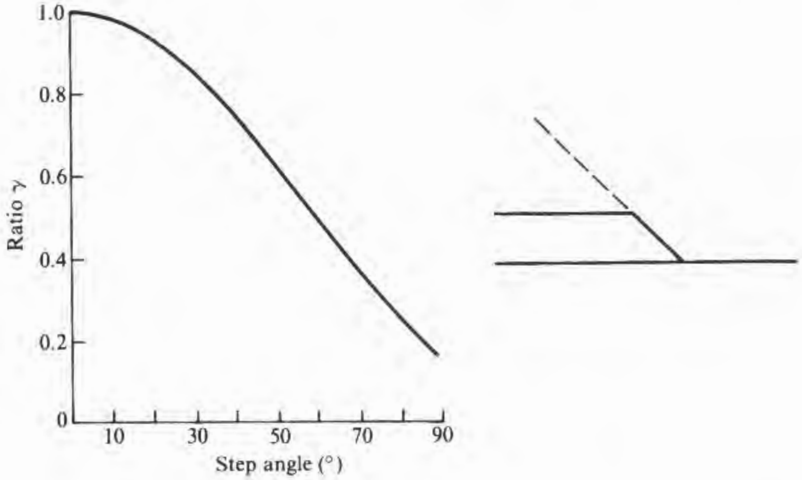


FIGURE 10.31

Calculated effect of step angle on ratio γ of metal thickness on step wall to that on an unshadowed horizontal surface. (Conventional planetary slice rotating tooling was assumed.) (*Source:* Adapted from P.B. Ghate, *Thin Solid Films 93*, p. 359, 1982.)



that resputtering of the depositing surface occurs, substantial redistribution and much better coverage occur (74). Plasma deposition atmospheres have short mean free paths and lead to an incoming angular spread very close to π. Chemical vapor deposition comes close to giving uniform coatings regardless of the profile. The mean free path of reacting atoms is very short, and atoms reach the surface, not by a line-of-sight path from the source, but by diffusion. Thus, even with overhangs, reacting atoms can reach the surface. In addition, there is usually a high degree of surface mobility of the reacting species, which further improves coverage. It is for this reason that the plug shown in Fig. 10.27 could be deposited. Not all materials and flows are amenable to CVD, however. Consequently, it is not a solution to all step coverage problems.

10.4.6 Protective Overcoats

Metal leads are generally thin, soft, and narrow and thus are subject to damage by scratching, which is most likely to occur either during multiprobe or during chip separation. To minimize scratch damage, a hard overcoat of SiO_2 had been added to the process flow by the late 1960s. The coating methods were originally E-beam evaporation or sputtering but later were expanded to include chemical vapor deposition of silicon dioxide and the plasma deposition of silicon nitride. All are low temperature in order to not damage aluminum leads, and, originally, most layers were cracked. The cracking was not particularly important for scratch protection nor for the subsequent use of protecting leads from being electrically shorted by loose particles in the package. However, with the realization that have been studied extensively for the aluminum–silicon system and are thought to be associated with changing interface properties, both at the aluminum–oxide and the silicon–oxide interface (78). In any event, changes in ϕ_{ms} will generally be small enough so that, when thresholds are adjusted by ion implantation, a change of gate material will only require a change of implant dosage and thus be easily accommodated.

10.6 MULTILEVEL METALLIZATION

As circuits became more complex, larger and larger portions of the surface area must be dedicated to leads, and, in many cases, the chip size determining factor is lead area. Without provisions for the crossing of leads, the problem of lead routing is insurmountable for all but the very simplest of circuits. Diffused crossunders have been used from the beginning, but such diffused paths add circuit capacitance and resistance, use appreciable area themselves, and still only allow one level.

An example of multilevel interconnects was shown in Fig. 10.1. The first level of metal is placed on the thermal oxide (Si ICs). Next, the wafer surface (thermal oxide and first-level leads) is covered by an insulator, usually CVD silicon dioxide, and via holes are cut through the CVD oxide where the second-level leads must contact the first level. Another layer of metal is then added and patterned. In some cases, yet another level is used, but the manufacturing problems increase with each level. In MOS circuits, if a polysilicon gate and polysilicon leads are to be used along with aluminum leads, the polysilicon must be deposited before the aluminum. Otherwise, the temperature required for the silicon deposition will ruin the aluminum–silicon contacts.

Multilevel capability has allowed more complex chips to be built in the same area and, in some cases, provided enhanced performance. The main process problem encountered in multilevel metallization is the difficulty in running leads up and down over the surface contours that would ordinarily get worse as the number of levels increases. Additionally, when aluminum is used, there can be difficulty in ensuring that all oxide is removed in the via contact area before the next level is added.

10.6.1 Interlevel Shorts

Interlevel shorts can occur either because of flaws in the interlevel oxide or because of hillock growth from the metal surface after it and the oxide have been deposited. Oxide flaws can occur because of particulate contamination during the oxide deposition or can stem from unwanted holes etched in the oxide due to a lithographic failure or from poor oxide step coverage (usually not a problem). Hillock growth is a serious problem, particularly when aluminum leads are used. Grain growth occurs during temperature cycling—for example, during deposition of the interlevel oxide—and because of film constraints, some of the growth is in the form of hillocks rising above the rest of the surface. Hillocks also occur as the result of electromigration (see section 10.8.1) and because of thermally induced or deposition-induced stresses in the film. These hillocks can penetrate the interlevel insulation and contact the next level of metal. Aluminum is a particularly poor material to use in this regard, but the addition of a few percent copper reduces the problem substantially.

10.6.2 Planarization

To prevent the worsening of surface contours as more layers are added,[13] various planarization processes are used. One partial solution, already mentioned, is the use of selectively deposited metal plugs in vias. Another is the use of CVD silicon dioxide heavily doped with boron and phosphorus (BPSG, or borophosphosilicate glass), which reduces its softening temperature and allows it to be reflowed. Reflowing allows both the sharp edges of openings and the hills formed by the oxide going over lower-level leads to be smoothed. Unfortunately, reflowing requires temperatures higher than those applicable to the commonly used aluminum, but such coatings are suitable for use over polysilicon.

Processes used for planarization after aluminum leads are generally based either on a simultaneous deposition/etchback process for the interlevel dielectric or on a chemically vapor deposited dielectric followed by a sacrificial conformal coating of either spin-on glass[14] or photoresist (79). Combined deposition/etching methods include biased sputtering (80) and plasma-enhanced CVD combined with ion etching (81). If a LPCVD process is available that gives a smooth surface as thickness increases, a very thick coating can be added and then etched back to the desired thickness by an anisotropic plasma etch (82). The process is simpler than a combination etch/deposition, and the results may be adequate. The sacrificial conformal coating process is based on the premise that the coating will be thinner over the high spots and thicker in depressions and that its upper surface will be much smoother than that of the underlying CVD oxide. If a plasma etch with the same etch rate for the spin-on coating and underlying CVD oxide is then used to remove all of the coating, the surface of the remaining CVD oxide will mirror the original surface contour of the coating (83).

[13] That steps will become more troublesome with each layer added can be seen from Fig. 10.23, which shows that as the thickness increases (which could come from additional layers) steps tend to become more pronounced.

[14] Usually a water-soluble silicate that, when heated, will decompose into volatiles and silica. The familiar water glass is one example of a water-soluble silicate.

10.7 INTERCONNECT AND GATE METAL PATTERN DEFINITION

Metal definition is done by a combination of lithography and wet etching, dry etching (plasma), or lift-off. Wet etching requires the least expensive equipment, is conceptually the simplest process, and has been used the longest. Plasma etching has little undercutting and thus is more applicable for fine geometries. Lift-off definition is a specialized process ordinarily used in fine-line applications. Both wet and dry etching procedures are covered in detail in Chapter 6. Lift-off is described here next.

When multiple layers are etched, the problem of preventing undercutting of at least one of the layers can be quite severe. Plasma etching is the most uniform, but the problem still remains, when multilayer systems are used, of providing the necessary series of reaction gases. Further, the requirement of minimal undercutting implies straight walls, and straight walls are more difficult to cover by the follow-on insulating layer. The lift-off process makes use of the fact that it is difficult to get good coverage over steps when most deposition systems are used (84–86). Thick resist is patterned so that it remains where the leads are unwanted. When the resist is covered with lead material, there will be a thinning or, in some cases, a break at the step. Thus, when the resist is removed, the metal on top of it will also be removed, leaving the leads in the areas where there was no resist. By using a multilayer resist, the lower layer can be undercut so that a pronounced shadowing of the edge of the leads occurs. This has the effect of producing a lead with sloped edges that will be easy to cover. Lift-off processing of metal is more widely used on gallium arsenide than on silicon.

10.8 FAILURE MODES

Metallization failure modes can be broken into five broad categories:

1. Electromigration-induced open leads that result from high lead currents causing a mass transport of lead material and ultimately a break in the lead.
2. Time-induced break in leads over steps caused by a combination of thermal cycling and stress in the film.
3. Electrical-overstress-induced open leads that result most often from an electrostatic discharge vaporizing a section of the lead.
4. Corrosion-induced open lead or a short between adjacent leads caused by moisture and/or poor cleaning.
5. Interlevel shorts caused either by breaks in the interlevel oxide or by growth of metal hillocks.

In addition, it is possible to have failures at the lead–wire-bond junction due either to an improper match of wire and lead materials causing undesirable intermetallic compound formation or to faulty bonding procedures. Neither of these problems will be discussed further.

10.8.1 Electromigration

Electromigration is the movement of lead material in the direction of electron flow because of momentum transfer from electrons to the metal ions. It is noticeable only when current density is very high, as it often is in IC leads. For example, a 1 μm thick strip 5 μm wide carrying 10 mA of current has a current density of 2×10^5 A/cm^2. This value is to be compared to a typical house wiring code that specifies a maximum of 30 A for a #14 wire, or less than 10^3 A/cm^2. Because of the possibility of such large current densities, electromigration properties must be considered both in circuit design and in the choice of lead material.

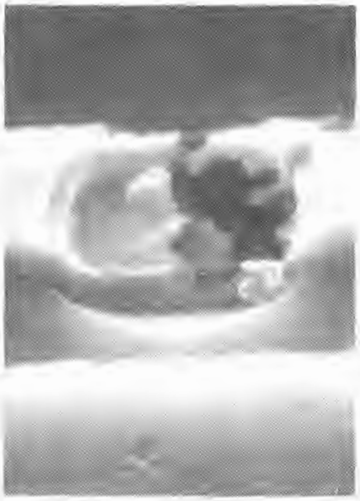
The metal ion transport flux J_m in the ideal case is given by (87)

$$J_m = \frac{NDZ^*e\rho j}{kT} \qquad 10.18$$

where N is the density of metal ions, D is the self-diffusion coefficient of the metal, Z^*e is the effective charge of the ion, ρ is the resistivity of the metal, and j is the density of the current flowing through the metal. For application to thin films with many grain boundaries, the expression is modified somewhat, but the form is the same.

FIGURE 10.33

Failure of an aluminum lead at a contact because of electromigration. (*Source:* Photograph courtesy of Dr. P.B. Ghate, Texas Instruments Incorporated.)



The effects of such migration are the growth of hillocks toward the positive end of the lead and the formation of voids and ultimately a break in the lead near the negative end. An example of a gap in a lead is shown in Fig. 10.33. That there can be enough transport to constitute a transistor reliability problem was recognized as early as 1965 (88) and has been studied extensively since (89, 90). Transport depends on the material and increases exponentially with temperature since D of Eq. 10.18 is of the form $D = D_o e^{-E/kT}$ where D_0 is a constant and E is the diffusion activation energy. For a given material, the more crystal defects, the more pronounced the migration. However, the addition of small amounts of another metal, such as 1% or 2% copper in aluminum, reduces the effect. In the case of aluminum, an overcoating of silica glass also reduces transport (91).

MTF, the median time to failure,[15] which in this case is the time to move enough material to cause a fatal flaw, is given by (90)

$$\text{MTF} = A'j^{\;m}e^{E/kT} \qquad 10.19$$

[15] The median time to failure is the time at which half the units have failed.

where A' is a constant and m is between 1 and 3. Experimentally, the MTF of aluminum leads evaporated onto a cold substrate has been found to be approximated by (89)

$$\mathrm{MTF} = 4.1 \times 10^{16}(ABj^{-2})e^{0.48/kT} \qquad 10.20$$

where MTF is in hours, T is the temperature in K, A is the cross section of the lead in cm^2, B is a units conversion constant and equals 1 (A^2·hour)/cm^6, and k is Boltzmann's constant expressed in eV/K. For aluminum evaporated onto a hot substrate so that large grains are produced, the activation energy typically increases to ~0.7 eV, and the MTF over the normal operating range is substantially increased. Other materials such as gold and tungsten show a pronounced increase in activation energy and in MTF.

EXAMPLE □ Compare MTF for leads 1 μm thick and 10 μm wide produced by evaporation onto a cold substrate if the IC chip is operating at 70°C and the lead is carrying first 30 mA and then 50 mA.

The cross-sectional area of the lead is $10^{-4} \times 10^{-3}$, or 10^{-7} cm^2. The current density j is either 3×10^5 A/cm^2 or 5×10^5 A/cm^2. The temperature is 343 K, and $k = 8.62 \times 10^{-5}$ eV/K. $ABj^{-2} = 1.1 \times 10^{-18}$ hours or 0.4×10^{-18} hours. Substituting these values and kT into Eq. 10.20 gives 46 and 17 years, respectively. Also, from Eq. 10.20, MTF varies as $(j_2/j_1)^2$, or as 25/9, which closely matches the ratio of 46/17. It should be remembered that this is median time to failure and that the actual time could be substantially less. Also, Eqs. 10.19 and 10.20 relate only to two particular sets of conditions. In actual manufacturing processes, a whole smear of activation energies ranging from 0.48 to almost 1.4 is observed. Also, the preexponential may vary over 6 or more orders of magnitude. □

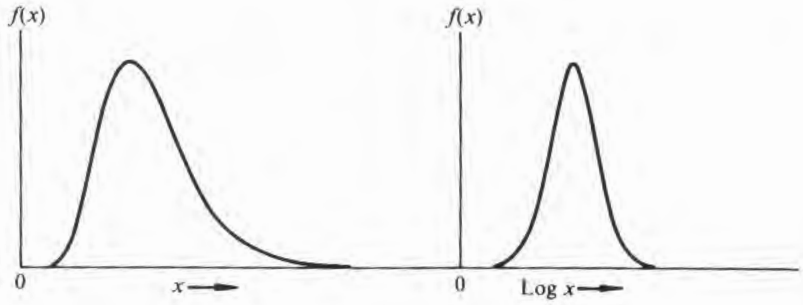
In most cases, the failure rate $\lambda(t)$ rather than the median time to failure is of more importance to a systems designer. The distribution of failures versus time has been experimentally observed to obey, not a normal,[16] but a log normal distribution.[17] Log normal distributions show a pronounced skew with a long tail extending to longer time, as shown in Fig. 10.34a, which by replotting using a logarithmic horizontal scale, then appears as the normal curve in Fig. 10.34b. In this transformation, by using $x = \log t$, a particular skewed distribution in t becomes a normal distribution in x. In this

[16] For a discussion of normal distributions, see Chapter 9.

[17] This kind of distribution is not peculiar to transistor and integrated circuit failures. It has, for example, been observed in one locale for the kilowatt hours of electrical power used versus the number of homes.

FIGURE 10.34

Effect of plotting $f(x)$ versus log x rather than $f(x)$ versus x when the distribution is log normal.



case, the fraction f that fails between time t and $t + dt$ as a function of time t is given by (92)

$$f(t) = \frac{0.4343}{\sigma t\sqrt{2\pi}}\, e^{-(\log t - \log t')^2/2\sigma^2} \qquad 10.21$$

where σ is the dispersion, not in time, but in log time and t' is the median life (MTF). When experimental data are used, log t' is given by $(\Sigma \log t_i)/N$ where the t_i's are the times associated with N experimental observations. This expression can be rewritten in the more conventional form as

$$\log [(t_1 \cdot t_2 \cdot t_3 \cdot \cdots \cdot t_N)^{1/N}] \qquad 10.22$$

where the bracketed term is referred to as the geometric mean. For the case of a log normal distribution, the geometric mean and the median are the same. Note that in Eq. 10.21 a $0.4343/t$ term appears that does not appear in the normal distribution. The t arises because during the transformation a $d(\log t)dt$ is involved. Since most experimental data are plotted on $\log_{10}$ graph paper, the 0.4343, or $\log_{10} e$, is needed to allow the direct use of $\log_{10}$ data.

The instantaneous fractional failure rate $\lambda(t)$ is the decrease in the fraction of surviving units per unit time at time t. That is (92),

$$\lambda(t) = \frac{f(t)}{1 - F(t)} \qquad 10.23$$

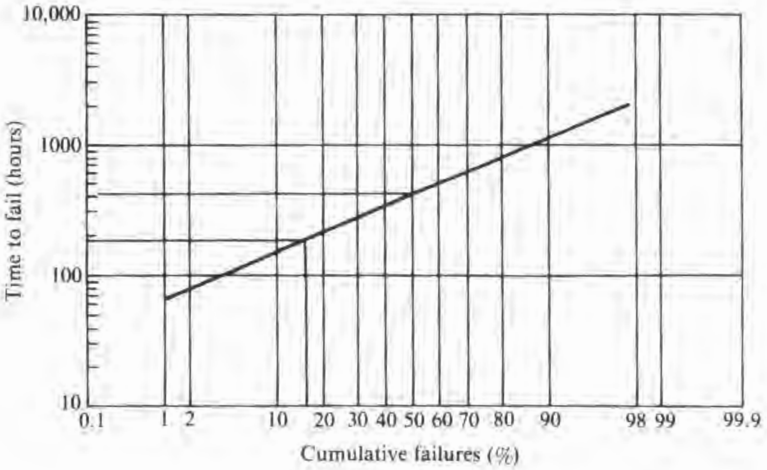
where $F(t)$ is the fraction of initial units that have failed at time t and is given by

$$F(t) = \int_0^t f(t)dt \qquad 10.24$$

The unit of λ from Eq. 10.23 is $1/t$, but typical terminology uses "device hours," with a common unit being FIT. One FIT is defined

FIGURE 10.35

Log normal plot of accelerated lift test data, where the median life (time for 50% failures) is 410 hours and σ for a log normal distribution is given by ln (time for 50% fail/time for 15.9% fail) = ln (410/180) = 0.82.



as one failure in 10^9 device hours and is, for example, equivalent to 0.1% failures per million hours.

While t' (MTF) can be, in principle, determined from Eq. 10.19, varying experimental conditions are hard to account for, and σ is not very amenable to calculation. Thus, failure rate predictions are almost always made from experimental cumulative failure data plotted on log probability paper[18] as shown in Fig. 10.35, from which both t' and σ can be read.

10.8.2 Corrosion

In the presence of moisture, corrosion of many metals used for leads can occur during IC operation. Polysilicon leads are immune to such effects, and silicides appear to be. Most corrosion is caused by electrolysis arising from a combination of moisture leaking into the chip, a water-soluble ionizable material either on the chip surface or along the moisture path, and electric current flow. Moisture most commonly reaches the chip by traveling along the leads of plastic packages to the bonding pads. Then, if cracks exist in the passivation layer or if the layer is a heavily doped phosphorus oxide (93), moisture can penetrate the coating and provide a conduction path between leads. Ions can be from contamination not removed from the chips before assembly, from phosphorus leaching from a phosphosilicate (PSG) overcoat (94), or from the package itself.

[18] One scale is based on cumulative percent, which, from Eq. 10.24, involves an integral of the form

$$\int_0^w e^{-z^2} dz = erf(w)$$

To make the time-to-fail plot linearly, a scale based on erf^{-1} of cumulative percent is used.

Aluminum electrolytic corrosion can be either anodic (lead-biased positive) or cathodic (lead-biased negative), but it is generally much more pronounced on negative leads (95, 96). In either case, the result is conversion of the aluminum to aluminum hydroxide [$Al(OH)_3$], a nonconductor. The anodic rate is temperature independent over the current density range studied (0.32–3.5 mA/cm^2), and one atom of aluminum is oxidized for each three electrons supplied. As the temperature increases, the cathodic reaction rate first increases with an activation energy of 0.47 eV and is relatively independent of current density. At some point, depending on the current density, the rate rather abruptly becomes temperature independent and depends only on current density, with one aluminum metal atom being converted to hydroxide for each electron supplied. For a current density of 0.32 mA/cm^2, the break occurs just below 40°C (97). Various chemical reactions have been suggested (90, 97, 98), but they all involve forming OH^- ions that then react with the Al to form $Al(OH)_3$.

Continued corrosion (oxidation) will ultimately break lead continuity, and, in addition, the increased volume of $Al(OH)_3$ will cause additional cracks in the protective overcoat. It has even been observed that wider leads sometimes fail sooner because they cause more overcoat cracking than narrow leads (99). Even if there are initially no cracks or moisture permeability of the layer, if moisture reaches the bonding pads, their oxidation can cause the raising and cracking of the adjacent overcoat, thus exposing the much smaller leads to corrosion as well.

Both gold and platinum grow dendrites in the presence of moisture, which can cause shorting of leads (100), and the titanium used in silver–titanium solar cell contacts gradually converts to TiO_2 (101).

10.8.3 Internal Stresses

Most lead materials will have high internal stresses induced during the deposition process. In addition, the layered construction, with constituent expansion coefficients ranging from 5×10^{-7}/°C (SiO_2) to 2.3×10^{-5}/°C (Al) will generate additional thermal stresses as the IC is thermally cycled. The ductility of lead materials such as aluminum will allow them to accommodate some stress, but if the stress is compressive, that accommodation will often be in the form of hillocks, which can punch through interlevel insulators and cause shorts. If the stress is tensile, as, for example, in going over steps, voids may result (90, 102).

10.9 DEPOSITION METHODS

The three usual methods of depositing lead material are evaporation, sputtering, and CVD. (For a general description of these processes and the kind of equipment used, see Chapter 4.) In some specialized cases, electrodeposition is used, and focused ion beams can be used for very slowly and directly depositing leads on the wafer in the desired pattern.

10.9.1 Surface Cleanup

Before deposition, the surface must be clean of oxide films that might prevent good contact, organic films that impair metal adhesion, and particulates that prevent good coverage or cause pinholes in overlaying interlevel insulators. (General surface cleanup procedures were discussed in Chapter 3.)

Contact windows are particularly difficult to clean since the oxide step tends to collect debris. When brushes are used for surface cleaning, if the water supply is not adequate or the brushes are too close to the surface, brush shavings will be left at the edge of the windows. Surface tension makes it difficult to satisfactorily wet etch very fine features, and plasma etching is often used instead. However, plasmas may not be able to remove some residues and often leave a thin oxide on silicon surfaces. If the contaminants are only organic, an ozone treatment (sometimes used for resist stripping) can be used for removal and not produce excessive oxide (103). In situ ion milling is also sometimes used as a final clean. More often, a quick dip in an weak HF etch, such as 1 HF to 100 H_2O, either buffered or unbuffered, followed by spin rinsing in DI water, is used just before insertion into the deposition chamber.

Rinse/dry equipment can reintroduce surface contamination if, for example, the drying gas (usually nitrogen) is contaminated with oil or particulates or if the rinse water is not completely removed (104). Particulates on the surface can also arise from broken wafers in spin-rinse/dryers and from particles that collect in the bottom of the deposition chamber and are then stirred up by gases entering during the venting operation. The latter occurs after the deposition, but often the metal film will be hot enough to cause the particles to adhere extraordinarily well. Spattering of an evaporation source due to excessive evaporation rates can cause globules of the evaporant to be attached to the film.

10.9.2 Metal Evaporation

While the vacuum equipment is essentially the same for evaporating any material, several kinds of evaporate sources may be used. Probably the oldest, and applicable only to metals, is the use of a filament heater comprised of several wires twisted together and heated by a high electric current. The metal to be evaporated is in wire form and wrapped around the filament. When the metal is heated, it melts, wets the filament surface, and evaporates. Heating materials must have a higher melting point than that of the evaporate and must not alloy appreciably with it. As a simple alternative, the filament may be a strip formed into a boat so that the evaporant can be contained, thus removing the restriction of wetability as well as allowing chunks or powders to be used. Both of these sources suffer from the inability to hold the large amounts of material necessary for rapid sequential evaporations while using loadlock systems. Further, filaments are prone to rapid burnout. To solve these problems, as well as filament/evaporant compatibility, large water-cooled crucibles with the center portion heated by an electron beam are often used (E-beam evaporation). Alternatively, a nonconductive, noncontaminating crucible can be used and a contained conductive evaporant heated inductively (IN-Source).[19] To provide yet more capacity, additional evaporant in the form of spooled wire can be housed inside the vacuum chamber and periodically reeled off into the crucible.

When multiple-component evaporations are desired–for example, for aluminum containing a few percent copper or silicon—control of the final composition is difficult because of the differences in vapor pressure. If the vapor pressures of the two constituents are not the same, then one will vaporize more rapidly than the other and gradually change the composition of the melt. Two independent sources can be used, with independently controlled rates to give the desired composition. Thin alternating layers can be deposited and then diffused together. Flash evaporation, in which small quantities of each component, such as a fine wire alloy of the correct composition, are continually fed to the evaporation chamber and immediately evaporated, is also sometimes used (105, 106). However, sputtering from a target of the desired composition is more common.

10.9.3 Sputtering

Sputtering is one common way of depositing metal alloys such as Al/Cu or Ti/W when sputtering targets are available. The problems in producing acceptable targets are threefold:

1. It is necessary to maintain a uniform composition throughout; otherwise, as the target is depleted, composition of the sputtered film will change.
2. The target must be physically strong enough to not break during usage and must be amenable to being attached to a backing plate.
3. The fabrication of the target, usually by powder metallurgy, must not introduce undesirable impurities. Pure or nearly pure metals such as platinum, gold, and doped aluminum seldom are a problem for any size target. However, brittle materials such as silicides are likely to have higher breakage rates.

[19] Trademark of Applied Materials, Inc.

When multiple layers are deposited, a series of sputter targets can be sequentially moved into place more easily than multiple evaporation sources. Sputtering also offers a distinct advantage for depositing low-vapor-pressure, high-melting-point materials like Mo, W, and Pt since source heaters of very high temperature are not required. Because deposition pressures are higher in sputtering than they are in evaporation, step coverage is slightly better. However, even though the pressure used is relatively high, provision must be made for pumping down to a hard vacuum in order to clean the system of oxygen and water vapor. In addition, the sputter gas, such as argon, must be very pure. Otherwise, the deposited films may not adhere well, will appear hazy, and will have higher than anticipated resistivity. It was such contamination that prevented early acceptance of sputtering for metallization, but after systems were improved to eliminate excessive residual gases, they were widely accepted.

10.9.4 Chemical Vapor Deposition

As has already been discussed, chemical vapor deposition provides better step coverage than sputtering or evaporation and is often used for interlevel insulator depositions. However, even though CVD of many metals is possible and has been studied for over 50 years, the application to semiconductor processing is quite recent. Metals germane to wafer fabrication that have been vapor phase deposited for other applications include Al, Au, Cr, Cu, Mo, Pt, Ta, Ti, and W (107). The fact that they can be deposited does not necessarily mean that the deposition temperature is low enough to be useful, that the film will adhere properly, that the electrical resistivity will be low enough, or that the surface quality will be satisfactory. The two metals that have been most studied for IC metallization are aluminum and tungsten. In both cases, low-pressure systems operating at a few hundred degrees C are being used.

Aluminum can be vapor deposited from $AlCl_3$ and from several aluminum-bearing organic compounds (107, 108). Tri-isobutyl aluminum thermal decomposition is used most (107–109) and is usually preceded by treating the wafer with $TiCl_4$ in order to promote nucleation. The overall reaction is

$$2Al(C_4H_9)_3 \rightarrow 2Al + 3H_2 + 6C_4H_8$$

and proceeds above about 220°C. Typical LPCVD pressures are in the 0.2–0.5 torr range (108). A rough and nonspecular surface, along with the lack of a process to co-deposit copper or some other material to improve the electromigration characteristics of aluminum, has thus far prevented commercial usage of the LPCVD process.

Tungsten can be deposited by the hydrogen reduction of WF_6, WCl_6, and WBr_6; by the silicon reduction of WF_6 ("displacement reaction"); and by the thermal decomposition of $W(CO)_6$ (tungsten carbonyl) (107, 108). It is relatively easy to produce good tungsten layers by depositing in the 250°C–500°C range, but the problems of integrating the procedure into an IC manufacturing flow have proven numerous. These kinds of problems are by no means unique to tungsten metallization and constitute a good case study of the way in which unexpected interactions can cause difficulties in the introduction of new processes.

WF_6 is the source commonly used for semiconductor application studies. Reactions include (a)

$$WF_6 + 3H_2 \rightarrow W + 6HF$$

and (b)

$$2WF_6 + 3Si \rightarrow 2W + 3SiF_4$$

The application most seriously considered is for via plugs. The deposition of a plug depends on selectively depositing tungsten only in the via openings. Thus, deposition conditions must be chosen so that nucleation does not occur on the interlevel oxide. In principle, reaction b just given is self-limiting at the thickness where appreciable WF_6 can no longer diffuse through the tungsten film to the silicon. However, for selectively depositing on silicon in contact openings, there is no covering of tungsten at the silicon–oxide interface, and considerable unwanted etching of the silicon at the window edges may occur. The effect can be suppressed but not entirely eliminated by the use of hydrogen (reaction a) since reaction b will proceed even with hydrogen present (108). One alternative is to use titanium silicide both as a barrier between the silicon and tungsten and as a means of making a very-low-resistance contact to the silicon. However, the reaction of WF_6 with the Ti in the $TiSi_2$ gives TiF_4, which causes high-resistance contacts (110, 111). When tungsten is used to contact aluminum rather than silicon, a problem similar to that of $TiSi_2$ occurs. That is, through the reaction

$$WF_6 + 2Al \rightarrow W + 2AlF_3$$

a small amount of aluminum fluoride is formed. Aluminum fluoride is a solid at deposition temperatures and forms a high-resistance interfacial layer. Higher temperatures and deposition rates minimize this effect (112).

Molybdenum and silicide CVD depositions for IC metallization applications have also been studied (113, 114). The silicides can, in principle, be deposited by the simultaneous reduction of a metal compound and a silicon compound. The thermal decomposition of silane (SiH_4) is a practical source of silicon since the reaction will proceed at temperatures of 400°C or less.

10.10 PROCESS CONTROL

The metallization parameters regularly monitored include the width of the leads (readily measured by commercial equipment) and the metal thickness and/or sheet resistance. In some cases, the metallization resistivity is itself of importance, but usually resistivity measurements are an alternative to more conventional thickness measurement. The resistivity ratio (defined in the next section), often used to check metal deposition processes as they are developed, also has merit as a process control tool.

Step height measurements using a stylus profilometer are easy when the step is more than a few thousand angstroms thick. Thus, after a typical metal leads layer has been patterned, its thickness can be readily determined. For very thin layers, however, such as platinum to be used for platinum silicide contacts, either optical transmissivity or sheet resistance measurements are more appropriate. Sheet resistance test patterns such as have been described for diffusion sheet resistance measurements can be used either on separate test wafers or on drop-in test chips (115). Optical transmissivity measurements require separate transparent wafers. Calibration curves relating the resistance or optical transmission to thickness are often not constructed. Rather, for process control, a range of the measured parameter that gives satisfactory results is specified.

When metallization composition is changed, bondability testing may also be required as part of the new metal qualification. The addition of silicon to aluminum metallization, for example, makes the bonding of wires to it considerably more difficult. The testing details will depend on the specific bonding procedure being used.

10.10.1 Resistivity Ratio

The resistivity ρ can be approximated by the sum of three terms:

$$\rho \cong \rho_1(T) + \rho_2(G) + \rho_3(I) \qquad 10.25$$

where ρ_1 is a function of temperature, ρ_2 depends on crystallographic disorder such as grain boundaries, and ρ_3 depends on the impurities present. For temperatures near absolute zero, ρ_1 becomes small, and ρ is dominated by ρ_2 and ρ_3. Normally, by room temperature, ρ_1 will be much larger. Only when disorder or impurities become excessive will ρ_2 and ρ_3 become noticeable. Thus, the resistivity ratio

$$RR = \frac{\rho_{\text{room}}}{\rho_{4.2\text{K}}} \cong \frac{\rho_1}{\rho_2 + \rho_3} \qquad 10.26$$

is a measure of the resistivity quality and can be used to evaluate film deposition methods (116). Typical ratios are around 30. Lower numbers indicate a more than normal amount of defects and/or impurities.

10.10.2 Current Capacity

Current-carrying capability testing is generally done on a continuing basis, with samples being periodically collected and subjected to accelerated electromigration testing. The need for routine checking arises because of the sensitivity of electromigration to various processing parameters such as film grain size and composition. Usually, electromigration testing is done by packaging test chips, but it can, in principle, be done in a limited fashion at the wafer level (117). Better control of the silicon temperature is possible because heat sinking is not through a package, subsequent examination and failure analysis are easier, and the time from wafer completion to test is much shorter. Unfortunately, the equipment cost per lead tested is much more expensive than that used in packaged testing.

The likelihood of burnout due to sudden current surges ("zotting") is not as amenable to routine testing and generally involves discharging a capacitor between probes on either side of suspected weak links. Such weak spots will usually be where the metal goes over steps, but some may be due to excessively narrow leads caused by lithography or etching problems. The capacitor size, the series resistance, and the capacitor voltage must be tailored to specific conditions. To simulate the electrostatic behavior of a person, a 150 pF capacitor in series with 2000 Ω is satisfactory (118). The voltage used may be as high as 10 kV, depending on just what the environmental conditions are. A much larger capacitor and a smaller resistance are needed when radiation-hardened circuits or those to be used in adverse industrial applications are tested.

10.10.3 Lead Adherence

On a much less frequent basis, film adhesion to its underlying substrate is measured. One of the oldest tests is the use of Scotch tape to see whether adhesion is better between tape and film or between film and substrate. Such a test is very subjective and can give only go/no-go results. The results depend on the adhesive qualities of the tape (which may vary from batch to batch), the peel rate, and the peel angle (119). A somewhat more scientific approach is to bond a wire to the surface with epoxy and then measure the force required to pull the film off the substrate (providing the glue–film interface does not fail first). The shear strength between film and substrate can, in principle, be determined quantitatively by dragging a sharp stylus across the film and increasing the force on the stylus until the film is scraped away at the bottom of the groove, leaving bare substrate. The shearing force F_s is given by (120)

$$F_s = \frac{a}{\sqrt{r^2 - a^2}} - P \qquad 10.27$$

where r is the radius of the stylus tip, P is the indentation hardness of the substrate. $a = \sqrt{w/\pi P}$ where w is the stylus force required for a clean scrape. The stylus must be very hard, and its tip very small in order to not require an excessive force. For aluminum, a minimum thickness of 1000 Å is recommended. The substrate should be a minimum of 1/8 inch thick to minimize deflection, and if the stylus is a diamond phonograph needle with a 0.7 mil radius, a force of less than 300 grams will be required (121).

10.10.4 Process Problems Requiring Test Patterns for Detection

The incidence of simple process-induced defects must be monitored in order to control the process and to assess the yield loss associated with the metallization steps. These defects primarily consist of open leads, shorts between adjacent leads, and high-resistivity metal-to-metal contacts at vias. For any practical process, such defects occur very infrequently, so any testing must involve long lengths of interconnect and many vias. Historically, patterns involving an appropriate number of potential failure sites have been used, with each pattern then being checked for failure. However, on a given pattern, the number of failures was not determined. As an alternative, by deliberately introducing a resistor between various segments of the pattern and then measuring the resistance after processing, the number of defects per pattern can be estimated (122).

CHAPTER 10 SAFETY

Safety precautions in metallization are primarily centered in making sure that the vacuum chambers do not collapse and that noxious fumes from CVD depositions are properly contained. At low pressure, approximately 15 pounds per square inch press the vacuum chamber inward. Metal chambers give little danger even if the pressure does cause the chamber to collapse. With the glass jars that used to be common, failure of the brittle glass caused the glass fragments to be propelled toward the center of the enclosure (an implosion). Their momentum then carried them on past the center and, unless an arresting metal safety shield was in place, out the other side, where extensive damage could occur. The CVD equipment sometimes used for metal deposition may use hydrogen as a carrier, in which case there is risk of fire or explosion. (These issues were discussed in Chapter 4.) In addition, many of the source materials and deposition by-products are corrosive and/or poisonous. Thus, appropriate storage and exhaust facilities must be provided.

CHAPTER 10 KEY IDEAS

□ The lower the resistivity of the semiconductor, the lower the metal–semiconductor contact resistance. To make useful contacts, the doping level should generally be greater than 10^{19} atoms/cc.

□ Aluminum is the most commonly used contact to silicon. To make good contact, it needs to be alloyed or sintered. These processes cause dissolution of the silicon and are crystallographic-plane sensitive.

□ Many silicides form lower-resistance contacts to silicon than does aluminum.

□ When either metals or silicides contact high-resistivity silicon or gallium arsenide, a Schottky barrier diode is formed. Schottky diodes are majority carrier devices and offer some circuit advantages over pn junction diodes.

□ As lead geometries get smaller, lower-resistivity lead and gate materials are required.

□ In many cases, it is desirable to subject leads to temperatures of several hundred degrees C after they are formed. This fact has lead to the use of polysilicon, silicides, or refractory metals for some interconnects.

□ The high current densities found in IC leads ($>10^5$ A/cm^2) make them susceptible to electromigration failure. Copper- or silicon-doped aluminum is better than pure aluminum, and gold is better than doped aluminum.

□ Plastic packages allow small amounts of moisture to penetrate to the chip. If the moisture contacts aluminum leads, electrolytic action will cause their oxidation and ultimate failure.

CHAPTER PROBLEMS 10

1. For the case of $\rho = 1\ \Omega$-cm and $a = 5\ \mu$m, plot r_s versus ℓ of Eq. 10.3 over enough range of ℓ for r_s to essentially vary from the value given by Eq. 10.1 to that given by Eq. 10.4. Based on this curve, what ℓ/a ratio is required to approximate an infinite thickness?

2. What would the pits left in sintered aluminum contacts on a (110) silicon wafer be expected to look like? Explain your answer.

3. If a 1 μm thick aluminum pad is used for a contact to silicon, what thickness of silicon will be removed if the contact is sintered at 450°C and if the removal is uniform over the entire contact area? Which wafer orientation would be most likely to produce a flat aluminum–silicon interface?

4. If two adjacent silicon Schottky diodes of the same area are made, one of Al and one of PtSi, if n of each of them is 1.05, and if the forward voltage of the aluminum diode is 1 μA at 0.35 V, what will be the PtSi diode voltage when 100 μA are flowing?

5. What is the thickness change when $MoSi_2$ is made from a 3000 Å layer of Mo on a 10,000 Å thick layer of polysilicon? The density of $MoSi_2$ is ~6 grams/cm^3.

6. If gold leads are to be used, what are two potential problems?

7. How much would the MTF be expected to decrease if the current through an interconnect were increased from 10μA to 1 mA?

8. If a lead (interconnect) system is experimentally observed to have a 2000 hours mean time to failure at 215°C and an activation energy of 0.7 eV, what would be the mean time to failure if the temperature were dropped to 55°C?

9. If, because of moisture, a thick aluminum lead is partially anodized to Al_2O_3, would a polyimide or an SiO_2 overcoat be most likely to crack during anodization? Why? List a lead material less susceptible to moisture-induced failure.

CHAPTER REFERENCES 10

1. Paul Ho, "Integrated Circuit Metallization," in 1985 CEI Semiconductor Materials and Process Technologies course.
2. D.P. Kennedy and P.C. Murley, *IBM J. Res. Develop. 12*, pp. 242–250, 1968.
3. G. D'Andrea and H. Murrmann, "Correction Terms for Contacts to Diffused Resistors," *IEEE Trans. on Electron Dev. ED-17*, pp. 481–482, 1970.
4. H.H. Berger, "Models for Contacts to Planar Devices," *Solid-State Electronics 15*, pp. 145–158, 1971.
5. M.P. Lepselter and J.M. Andrews, "Ohmic Contacts to Silicon," pp. 159–186, in Bertram Schwartz, ed., *Ohmic Contacts to Semiconductors*, Electrochemical Society, New York, 1969.
6. C.Y. Chang, Y.K. Fang, and S.M. Sze, "Specific Contact Resistance of Metal–Semiconductor Barriers," *Solid-State Electronics 14*, pp. 541–550, 1971.
7. N. Braslau, "Alloyed Ohmic Contacts to GaAs," *J. Vac. Sci. Technol. 19*, pp. 803–807, 1981.
8. H.H. Berger, "Contact Resistance and Contact Resistivity," *J. Electrochem. Soc. 119*, pp. 507–514, 1972. (This paper includes a review of other methods of measuring contact resistance. See included references.)
9. R. Stall et al., "Ultra Low Resistance Ohmic Contacts to n-GaAs," *Electronics Letters 15*, pp. 800–801, 1979.
10. J.W. Faust et al., "Molten Metal Etches for the Orientation of Semiconductors by Optical Techniques," *J. Electrochem. Soc. 109*, pp. 824–828, 1962.
11. Tchang-Il Chung, "Study of Aluminum Fusion into Silicon," *J. Electrochem. Soc. 109*, pp. 229–234, 1962.
12. P.A. Totta and R.P. Sopher, "SLT Device Metallurgy and Its Monolithic Extension," *IBM J. Res. Develop. 13*, pp. 226–238, 1969.
13. Harry Sello, "Ohmic Contacts and Integrated Circuits," pp. 277–298, in Bertram Schwartz, ed., *Ohmic Contacts to Semiconductors*, Electrochemical Society, New York, 1969.
14. R.J. Anstead and S.R. Floyd, "Thermal Effects on the Integrity of Aluminum to Silicon Contacts in Silicon Integrated Circuits," *IEEE Trans. on Electron Dev. ED-16*, pp. 381–386, 1969.
15. George L. Schnable and Ralph S. Keen, "Aluminum Metallization—Advantages and Limitations for Integrated Circuit Applications," *Proc. IEEE 57*, pp. 1570–1580, 1969.
16. J.M. McCarthy, "Failure of Aluminum Contacts to Silicon in Shallow Diffused Transistors," *Microelectronics and Reliability 9*, pp. 187–188, 1979.
17. Arthur J. Learn, "Evolution and Current Status of Aluminum Metallization," *J. Electrochem. Soc. 123*, pp. 894–906, 1976. (This review paper has over 250 references in it.)
18. L.A. Berthoud, "Aluminium Alloying in Silicon Integrated Circuits," *Thin Solid Films 43*, pp. 219–327, 1977.
19. R.A. Levy et al., "In-Source Al–0.5% Cu Metallization for CMOS Devices," *J. Electrochem. Soc. 132*, pp. 159–168, 1985.
20. B.L. Kuiper, U.S. Patent 3,382,568.
21. S.P. Bellier and L.B. Ehlert, "An Improved Metallization Process for Silicon Transistors," pp. 304–314, in Howard R. Huff and Ronald R. Burgess, eds., *Semiconductor Silicon/73*, Electrochemical Society, Princeton, N.J., 1973.
22. H.M. Naguib and L.H. Hobbs, "Al/Si and Al/Poly-Si Contact Resistance in Integrated Circuits," *J. Electrochem. Soc. 124*, pp. 573–577, 1977.
23. H.M. Naguib and L.H. Hobbs, "The Reduction of Poly-Si Dissolution and Contact Resistance at Al/n-Poly-Si Interfaces in Integrated Circuits," *J. Electrochem. Soc. 125*, pp. 169–171, 1978.
24. M. Eizenberg, "Applications of Thin Alloy Films in Silicon Contacts," pp. 348–360, in Kenneth E. Bean and George A. Rozgonyi, eds., *VLSI Science and Technology/84*, Electrochemical Society, Princeton, N.J., 1984.
25. A.S. Berezhnoi, *Silicon and Its Binary Systems*, Consultants Bureau, New York, 1960.
26. M.P. Lepselter, "Beam-Lead Technology," *Bell Syst. Tech. J. 45*, pp. 233–253, 1966.

27. S. Yanagisawa and T. Fukuyama, "Reaction of Mo Thin Films on Si (100) Surfaces," *J. Electrochem. Soc. 127*, pp. 1150–1156, 1980.
28. V.I. Rideout, "A Review of the Theory and Technology for Ohmic Contacts to Group III–V Compound Semiconductors," *Solid-State Electronics 18*, pp. 541–550, 1975.
29. R.H. Cox and H. Strach, "Ohmic Contacts for GaAs Devices," *Solid-State Electronics 10*, pp. 1213–1218, 1967.
30. N. Braslau, J.B. Gunn, and J.L. Staples, "Metal–Semiconductor Contacts for GaAs Bulk-Effect Devices," *Solid-State Electronics 10*, pp. 381–385, 1967.
31. Ajit Rode and J. Gordon Roper, "Gallium Arsenide Digital IC Processing—A Manufacturing Perspective," *Solid State Technology*, pp. 209–215, February 1985.
32. M. Heiblum et al., "Characteristics of AuGeNi Ohmic Contacts to GaAs," *Solid-State Electronics 25*, pp. 185–195, 1982.
33. P.A. Barnes et al., "Ohmic Contacts Produced by Laser Annealing Te-Implanted GaAs," *Appl. Phys. Lett. 33*, pp. 965–967, 1978.
34. P.A. Pianetta et al., "Non-Alloyed Ohmic Contacts to Electron-Beam Annealed Se-Ion-Implanted GaAs," *Appl. Phys. Lett. 36*, pp. 597–599, 1980.
35. Y.I. Nissim et al., "Non-Alloyed Contacts in n-GaAs by CW Laser Assisted Diffusion from a SnO_2/SiO_2 Source," *IEEE Trans. on Electron Dev. ED-28*, pp. 607–609, 1981.
36. H.R. Grinolds and G.V. Robinson, "Pd/Ge Contacts to n-Type GaAs," *Solid-State Electronics 23*, pp. 573–585, 1980.
37. E.D. Marshall et al., "Non-Alloyed Ohmic Contacts to n-GaAs by Solid Phase Epitaxy," *Appl. Phys. Lett. 47*, pp. 298–300, 1985.
38. P.A. Barnes and A.Y. Chou, "Non-Alloyed Ohmic Contacts to n-GaAs by Molecular Beam Epitaxy," *Appl. Phys. Lett. 33*, pp. 651–653, 1978.
39. W.T. Tsang, "In Situ Ohmic-Contact Formation to n- and p-GaAs by Molecular Beam Epitaxy," *Appl. Phys. Lett. 33*, pp. 1022–1025, 1978.
40. J.V. DiLorenzo et al., "Non-Alloyed and In Situ Ohmic Contacts to Highly Doped n-Type GaAs Grown by Molecular Beam Epitaxy for Field Effect Transistors," *J. Appl. Phys. 50*, pp. 951–954, 1979.
41. W.T. Anderson et al., "Development of Ohmic Contacts for GaAs Devices Using Epitaxial Germanium Films," *IEEE J. Solid-State Circuits SC-13*, pp. 430–435, 1978.
42. R.T. Galla et al., "Evaluation of the Interfacial Resistance of Thin Film Interconnections," *Microelectronics and Reliability 7*, pp. 185–212, 1968.
43. A.E. Michel et al., "Base Contacts for High-Speed Germanium Transistors," pp. 243–252, in Bertram Schwartz, ed., *Ohmic Contacts to Semiconductors*, Electrochemical Society, New York, 1969.
44. W. Shockley, "Research and Investigation of Inverse Epitaxial UHF Power Transistors," *Final Tech. Rpt. Al-TDR-64-207*, Air Force Avionics Laboratory, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio, 1964.
45. G.K. Reeves, "Specific Contact Resistance Using a Circular Transmission Line Model," *Solid-State Electronics 23*, pp. 487–490, 1980.
46. Chung-Yu Ting and Charles Y. Chen, "A Study of the Contacts of a Diffused Resistor," *Solid-State Electronics 14*, pp. 433–438, 1971.
47. A.Y.C. Yu and E.H. Snow, "Surface Effects on Metal–Silicon Contacts," *J. Appl. Phys. 39*, pp. 3008–3016, 1968.
48. M.P. Lepselter and S.M. Sze, "Silicon Schottky Barrier Diode with Near-Ideal *I–V* Characteristics," *Bell Syst. Tech. J. 47*, pp. 195–208, 1968.
49. A.Y.C. Yu and C.A. Mead, "Characteristics of Aluminum–Silicon Schottky Barrier Diode," *Solid-State Electronics 13*, pp. 97–104, 1970.
50. S.M. Sze, *Physics of Semiconductor Devices*, 2d ed., John Wiley & Sons, New York, 1981.
51. Howard C. Card, "Aluminum–Silicon Schottky Barriers and Ohmic Contacts in Integrated Circuits," *IEEE Trans. on Electron Dev. ED-23*, pp. 538–544, 1976.
52. J.M. Shannon, "Control of Schottky Barrier Height Using Highly Doped Surface Layers," *Solid-State Electronics 19*, pp. 537–543, 1976.
53. P.B. Ghate, "Electromigration-Induced Failures in VLSI Interconnects," pp. 292–299, in *Proc. 20th Annual IEEE Reliability Physics Symposium*, March 1982.
54. R.W. Keyes, "The Evolution of Digital Electron-

ics toward VLSI," *IEEE Trans. on Electron Dev. ED-26*, pp. 271–278, 1979.

55. Krishna C. Saraswat, "Effect of Scaling of Interconnections on the Time Delay of VLSI Circuits," *IEEE Trans. on Electron Dev. ED-29*, pp. 645–650, 1982.
56. William E. Engeler and Dale M. Brown, "Performance of Refractory Metal Multilevel Interconnection System," *IEEE Trans. on Electron Dev. ED-19*, pp. 54–61, 1972.
57. N. Yamamoto et al., "Fabrication of Highly Reliable Tungsten Gate MOS VLSIs," *J. Electrochem. Soc. 133*, pp. 401–407, 1986.
58. L.L. Vadasz et al., "Silicon Gate Technology," *IEEE Spectrum 6*, pp. 28–35, 1969.
59. V. Ramakrishna et al., "Future Requirements for High-Speed VLSI Interconnections," pp. 27–32, in *Proc. IEEE VLSI Interconnection Conference*, Santa Clara, Calif., 1987.
60. Malcomb Beasley, "Superconductor Material Development for VLSI Interconnect," Multilevel Interconnection State-of-the-Art Seminar, Santa Clara, Calif., 1987.
61. Joseph W. Goodman et al., "Optical Interconnections for VLSI Systems," *Proc. IEEE 72*, pp. 850–866, 1984.
62. R. Selvaraj et al., "Optical Interconnections Using Integrated Waveguides in Polyimide for Wafer Scale Integration," pp. 306–313, in *Proc. IEEE VLSI Interconnection Conference*, Santa Clara, Calif., 1987.
63. P.B. Ghate et al., "Application of Ti:W Barrier Metallization for Integrated Circuits," *Thin Solid Films 53*, pp. 117–128, 1978.
64. P.B. Ghate, "Metallization for Very Large-Scale Integrated Circuits," *Thin Solid Films 93*, pp. 359–383, 1982.
65. M.A. Nicolet and M. Bartur, "Diffusion Barriers in Layered Contact Structures," *J. Vac. Sci. Technol. 19*, pp. 786–793 and references therein, 1981.
66. B. Lee et al., "Effect of Oxygen on the Diffusion Properties of TiN," pp. 344–350, in *Proc. IEEE VLSI Interconnection Conference*, Santa Clara, Calif., 1987.
67. M.-A. Nicolet, "Diffusion Barriers in Thin Films," *Thin Solid Films 52*, pp. 415–554, 1978.
68. Joel R. Shappirio, "Diffusion Barriers in Advanced Semiconductor Device Technology," *Solid State Technology*, pp. 161–166 and references therein, October 1985.
69. N. Kumar et al., "Fabrication of RF Reactively Sputtered TiN Thin Films," *Semiconductor International*, pp. 100–104, April 1987.
70. M. Whittmer et al., "Effect of Cu on the Kinetics and Microstructure of Al_3Ti Formation," *J. Electrochem. Soc. 132*, pp. 1450–1455, 1985.
71. I.A. Blech, "Evaporated Film Profiles over Steps in Substrates," *Thin Solid Films 6*, pp. 113–118, 1970.
72. W.G. Oldham et al., "A General Simulator for VLSI Lithography and Etching Processes: Part II—Application to Deposition and Etching," *IEEE Trans. on Electron Dev. ED-27*, pp. 1455–1459, 1980.
73. A.C. Adams, "Plasma Deposition of Inorganic Films," *Solid State Technology*, pp. 135–139, April 1983.
74. Yoshio Homma and Sukeyoshi Tsunekawa, "Planar Deposition of Aluminum by RF/DC Sputtering with RF Bias," *J. Electrochem. Soc. 132*, pp. 1466–1472, 1985.
75. C.Y. Ting, "Silicides for Contacts and Interconnects," Tech. Digest Int. Electron Dev. Meeting, San Francisco, December 1984.
76. Pieter Burggraaf, "Silicide Technology Spotlight," *Semiconductor International*, pp. 293–298, May 1985.
77. A.E. Morgan et al., "Characterization of a Self-Aligned Cobalt Silicide Process," *J. Electrochem. Soc. 134*, pp. 925–935 and references therein, 1987.
78. A.I. Akinwande and J.D. Plummer, "Process Dependence of the Metal Semiconductor Work Function Difference," *J. Electrochem. Soc. 134*, pp. 2297–2303 and references therein, 1987.
79. Papers in *Proc. IEEE VLSI Interconnection Conference*, Santa Clara, Calif., 1987.
80. C.Y. Ting et al., "Study of Planarized Sputter-Deposited SiO_2," *J. Vac. Sci. Technol. 15*, pp. 1105–1112, 1978.
81. Gregory C. Smith and Andrew J. Purdes, "Sidewall-Tapered Oxide by Plasma-Enhanced Chemical Vapor Deposition," *J. Electrochem. Soc. 132*, pp. 2721–2725, 1985.
82. J.S. Mercier et al., "Dry Etch-Back of Overthick PSG Films for Step-Coverage Improvement," *J. Electrochem. Soc. 132*, pp. 1219–1222, 1985.

83. A.C. Adams and C.P. Capio, "Planarization of Phosphorus-Doped Silicon Dioxide," *J. Electrochem. Soc. 128*, pp. 423–429, 1981.
84. O. Wadi et al., "Mask Preparation for Small Dimension Ion Milling by Two-Step Lift-Off Process," *J. Electrochem. Soc. 124*, pp. 959–960, 1977.
85. T. Sakurai and T. Serikawa, "Lift-Off Metallization of Sputtered Al Alloy Films," *J. Electrochem. Soc. 126*, pp. 1257–1260, 1979.
86. Moshe Oren and A.N.M. Masum Choudhury, "Interconnect Metallization Technique for GaAs Digital ICs," *J. Electrochem. Soc. 134*, pp. 750–752, 1987.
87. H.B. Huntington and A.R. Grone, "Current-Induced Marker Motion in Gold Wires," *J. Phys. Chem. Solids 20*, pp. 88–98, 1961.
88. I.A. Blech et al., "A Study of Failure Mechanisms in Silicon Planar Epitaxial Transistors," *Rome Air Development Center Tech. Rpt. TR 66–31*, December 1965.
89. James R. Black, "Electromigration Failure Modes in Aluminum Metallization for Semiconductor Devices," *Proc. IEEE 57*, pp. 1587–1594, 1969. (See also the included references for background on the work prior to 1969.)
90. P.B. Ghate, "Reliability of VLSI Interconnections," pp. 321–337, in *Proc. American Institute of Physics Conference*, no. 138, New York, 1986.
91. James R. Black, "RF Power Transistor Metallization Failure," *IEEE Trans. on Electron Dev. ED-17*, pp. 800–803, 1970.
92. Frederick H. Reynolds, "Thermally Accelerated Aging of Semiconductor Components," *Proc. IEEE 62*, pp. 212–222, 1974.
93. Robert B. Comizzoli, "Bulk and Surface Conduction in CVD SiO_2 and PSG Passivation Layers," *J. Electrochem. Soc. 123*, pp. 386–391, 1976.
94. Naoyuki Nagasima et al., "Interaction between Phosphosilicate Glass Films and Water," *J. Electrochem. Soc. 121*, pp. 434–438, 1974.
95. H. Koelmans, "Metallization Corrosion in Si Devices by Moisture Induced Electrolysis," pp. 168–171, in *Proc. 12th Annual IEEE Reliability Physics Symposium*, New York, 1974.
96. W.M. Paulson and R.W. Kirk, "The Effects of Phosphorus-Doped Passivation Glasses on the Corrosion of Aluminum," pp. 172–179, in *Proc. 12th Annual IEEE Reliability Physics Symposium*, New York, 1974.
97. E.P.G.T. van de Ven and H. Koelmans, "The Cathodic Corrosion of Aluminum," *J. Electrochem. Soc. 123*, pp. 143–144, 1976.
98. Nicholas Lycoudes, "The Reliability of Plastic Microcircuits in Moist Environments," *Solid State Technology*, pp. 53–62, October 1978.
99. T. Wada et al., "Relationship between Width and Spacing of Aluminum Electrodes and Aluminum Corrosion on Simulated Microelectronic Circuit Patterns," *J. Electrochem. Soc. 134*, pp. 649–653, 1987.
100. E.B. Hakim and J.R. Shappiro, "Failure Mechanisms in Gold Metallized Sealed Junction Devices," *Solid State Technology*, pp. 66–68, April 1975.
101. W.H. Becker and S.R. Pollack, "The Formation and Degradation of Ti–Ag and Ti–Pd–Ag Solar Cell Contacts," pp. 40–50, in *Proc. 8th IEEE Photovoltaic Specialists Conference*, Seattle, 1970.
102. S.K. Groothuis and W.H. Schroen, "Stress Related Failures Causing Open Metallization," pp. 1–8, in *Proc. 25th Annual IEEE International Reliability Physics Symposium*, 1987.
103. H. Norstrom et al., "Dry Cleaning of Contact Holes Using Ultraviolet (UV) Generated Ozone," *J. Electrochem. Soc. 132*, pp. 2285–2287, 1985.
104. Vance Hoffman, "Practical Troubleshooting of Vacuum Deposition Processes and Equipment for Aluminum Metallization," *Solid State Technology*, pp. 47–56, December 1978.
105. Arthur J. Learned, "Aluminum Alloy Film Deposition and Characterization," *Thin Solid Films 20*, pp. 261–279, 1974.
106. Tom Strahl, "Flash Evaporation, An Alternative to Magnetron Sputtering in the Production of High-Quality Aluminum Alloy Films," *Solid State Technology*, pp. 78–82, December 1978.
107. Carroll F. Powell, Chap. 10, "Chemically Deposited Metals," in Carroll F. Powell et al., eds., *Vapor Deposition*, John Wiley & Sons, New York, 1966.
108. R.A. Levy and M.L. Green, "Low Pressure Chemical Vapor Deposition of Tungsten and Aluminum for VLSI Applications," *J. Electro-*

chem. Soc. 134, pp. 37C–49C and references therein, 1987.

109. M.J. Cooke et al., "LPCVD of Aluminum and Al–Si Alloys for Semiconductor Metallization," *Solid State Technology*, pp. 62–65, December 1982.
110. E.K. Broadbent et al., "Growth of Selective Tungsten on Self-Aligned Ti and PtNi Silicides by Low Pressure Chemical Vapor Deposition," *J. Electrochem. Soc. 133*, pp. 1715–1721, 1986.
111. Gregory C. Smith et al., "Damage to $TiSi_2$ Clad Doped Silicon due to CVD Selective Tungsten Deposition," pp. 155–161, in *Proc. IEEE VLSI Interconnection Conference*, Santa Clara, Calif., 1987.
112. R. Chow and S. Kang, "Selective Chemical Vapor Deposition of Tungsten on Aluminum," pp. 208–215, in *Proc. IEEE VLSI Interconnection Conference*, Santa Clara, Calif., 1987.
113. Jim Crawford, "Refractory Metals Pace IC Complexity," *Semiconductor International*, pp. 84–86, March 1987.
114. Suresh Sachdev and Robert Castellano, "CVD Tungsten and Tungsten Silicide for VLSI Applications," *Semiconductor International*, pp. 306–310, May 1985.
115. Sheldon C.P. Lem and Doug Ridley, "An Overview of Thickness Measurement Techniques for Metallic Thin Films," *Solid State Technology*, pp. 99–103, February 1983.
116. C.R. Fuller and P.B. Ghate, "Magnetron-Sputtered Aluminum Films for Integrated Circuit Interconnections," *Thin Solid Films 64*, pp. 25–37, 1979.
117. Janet M. Towner, "Electromigration of Thin Films at the Wafer Level," *Solid State Technology*, pp. 197–200, October 1984.
118. T.M. Madzy and L.A. Price, "Module Electrostatic Discharge Simulator," pp. 36–40, in *Proc. Electrical Overstress/Electrostatic Discharge Symposium*, Rome Air Development Reliability Analysis Center, 1979.
119. D.W. Aubrey et al., *J. Appl. Polymer Sci. 13*, p. 2193, 1969.
120. P. Benjamin and C. Weaver, "Measurement of Adhesion of Thin Films," *Proc. Roy. Soc. A. 254*, pp. 163–176, 1960.
121. Murray Bloom, "Development of Improved Test Standards for Monolithic Circuits," *TRW Final Rpt. NAS8-24388*, 1970.
122. Richard Spencer, "Novel IC Metallization Test Structures for Drop-In Process Monitors," *Solid State Technology*, pp. 201–205, September 1983.

CHAPTER 11

Yields and Yield Analysis

11.1 INTRODUCTION

Three levels of wafer and circuit testing have evolved, two of which are performed in the wafer fabrication facility. The first level of testing consists of measuring a few parameters as soon as possible in order to provide quick feedback on process behavior. These measurements may be either wafer oriented, as are, for example, ion implant sheet resistances, or device oriented, as is the threshold voltage of a MOS test transistor. The results of these tests are generally reflected in the number of wafers passed at a given process inspection point (yield at that process step). The next level is the automatic DC testing (multiprobing) done on each chip after wafer processing is completed but before the wafer is broken into individual chips. The tests performed at this level are primarily designed to screen out chips that are likely to fail the more comprehensive AC/DC/temperature range third level of testing. However, the results are also an important measure of wafer-fab performance (multiprobe yield). Third-level testing is performed in the assembly/test area after packaging, and the primary goal is to ensure that the product meets customer specifications.

In addition to these test points, tests, and yields, there are also several materials-oriented inspection points and yields prior to a slice's entering the wafer fabrication facility. The more salient of these are included in the yield definitions of the next section.

11.2 YIELD DEFINITIONS

Crystal yield, the grams of crystal within specification divided by the grams of polysilicon required, is primarily a measure of how well the crystal resistivity, perfection, and oxygen content are kept within the required range. It is also a measure of how well the equipment is maintained (so that it does not fail during the 24 hours or so that it takes to pull a crystal) and of how wide a range of resistivities the manufacturer is able to sell. A typical value is about 50%. By recycling (remelting, redoping, and repulling) out-of-specification crystals, the silicon yield can be raised to perhaps 65%.

The next yield to be considered is that of converting from silicon crystal to slice. *Crystal-to-slice yield* may be expressed in grams/gram, grams/slice, or grams/cm^2 of slice and depends on how much silicon is lost when the crystal is ground from its as-grown diameter to the slice diameter, how thick the saw blade used for slicing is relative to the slice thickness, and how many slices are broken or rejected for edge chipping, poor polishing, and so on. A typical value is 50%. Thus, about 4 grams of polysilicon for each gram of slice is to be expected.

At this point, the slice enters the wafer processing area, where it becomes a wafer. One definition of *process yield* is the number of wafers to multiprobe produced in a specified time (often one month) divided by the number of starting slices used to produce those wafers. The overall process yield is generally broken into half a dozen or more intermediate yields relating to each of the major process steps. Typical overall process yield is 90%–95%. However, this yield is based only on those slices that are initially earmarked for finished salable units. Substantial quantities of incoming slices are also used for pilots, engineering runs, and so on. When these slices are considered, the overall wafer-fab area wafer out per slice in yield is closer to 70%. Thus, depending on whether process performance is being evaluated or the total number of slices required for some quantity of ICs is being estimated, two different yield numbers are required.

The individual chips on each wafer are checked electrically at multiprobe. The ratio of good chips to the total number of chips on a wafer is that wafer's *multiprobe yield*. The average value for a given circuit can range from a few percent to over 90%, depending on the complexity of the circuit and the proficiency of the wafer-fab area. There are a few pitfalls in determining this yield, and they stem from the determination of how many potential chips are on a wafer. One method of determining total chips is by means of the number of times the probe head comes down to make a measurement. However, this number will often include all of the partial chips on the periphery of the wafer as well as the whole ones over the rest of the wafer. Using this number for the potential chips will give a pessimistic yield number. Using the calculated number[1] of whole chips that can be placed on a wafer gives the most realistic yield number. Often, however, the chips in the outer few millimeters of the edge will be excluded from the potential count and thus give an inflated

[1] Simple programs can be written for either programmable calculators or desktop computers that will give this number. For large chips, their placement on the wafer may make a substantial difference in the number.

TABLE 11.1

Major Semiconductor Manufacturing Yield Points

Yield	*Definition*
Crystal yield*	Polycrystal to single-crystal conversion
Slice yield	Single-crystal boule to polished slice conversion
Process yield	Slice to finished wafer conversion
Multiprobe yield	Ratio of good chips per lot to total chips per lot
Assembly/test yield	Ratio of good finished units to good chips

*In the case of compound semiconductors, such as gallium arsenide, an additional loss point occurs prior to crystal yield. It is at the compounding step, where the constituents gallium and arsenic are combined to give polycrystalline gallium arsenide.

yield number. Multiprobe yield can be considered as the product of the functional yield and the parametric yield. A functional chip is one that "wiggles" (one that will roughly perform the function for which it was intended but that will not necessarily meet all of the specifications required to have a useful circuit). For example, a gate output might go from high to low as the appropriate inputs are changed and thus be functional. However, it might use too much power, switch too slowly, have too much leakage, or not have enough voltage swing and thus not meet the total set of specifications. Such a gate would then be a parametric failure. Ordinarily, the two yields will not be separated, but occasionally they will be for some yield analysis procedures.

After the wafers leave the wafer fabrication facility, they are broken into individual chips. The good chips are packaged and electrically tested to the finished-unit specifications. The *assembly/final test yield* is the number of units passing this final test in some time period divided by the number of chips required to produce those units. This final testing will check both DC and AC performance, usually at several temperatures. The assembly/final test yield will usually be in the 90%–95% range.

All of the yields that have been discussed, along with their definitions, are summarized in Table 11.1.

11.3 METHODS OF YIELD MEASUREMENT

Most wafer fabrication yield determinations are based on either visual examinations, special processing-oriented tests such as were described in the various processing chapters, curve tracer measurements, or multiprobe results. The main use of curve tracer examinations, however, is in follow-up defect analysis. By far, the largest body of yield data is taken by automatic multiprobe equipment.

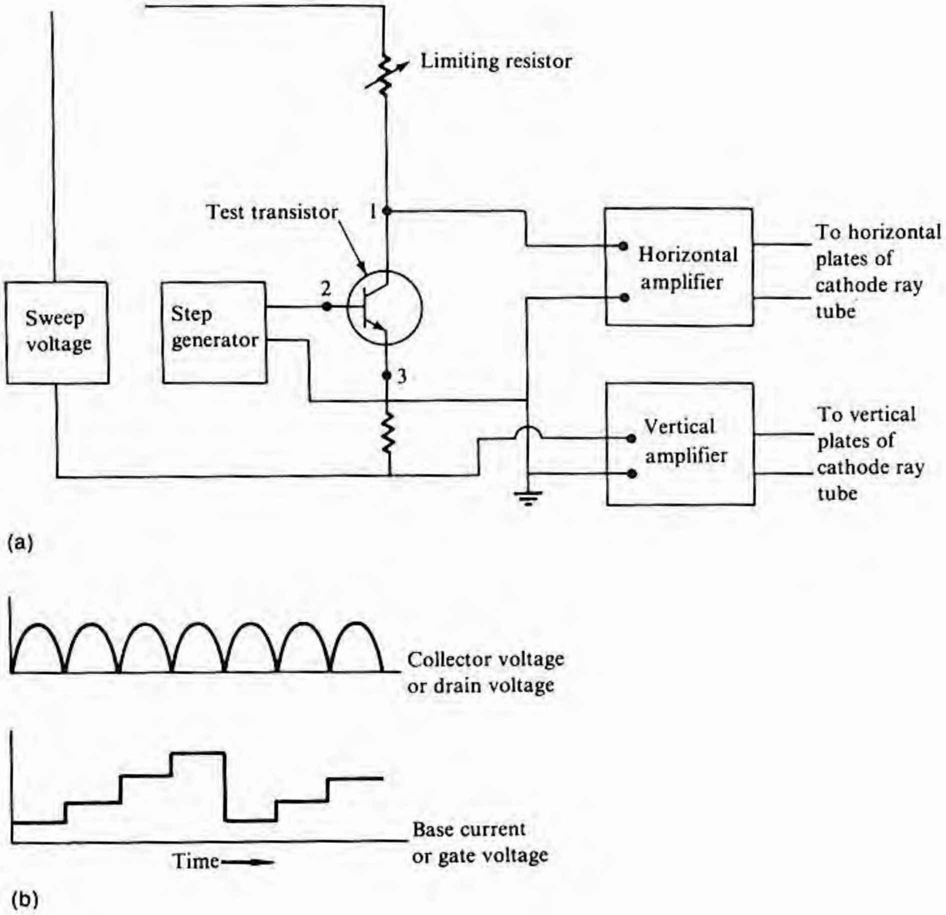
11.3.1 The Curve Tracer

The basic elements of a curve tracer are shown in Fig. 11.1. The sweep voltage will usually be rectified 60 Hz, and the current or voltage step generation synchronized with it as shown in Fig. 11.1. Both the number of steps and the step amplitude may be varied. For diode junction *I–V* measurements, the diode is connected between terminals 1 and 3, and the step generator is not used. For bipolar transistor characteristics, a current step generator is connected to the base. For MOS transistor characteristics, a voltage step generator is connected to the gate.

To facilitate contacting the wafer, a probe station is used that has a chuck to hold the wafer and allow for *x* and *y* motion; two or three fine-wire probes with *x*, *y*, and *z* motion; and a stereoscopic microscope. A microscope and fine-probe wires with sharp points

FIGURE 11.1

Schematic of a typical curve tracer.


Limiting resistor
Test transistor
1
2
3
Sweep voltage
Step generator
Horizontal amplifier
To horizontal plates of cathode ray tube
Vertical amplifier
To vertical plates of cathode ray tube
(a)
Collector voltage or drain voltage
Base current or gate voltage
Time
(b)

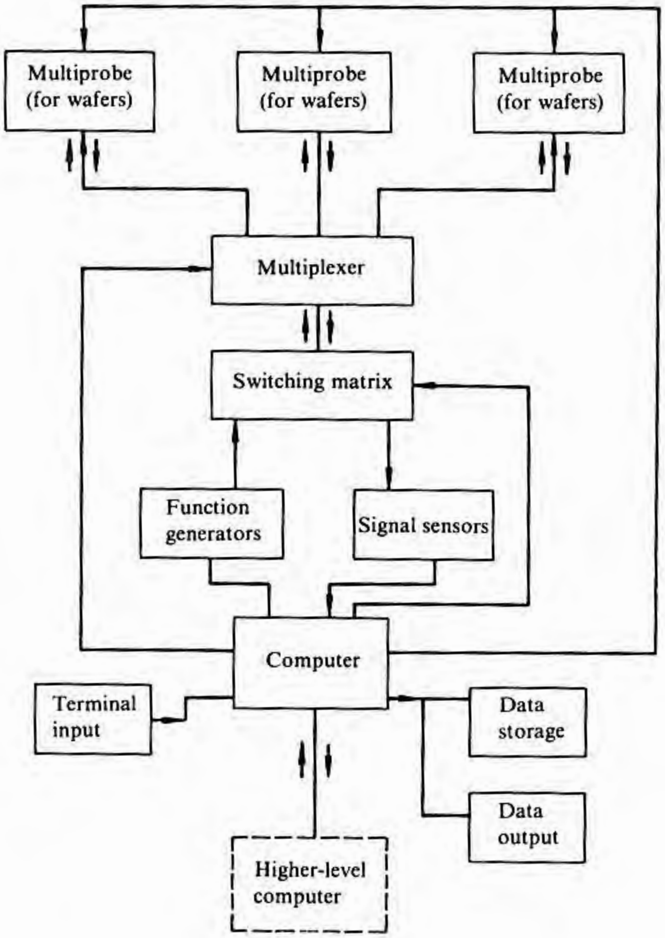
are required since leads may be only 1–3 μm across. When it is necessary to electrically isolate some component so that it can be individually examined, the appropriate interconnecting leads must be broken. A common method is to place two probes close together on a lead and use a capacitor discharge to vaporize the section of lead between the probes. An alternative method of breaking leads is to use an isolation mask (bipolar circuits) or a moat mask (MOS circuits) and etch through all interconnections.

11.3.2 Automatic Multiprobe Testing

Computer-controlled testers are required both to provide the speed necessary to process large volumes of chips and to direct the intricate tests of complex circuits. A block diagram of an automatic test system is shown in Fig. 11.2. The multiprobe station is similar to the

FIGURE 11.2

Block diagram of an automatic test system.



hand prober just described except that it has many more contacting probes, and the chucked wafer can be lowered away from the probes, shifted laterally, and raised again. Thus, the probes can be adjusted once to contact all of the appropriate pads on a chip, and then after each chip is tested, the wafer can automatically be moved to the next position. Initial positioning can either be manual or automatic through the use of pattern recognition technique. As indicated in Fig. 11.2, more than one multiprobe station may be connected to the test system via a multiplexer. Feeding into the multiplexer through a switching matrix is a multitude of function generators and signal sensors. The function generators include such items as DC power supplies and pulse generators.[2] The sensors are digital voltmeters, ammeters, comparators, and so on. The matrix directs the outputs of the function generators to the appropriate probes and connects the required sensors to proper probes. The local computer supplies timing and direction to all components of the system and may itself be tied to a higher-level management computer to provide test specifications and more complex data processing.

A substantial amount of computation may be required during testing since the fail/pass decision may be based on the outcome of various calculations. These may range from a simple calculation of transistor gain from measured currents to a regression analysis of a number of room-temperature measurements in order to predict whether or not the final packaged unit will pass high-temperature specifications.

As chips are tested, rejects are indicated by an ink dot placed in the middle of the chip. Often, two inkers with a different color ink in each are used so that a category in addition to good/bad can be indicated. For example, operational amplifier chips might be sorted into two good categories, depending on open loop gain. When the parameters are read and recorded for all chips on each wafer, the test order will not matter. However, to conserve test time, a "home-on-first-fail" procedure is often used so that valuable test time is not wasted in testing further a unit that has already failed. When this test philosophy is being followed, it is important to order the tests so that the ones most likely to fail are done first. Printouts of each lot's results are usually provided, but if a home-on-first-fail strategy is used, complete data will not be available. From a diagnostic standpoint, since incomplete data is a handicap, a full characteri-

[2] Multiprobe testing is generally referred to as "DC," but in the interest of minimizing test time, the inputs and outputs are really short pulses. The minimum pulse length is determined by the amount of time for the device to reach electrical equilibrium and the *RC* time constant of the connecting lines.

zation program will often be routinely run on a small percentage of wafers.

11.4 YIELD TRACKING

11.4.1 Process and Multiprobe Yield Tracking

Process yield or multiprobe yield can be plotted on a daily or a lot-by-lot basis. Either method gives a time series that can then be fitted with a trend line. Because of the rather substantial daily fluctuations, it is sometimes difficult to tell immediately when a break in the trend line occurs. One method of enhancing the breakpoint is to use the Cumulative Sum method (1, 2), in which not the yield Y_i but a sum S_i of the form

$$S = \sum_{i=1}^{n}(Y_i - Y_0) \qquad 11.1$$

is plotted. When the yields are constant, the curve is a straight line that will have a zero, positive, or negative slope, depending on the value chosen for Y_0. Ordinarily, a long-term yield average would be used for Y_0, although it may be slightly adjusted to give the desired initial slope. With this approach, if yields start drifting downward, there will almost immediately be a sharp break in the curve.

For wafer-fab areas that process many different device types, it may be impossible to maintain device continuity from week to week and thus difficult to adequately track multiprobe yield. One alternative is to track defect density, which is discussed in the next section. A defect density can be determined from multiprobe yield, IC area, and process complexity, which, in principle, is independent of the specific device being built.

11.4.2 Split Lot Testing

When process changes are considered, one of the more common ways of deciding whether such a change will in fact help yields (or costs) is to run a series of split lots. Split wafer lots all have common processing up to some particular step to be evaluated. At that point, the mother lot is split, and the various sublots are treated differently. Then, after the step or steps being evaluated are completed, the sublots are again merged for the remainder of the process. At the end, the various sublots are multiprobed separately, and decisions on the efficacy of the process changes are made on the basis of these yields. An improper interpretation could be very costly. Consequently, it is very important to accurately determine whether or not significant differences exist between various splits. What constitutes a significant difference is not defined in absolute terms, but rather only in how much confidence one must have in order to be satisfied that there is or is not a difference.

The distribution of multiprobe yield (MPY) by wafer for a full 50 wafer lot will ordinarily closely approach a normal distribution and have 1s limits of about 5 yield points.[3] The standard deviation of the multiprobe yield of the lot can be calculated from

$$s = \left[\frac{\Sigma(M - \overline{M})^2}{n}\right]^{1/2} \qquad 11.2$$

It is, however, more easily calculated from

$$s = \left[\frac{\Sigma M^2}{n} - \left(\frac{\Sigma M}{n}\right)^2\right]^{1/2} \qquad 11.3$$

where M is the multiprobe yield of each wafer, $\overline{M}$ is the arithmetic mean of the multiprobe yield of all of the wafers being considered, and n is the number of wafers. When the sample size is small enough for $n - 1$ to be appreciably less than n, $n - 1$ should be substituted for n in order to give an "unbiased" estimate of s.

If a random sample of N wafers is taken from a larger body of wafers with a normal distribution, it too should have a normal distribution. Its mean will not necessarily be the same as that of the whole lot. The dispersion[4] of the sample lot means will be given by

$$s_N = \frac{s}{\sqrt{N}} \qquad 11.4$$

For large N, the dispersion will be very low, and the sample mean will be essentially that of the lot. For $N = 1$, the mean of a lot of 1 is its MPY, and that mean's dispersion is s, or the same as that of the multiprobe yield of the lot. Thus, when the yield of a small split from a considerably larger lot is examined, the expected dispersion of the mean of a split, assuming that it is no different from the rest of the lot, can be calculated by Eq. 11.4. Then, the actual value of the mean of the split can be compared to that of the rest of the lot, and any difference is judged in light of the expected spread.

EXAMPLE □ Suppose that 9 wafers of a full 50 wafer lot had a different flow and that the multiprobe average of the other 41 wafers was 70% with a 1s value of 6 percentage points. What is the significance of a 9 slice MPY average of 74%?

[3] Remember that σ is the standard deviation of some property of a large population of objects and is defined so that $\pm 1\sigma$ from the arithmetic mean of a normal distribution will include 68.3% of the population, $\pm 2\sigma$ will include 95.5%, and $\pm 3\sigma$ will include 99.7%. In this case, the population is a collection of wafers, and the property is the multiprobe yield of each wafer. For samples from the large population, s rather than σ is used to denote standard deviation.

[4] One measure of dispersion is standard deviation. Relative dispersion is given by $s/\overline{M}$.

From Eq. 11.2, the expected dispersion s_N of the multiprobe average of lots of 9 wafers each would be

$$s_N = \frac{6}{\sqrt{9}} = 2 \text{ points}$$

The MPY of 74% is 4 points, or $2s$ away from the main lot yield of 70%. If these 9 slices were of the same population as the other 41, then there would have been only a 2.2% chance that their average yield would have been $2s$ away from the main average on the high side since 2.2% of the distribution is outside $2s$ on the high side. One can conclude that there is a 98% chance that these 9 wafers are not of the same population—that is, that the process change actually helped the yield. However, there is no implication that there is a 98% chance that the average yield of the new process will be *4 points* better than the old. The only implication is that there is a good chance that the new process is somewhat better than the old one.[5] □

Occasionally, a test lot will be completely split into small groups. There might, for example, be 5 lots of 9 wafers each. Again, the problem is one of determining whether or not any observed differences in multiprobe yield are significant. In this case, there is no major population to compare against, but it can first be postulated that all lots are random samples from the same population. If this is true, the standard deviation of the population from whence they came (pooled estimate) can be approximated by (3)

$$s_{\text{ap}}^2 = \frac{\Sigma M_{i1}^2 - [(\Sigma M_{i1})^2/N_1] + \Sigma M_{i2}^2 - [(\Sigma M_{i2})^2/N_2] + \ldots}{(N_1 - 1) + (N_2 - 1) + \ldots} \qquad 11.5$$

where M_{ij} is the multiprobe yield of the ith wafer of the jth split. The standard deviation s_{m1-m2} (dispersion) of the means of the multiprobe yields M_1 and M_2 of two samples picked from a sample with standard deviation of s is given by

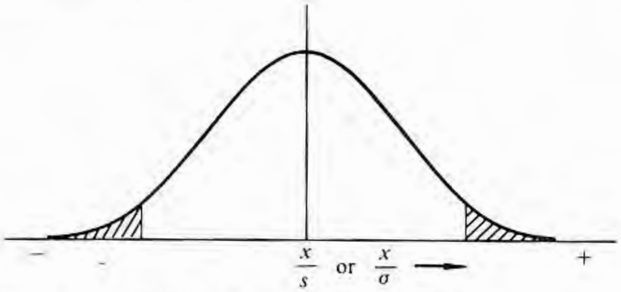
$$s_{M1-M2} = s\left(\frac{1}{N_1} + \frac{1}{N_2}\right)^{1/2} \qquad 11.6$$

where N_1 and N_2 are the number of wafers in each lot. In this case, s is not available, but an approximate value s_{ap} can be calculated from Eq. 11.5 and used instead. The observed difference between

[5] For procedures on estimating the probable range of the improvement and for determining the size of sample needed for a given degree of confidence, the reader should see, for example, A.R. Alvarez et al., *Solid State Technology*, pp. 127–133, July 1983, or a standard statistics text.

TABLE 11.2

Area in One Tail of the Distribution Curve



x/σ *or* t	*Normal Curve Area*	*20 Wafer Curve Area**	*10 Wafer Curve Area**
1	0.159	0.16	0.16
1.2	0.115	0.12	0.14
1.4	0.081	0.086	0.10
1.6	0.055	0.062	0.074
1.8	0.036	0.044	0.054
2.0	0.023	0.030	0.040
2.2	0.014	0.021	0.029
2.4	0.0082	0.014	0.023
2.6	0.0047	0.0090	0.016
2.8	0.0026	0.0058	0.012
3.0	0.0014	0.0038	0.0082

*Numbers interpolated from data in reference 3.

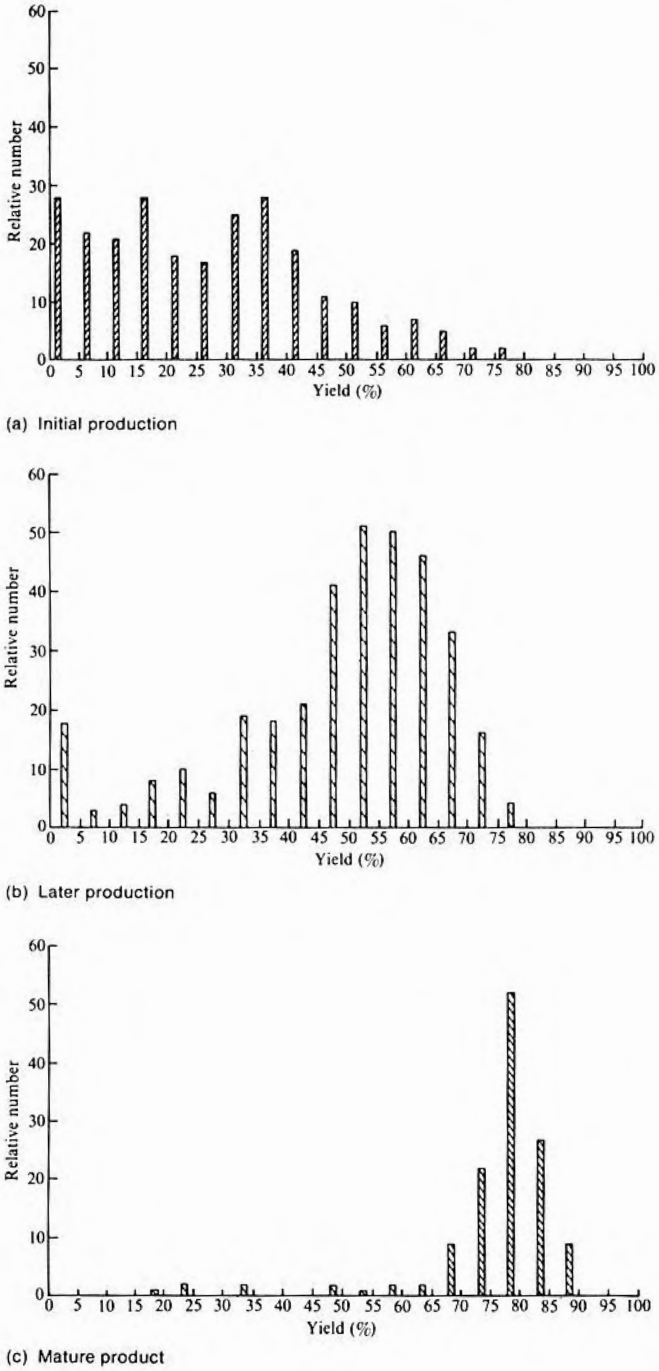
the two yields can then be compared to $s(M_1 - M_2)$ and a judgment as to its significance made. However, the normal probability curve data given in Table 11.2 do not, in principle, apply in this case. Instead, a "t" table, where t is the ratio of the yield difference to the standard error, is used with the approximate s values. The t table probabilities change with the number of wafers in the samples (N_1 and N_2 in Eq. 11.6). For demonstration purposes and because 5 and 10 slice splits are often used, Table 11.2 gives t only for $N_1 + N_2 = 20$ and $N_1 + N_2 = 10$. From a comparison of the t values with the x/s values in the table, it can be seen that, for a given difference in yield, the t table would predict a higher chance that the difference was insignificant and that the smaller the total number of wafers, the larger the yield difference required for significance. It can also be seen that if confidences in the 95% range are acceptable, the normal distribution curve can still be used.

11.5 DEFECT DENSITY

Two general kinds of multiprobe yield losses are observed in IC chips. One, which tends to extend over large areas of a wafer, occurs when one or more process parameters such as sheet resistance or diffusion depth drift out of range. The other occurs when localized defects cause either a single or a small cluster of failures. Examples are dirty contact windows and resist pinhole-induced diffusion defects. The defect density D is a measure of the density of localized defects and has proven quite useful in multiprobe yield prediction. Once a product is in production, it is usually possible to neglect the effect of broad-area process drift-induced defects because putting a process in production means matching the process variability to the required device parameter spread. Fig. 11.3 illus-

FIGURE 11.3

Multiprobe yield distribution for a large number of wafers.



(a) Initial production

(b) Later production

(c) Mature product

trates this point. The histogram in Fig. 11.3a shows the distribution of multiprobe yield over several lots when a new product with a completely new process flow is first introduced. The large congregation of yields near zero is caused by the actual process and the process required for high device yield not properly overlapping. The histogram in Fig. 11.3c shows the yield distribution after the product matures. The histogram in Fig. 11.3b will be discussed later.

11.5.1 Theory

Defect density theory is based on the existence of a number of "killer defects" very small in area and scattered over the wafer. The assumed distribution of these defects, $f(D)$, over the area of material processed and the mathematics used to convert this distribution to yield determine the way in which yield is predicted to vary with defect density and IC area. Some investigators restrict the use of defect-density-related equations to defects involving flaws in the IC physical structure and exclude flaws causing parametric failure. That is, yield equations are used only with functional failures and not with parametric failures. Experimentally, the equations seem to be equally applicable whether or not parametric failures are excluded. Mathematically, there appears to be little difference between a defect caused by a physical flaw in the IC structure and a defect defined by the intersection of the n-dimensional actual circuit parameter space with the envelope defined by the IC electrical specifications.

If each defect is considered to be identifiable and randomly distributed over a large area, the probability of finding a small area A with no defect (a good chip) is given by

$$P = e^{-AD} \tag{11.7}$$

where D is the density of the defects. This expression is generally referred to either as a Poisson distribution or as a Boltzmann distribution. If the defects are considered to be indistinguishable, the probability of finding a good chip in a small area A is given by

$$P = \frac{1}{1 + AD} \tag{11.8}$$

and is referred to as a Bose–Einstein distribution. A more general form of the Boltzmann distribution gives

$$P = \int e^{-D'A} f(D')dD' \tag{11.9}$$

where a single-valued defect density D is replaced with a defect density function $f(D')$, which has an average value D.

Early work assumed that the defect density was everywhere equal to D—that is, that $f(D') = 0$, except at $D' = D$ where it equaled 1. Then, the probability P of finding a chip area overlapping no defect (the yield Y of good chips) is given by (4, 5)

$$P = Y = e^{-DA} \qquad 11.10$$

or

$$Y(\%) = 100 \times e^{-DA}$$

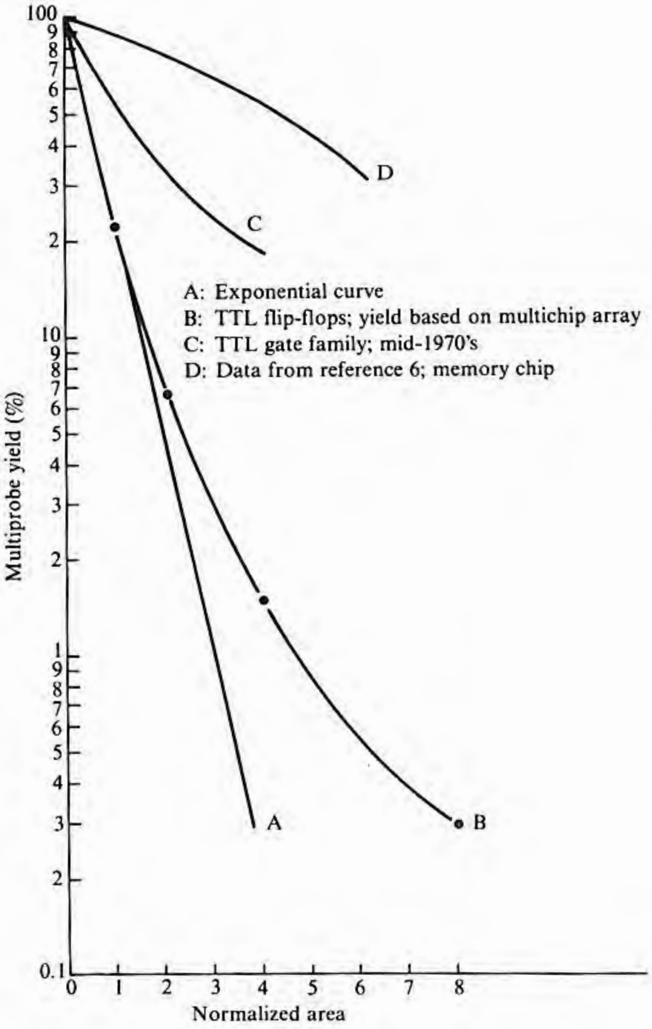
Eq. 11.10 predicts that if a chip of area A has a yield Y_0, then a chip of size NA should have a yield Y of $(Y_0)^N$. This approach to yield projection was first proposed when the IC was just making its appearance. The yields of all chips were low at that time, and the projected yields of larger, more complex circuits looked hopeless. However, it soon became clear that the exponential gave a pessimistic view. Curves of yield versus area are illustrated in Fig. 11.4. Curve A is an Eq. 11.10 projection, based on an actual IC yield of 23% for a normalized chip area of 1. The data for curve B were obtained by determining the yield of a single circuit chip and then, on the same set of wafers, redetermining the yield of arrays of 2, 4, and 8 good adjacent chips. It can be argued that a multichip array is not a proper way to perform such an experiment since the yield of different parts of the same circuit can be quite different.[6] For example, the yield of the portion of a chip used for bonding pads should be very good and, for a small chip, is a much larger percentage of the total chip area than for a large chip. An array of adjacent chips would maintain the same high percentage of high-yielding bonding pad area and thus have artificially high yields at large areas. In most cases, however, this does not happen. As an example, curve C of Fig. 11.4 shows yields for a series of ICs that comprise a TTL gate family and that have the same yield–area trend as curve B. All of the ICs in the series have the same number of bonding pads but an increasing circuit complexity and area. The data were taken somewhat later in time than the data of curve B and thus explain why the yields are higher. For some classes of devices, a downward curvature has been observed (6), as shown in curve D. Thus, the argument is sometimes valid and must be considered.

If the defect density were uniform, the yield of all slices will be very closely grouped when the chip area is very small compared with the wafer area. That this fact is not observed can be seen from Fig. 11.3, and Murphy (7) in 1964 suggested that there could be a spread of defect densities. For example, the probability of finding a particular density might be highest for the value D and tail off on

[6] Actually, A is often defined not as the area of the chip, but as the portion that would be susceptible to defects. In some layouts, some substantial surface area has no active elements and is very insensitive to process-induced defects.

FIGURE 11.4

Multiprobe yield versus normalized area.



either side in the manner of a Gaussian curve. As approximations that made integration of Eq. 11.9 easier, a triangular distribution and a rectangular distribution, as shown in Fig. 11.5, were used in the calculations. The distribution smearing has the effect of changing the character of the curve from that of curve A in Fig. 11.4 to something more like that of curve B in Fig. 11.4. The triangular distribution, with the peak centered at the average defect density D and having a half width of D was the one most favored by Murphy and indeed is still in use. It results in

$$Y = \left(\frac{1 - e^{-DA}}{DA}\right)^2 \tag{11.11}$$

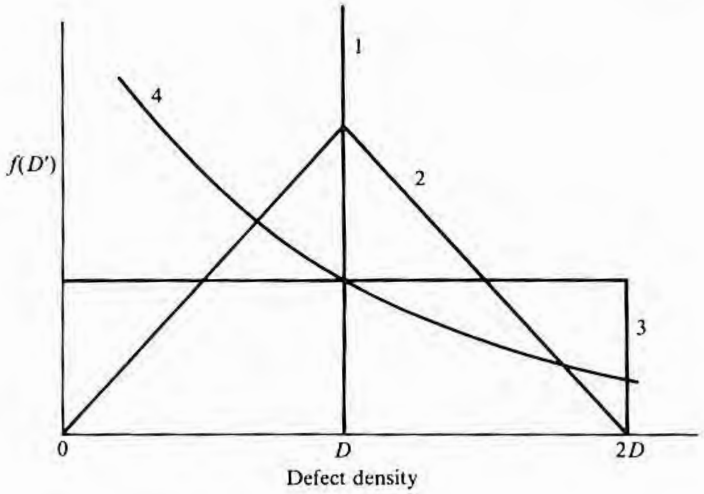


FIGURE 11.5

Defect distribution functions.

Use of the rectangular distribution of width twice the average defect density D gives

$$Y = \frac{1 - e^{-2DA}}{2DA} \qquad 11.12$$

A few years later, Seeds (8) suggested that there was a higher probability of a low defect density than a high one and used $(1/D)e^{-D'/D}$ for $f(D')$ (curve 4 of Fig. 11.5), which gives

$$Y = e^{-\sqrt{DA}} \qquad 11.13$$

and again results in a higher yield prediction at larger areas than does the exponential. Often, a noticeable reduction in yield occurs in going from the center to the edge of a wafer. Some approaches in handling this problem fold the radial variation in defect density into Eq. 11.10 in an attempt to provide more realistic yield projections (9, 10). In these cases, it is assumed that the average chip yield Y_w for a given wafer can be described by a series of yields associated with different areas of the wafer. That is,

$$Y_w = F_1 e^{-D1} + F_2 e^{-D2} + \cdots \qquad 11.14$$

where F_i is the fraction of the slice with an average defect density Di. A more generalized approach is to assume that $f(D')$ in Eq. 11.9 is a gamma distribution (11). A negative binomial distribution results, and the yield is given by

$$Y = (1 + \overline{D}AV)^{-1/V} \qquad 11.15$$

where V is the square of the coefficient of variance of the defect distribution and $\overline{D}$ is the defect density mean. The coefficient of variance is given by the standard deviation of the defect density divided by the mean $\overline{D}$.

Rather than choosing different density distribution functions, different statistics can also be used. Instead of the Boltzmann probability, Price (12), considering that the defects were indistinguishable, used Bose–Einstein statistics, which give for the yield

$$Y = \frac{1}{1 + AD} \qquad 11.16$$

This equation gives results similar to those of Eqs. 11.11 through 11.13. If several independent defect-producing mechanisms exist, as, for example, would be found in the different processing steps, then Eq. 11.16 can be generalized as follows:

$$Y = \frac{1}{(1 + AD_1)(1 + AD_2) \cdots} \qquad 11.17$$

The various equations just discussed are summarized in Table 11.3, and Fig. 11.6 shows a comparison of their projections. In most cases, it appears that while a particular equation may be most realistic for a given wafer-fab area it by no means has universal applicability. It should be noted that in all of these expressions, Y is a function of the DA product, not of D or A alone. Thus, the same differences among the curves showing up at the large areas of Fig.

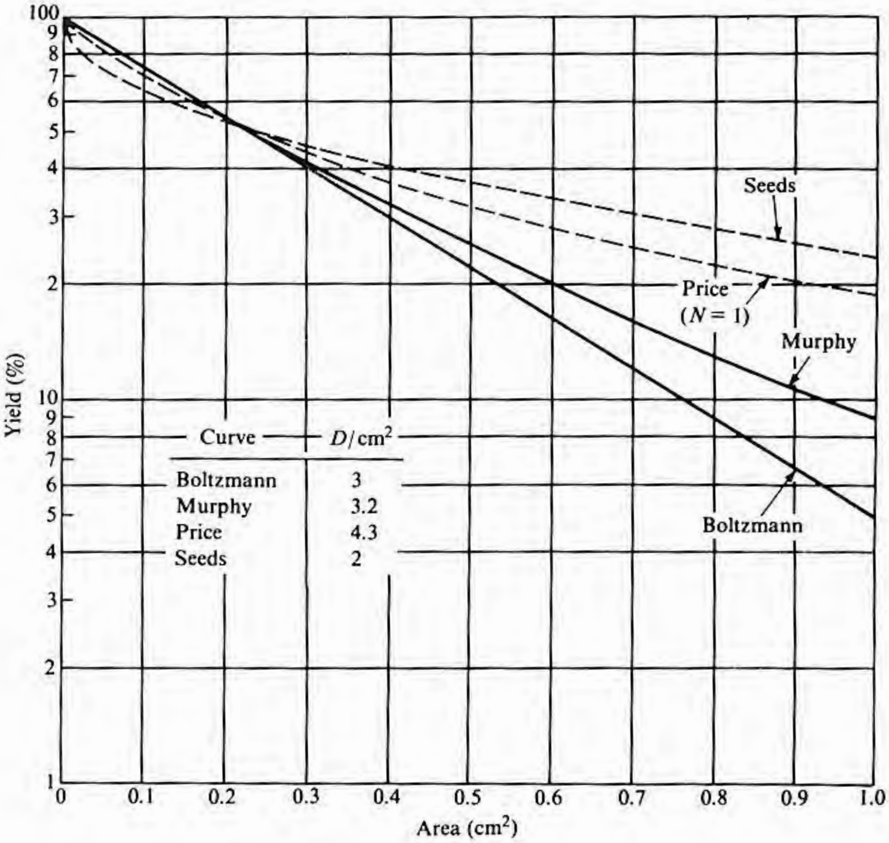
TABLE 11.3

Defect Density–Yield Prediction Equations

Eq. No.	*Equation*	*Approach*	*Reference*
T1a	$Y = e^{-DA}$	Exponential (Poisson or Boltzmann)	(4, 5)
T1b	$Y = e^{-D1A}e^{-D2A} \cdots e^{-DnA}$		
T1c	$Y = e^{-NDA}$		
T2	$Y = \left(\frac{1 - e^{-DA}}{DA}\right)^2$	Murphy	(7)
T3	$Y = e^{-\sqrt{DA}}$	Seeds	(8)
T4a	$Y = \frac{1}{1 + DA}$	Price (Bose–Einstein)	(12)
T4b	$Y = \frac{1}{(1 + DA)^N}$		
T4c	$Y = \frac{1}{(1 + D_1A)(1 + D_2A) \cdots (1 + D_NA)}$		
T5	$Y = (1 + \overline{D}AV)^{-1/V}$	Negative binomial	(11)

FIGURE 11.6

Comparison of yield projection theories, with curves matched at yield of 50% and area of 0.23 cm². (The *D*'s required for this match are tabulated on graph.)



11.6 could just as easily occur for smaller areas and larger defect densities. If a given area and yield are used to calculate *D* through the various equations of Table 11.3, the *D* values may be quite different, as can be seen in the examples in Fig. 11.6. Despite the disparity in *D* values, if the yields are above perhaps 30%, the differences among the predictions of the various expressions are probably not as great as the daily experimental range of values that are observed. One practical approach to the choice of equations is to use regression analysis to see which equation best fits the data of a given wafer-fab area and then use that equation for that area.

Since defects are not all generated at the same time, it is helpful to have an expression that relates the number introduced at each process step to the final yield (a thing difficult to do with a simple polynomial fit of data). Generally, most of the defects will be generated at only a few of the many manufacturing steps. These steps will closely follow the major ones listed in the process flowcharts of Chapter 2 but may not include all of them. Those considered major defect generators are usually referred to as "critical" steps or levels. It can be postulated that the overall yield Y is the product of the yields of the individual steps. That is,

$$Y = Y_1 Y_2 \cdots Y_n \qquad 11.18$$

Eq. T1a of Table 11.3 then becomes

$$Y = e^{-D1A} e^{-D2A} \cdots \qquad 11.19$$

If the number of defects introduced are considered to be the same for each level, then Eq. 11.19 reduces to

$$Y = e^{-NDA} \qquad 11.20$$

where N is the number of critical levels. The derivation of Eq. T4a included this provision (Eq. 11.17) and gave rise to Eqs. T4b and T4c. It should be noted, however, that Eqs. T4b and T4c cannot be used interchangeably by making the D of T4a equal to the sum of the D's of T4c. It is also a property of Eqs. T4b and T4c that with the same defect density for each level, as N increases, the predicted Y more and more closely matches the Y of Eq. T1a. By the time $N = 8$ (not an unreasonable number of steps), the shape of the curves of Eqs. T1 and T4 are nearly indistinguishable.

11.5.2 Defect Density Determination

Before any projections can be made using an equation from Table 11.3, a value for D must be determined. Ordinarily, multiprobe yields for wafers made by the same processes but having chips of different sizes will be collected and a D determined from a best-fit curve to the data. Occasionally, however, monitor wafers with special test patterns whose yields can be correlated with a defect density for that step are processed together with production wafer lots. This method allows an update on the projected yield of the lots to be made after each process step that used the special monitor wafers. This aspect will be discussed separately.

The "area" in the various equations is usually taken to be that of the whole circuit, including bonding pads. In some cases, however, it may be more useful to consider an effective area more closely related to the specific area of the chip in which a killing defect could occur. The yield normally is the total multiprobe yield, although functional yield is sometimes used. The complexity factor N can be defined as 1 if all products to be examined are made by the same process and the goal of the study does not require defects by level. Since most defects are associated with the lithography steps, the number of patterning steps is often used for N. Sometimes, implant steps will each be counted as 1/2 since no etching is involved.

The usual presumption is that when a large number of devices are run by the same process (all of the same N value), a defect density applicable to the whole collection can be determined by best fitting all of the data points (yield and area of each individual IC type) with the chosen equation. Of the equations of Table 11.3, the exponential is, in principle, the easiest to fit. Standard expressions are available for an exponential best fit in which the differences among the observed ln Y and the best-fit predicted ln Y are minimized. Unfortunately, this approach does not constrain the yield to be 100% at zero area. The problem can be eliminated by deriving the best-fit expression based on the fitted equation $Y = e^{-bx}$ rather than $Y = ae^{-bx}$, as is normally done. In this case, the best-fit D is given by (13)[7]

$$D = \frac{-\Sigma \ln Y_i}{\Sigma N_i A_i} \qquad 11.21$$

where Y is used as a decimal fraction, not as a percentage. It may be tempting to calculate a D for each pair of points and average them. That is,

$$D_{\text{av}} = \frac{1}{N} \Sigma \frac{\ln Y_i}{N_i A_i} \qquad 11.22$$

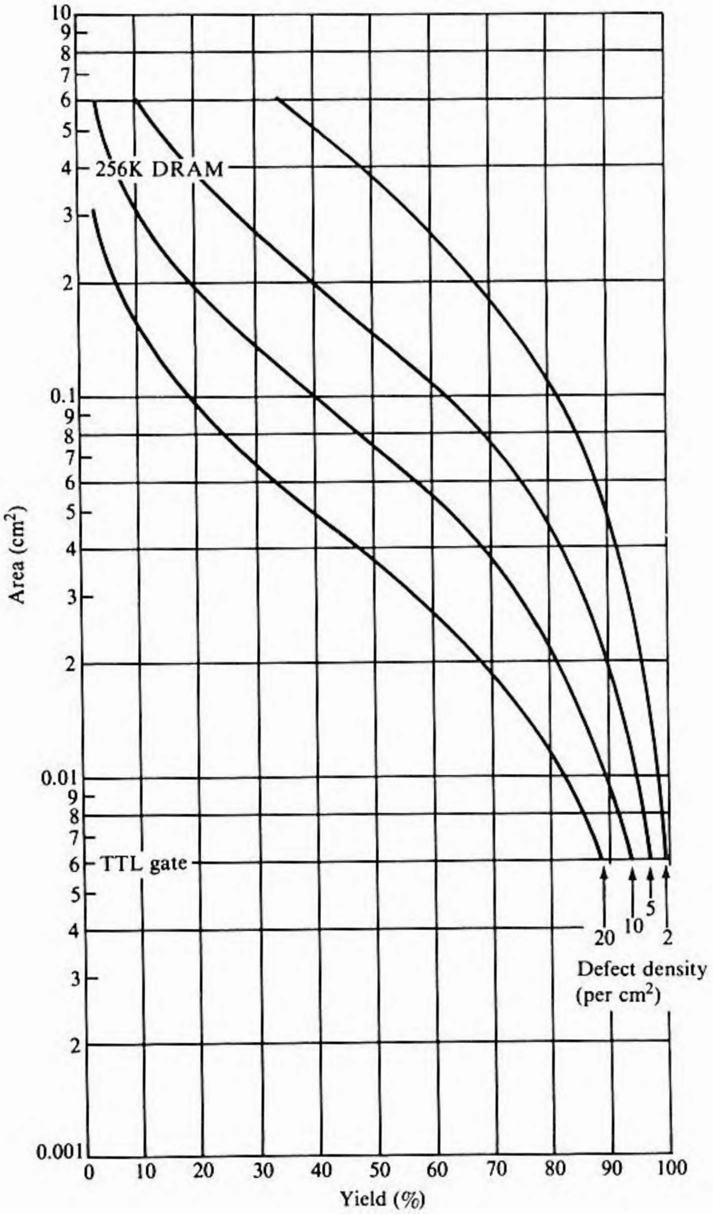
Such a value will not match the best-fit D of Eq. 11.21, although it may be reasonably close.

Expression T2 of Table 11.3 is more difficult to use. An easy way to estimate a D for a series of data points is to plot the points onto paper with a family of curves already drawn on it as shown in Fig. 11.7. After connecting the points with a smooth curve, the D value can be estimated from the position of this curve relative to the

[7] What constitutes a "best fit" depends in part on the use to which it will be put. In the case just discussed, one could have, for example, decided on a procedure that would have minimized the sum of the deviations of the yields rather than deviations of the log of the yields.

FIGURE 11.7

Area–yield plot for Murphy's equation.



location of the background curves. For calculating D from an individual Y–A value, a power series expansion can be used (14):

$$D = \frac{2Z + 1.33Z^2 + 1.11Z^3 + 1.007Z^4 + 0.953Z^5}{A} \qquad 11.23$$

where $Z = 1 - \sqrt{Y}$, and Y is used as a decimal fraction, not as a percentage.

In the case of Eq. T3, $Y = e^{-\sqrt{DA}}$, D for a single pair of Y–A values can be obtained from

$$D = \frac{(\ln Y)^2}{A} \qquad 11.24$$

For the D corresponding to a best-fit curve, the data points are plotted on semilog paper with yield along the log axis, and the square root of area (linear dimension of a square chip) is plotted on the other. The slope of the line will be $-\sqrt{D}$.

A best-fit D for expression T4b, $Y = 1/(1 + DA)^N$, can be obtained by rearranging as follows:

$$Y^{-1/N} = Z = 1 + DA \qquad 11.25$$

With this form, standard linear regression can be used for curve fitting. There is, however, one potential pitfall in that when Y is very small, Z becomes inordinately large and may weight the best fit in an undesirable manner. As in the case of Eq. T1c, yields associated with different N's but the same D value can all be combined into one curve in order to improve statistics.

11.6 CHIP YIELD PREDICTIONS BASED ON DEFECT DENSITY

Yield prediction is a necessary part of both IC design and the wafer fabrication operation. It can be used to estimate the yield of new devices and help ensure that the design will provide a part that will be economically successful. It can also be used to evaluate the yield of mature chips relative to others being run in the same wafer-fab shop.

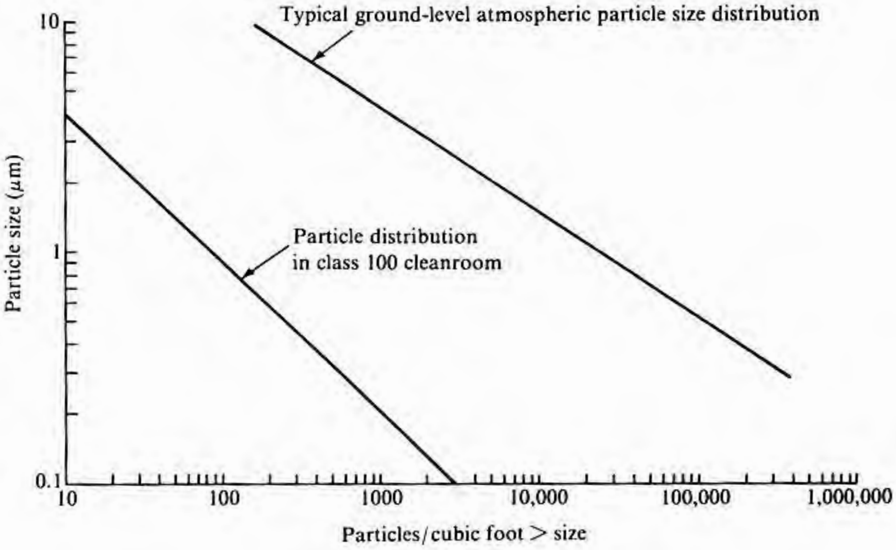
11.6.1 New Devices

When a new part in which the only changes are chip size and circuit function is put into production, the use of historical data, plotted in the format of Fig. 11.6, can be used directly. A defect density is calculated, and from it and the new chip area, an expected yield can be determined. If the actual yield falls below the curve, it is almost always due to a layout flaw, such as a lead spacing inadvertently made too small. When a new "shrunk" device is introduced, it must be remembered that the yield is driven by two opposing forces. First, since it is a smaller part, there is less likelihood of a defect on it. Second, however, the tighter geometry makes the photolithography steps more susceptible to small-area defects.[8] Attempts have been made to fold the defect size distribution into the yield equation (15), but data on the distribution to use are difficult to surmise.

One approach to the problem is to assume that since most lithography defects are caused by particulates collecting on the wafer or mask surface, the defect size will be proportional to the particle size. Particles may be carried in the ambient air, held in suspension in the processing liquids and gases, and/or generated by moving parts of the slice-handling equipment. The airborne particulate density increases with decreasing particle size as shown in Fig. 11.8 as approximately $1/d^2$ where d is the diameter. The density of liquid-born particles follows the same general trend. A reasonable device sensitivity approximation is that a particle of diameter 1/2 the size of a feature (a line spacing or contact window opening) on the wafer surface can cause a killing photolithographic defect. A calculation of the defect density for a given level then gives a measure of defects greater than 1/2 the feature size ζ. Using the particulate size distribution of Fig. 11.8, the defect density can be calculated for some other feature size. Thus, all processing steps (levels) that have losses

FIGURE 11.8

Particulate density above a given size versus size. (*Source:* Adapted from data in Philip W. Morrison, ed., *Environmental Control in Electronic Manufacturing*, Van Nostrand Reinhold Co., New York, 1973.)



[8] The defect density formulation discussed earlier assumed small-area "point" defects.

determined primarily by particulates, or photolithographic steps,[9] can be expected to show defect densities that will increase approximately[10] as $1/\zeta^2$. That is, for estimating purposes,

$$D(\zeta_2) \cong D(\zeta_1)\left(\frac{\zeta_1}{\zeta_2}\right)^2 \qquad 11.26$$

EXAMPLE □ Assume, in a 9 level process, that 6 levels are particle limited and that only one part type is run in the wafer-fab area. The chip has an area of 0.12 cm² and is running a 75% multiprobe yield. After a 15% linear shrink, what is the estimated new yield?

Since there is little extrapolation in area, choose the equation from Table 11.3 that is easiest to use. Further, assume that all levels have the same defect density. By using Eq. T1c, the initial D for *one level* is

$$D = \frac{\ln Y}{9 \times 0.12} = 0.27/\text{cm}^2$$

The new D_ζ for the shrunk lithographic levels is

$$D_\zeta = D\left(\frac{1}{0.85}\right)^2 = 0.27 \times 1.38 = 0.37$$

The new area is 0.087 cm², and

$$Y = (e^{-6AD_\zeta})(e^{-3AD}) = e^{-0.283} = 76\%$$

The projection in this case is that little change should occur in the multiprobe yield, but because of the smaller size, 38% more chips would be available. □

11.6.2 New IC Designs with New Processes

Projecting the yields of new designs with new processes is much more difficult than the case just discussed. There, the parametric yield was assumed to be very high and was neglected. However, in the case of new designs, it may well be the limiting factor. In any event,

$$Y = Y_P Y_D \qquad 11.27$$

[9] Assuming that the lithographic process is capable of printing the smaller geometries. If the process is already marginal or if the shrink is more than a few percent, yields may drop abruptly, not because of increased particles, but because of poor printing resolution.

[10] Allusion in reference 16 to unpublished work suggests that $1/\zeta^3$ might be more appropriate.

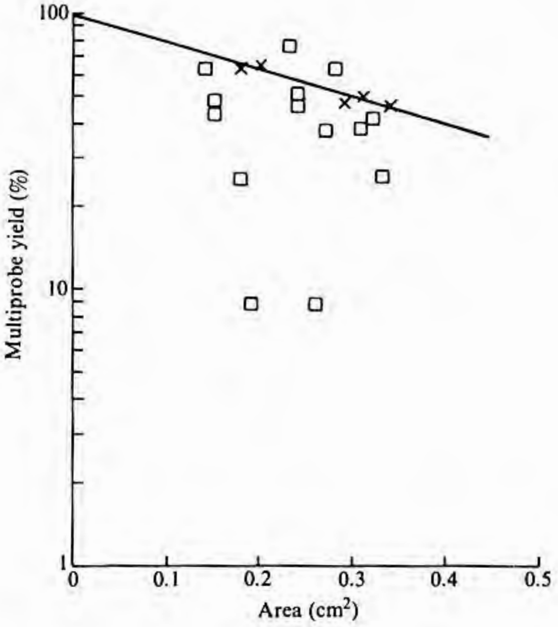
where Y_P is the parametric yield and Y_D is that due to defects. Devices with known high Y_P values can be run, and the wafer-fab defect density can be determined as before. Next, the process-induced variation of critical parameters important to the functioning of the new device is characterized in terms of a median value and standard deviation. For a MOS digital device, the channel length and width, the gate oxide thickness, and the flatband voltage are usually adequate. Then, the region in the space defined by those critical parameters that allow the IC to function properly is mapped out. The parametric yield can be estimated by integration of the probability distribution functions over this space (17). Such calculations are generally done in conjunction with SPICE or some similar circuit analysis program and will not be considered further in this book. The projected multiprobe yield (Eq. 11.27) is the product of this parametric yield and the previously determined defect-limited yield Y_D.

11.6.3 Devices in Production

Once an integrated circuit is in production, yield prediction is useful to see whether that circuit is running as well as it should. By examining a yield–area plot such as that of Fig. 11.9, low-yielding device types can be easily pinpointed and their expected yields projected. This particular figure shows the yields of 14 devices (□'s) that were just transferred into a wafer-fab area, along with 5 devices

FIGURE 11.9

Multiprobe yield versus area.



(×'s) that were already running and whose yields were considered representative of the wafer-fab area. The 5 devices were used to draw the line used for projecting acceptable yields of the other 14 chips. Sometimes, it is necessary to switch production from one wafer-fab area to another, and it is desirable to predict the yield in the new area. Then, the determination of a defect density representative of each facility will allow the yield of transferred devices to be satisfactorily estimated.

11.6.4 Work in Progress

Since the cycle time for wafers in a fabrication area will generally be in the 3–4 week range, a means of projecting the multiprobe yield of lots in progress is sometimes helpful. Such a projection can afford a more accurate means of predicting output volume and can also give early warning of process problems. The projections can be made either on the basis of the yield of monitor wafers (pilot wafers) pulled at various stages of the process or from the results of regular quality control (QC) inspections for visual defects. The general procedure is to assume historical yields as lots begin and then, as the results come in for the intermediate yields, upgrade the projection. That is, when the lot begins,

$$Y_{\text{est}} = Y_{h1} \times Y_{h2} \times Y_{h3} \cdots \qquad 11.28$$

and, as time progresses,

$$Y_{\text{est}} = Y_{a1} \times Y_{h2} \times Y_{h3} \cdots$$

$$Y_{\text{est}} = Y_{a1} \times Y_{a2} \times Y_{h3} \cdots$$

where Y_{hi} is the historical yield of step i and Y_{ai} is the actual yield of step i.

11.6.5 When Yield Is Low

When the average multiprobe yield is low and the distribution by slice looks like that shown in Fig. 11.3a, yield projections based on defect density calculations are not very helpful since defects in the normal sense of the word are probably not limiting yield. The reason for the yield loss is more likely due either to the fact that the process is not matched to the device design or else to the fact that the process is in very poor control. However, the yields to be expected when the process is controlled are reasonably well predicted by the maximum yield observed, which is illustrated in the sequence of parts a, b, and c in Fig. 11.3. In part a, while the peak is at 35%–40%, there is some yield at 75%–80%. In part b, the peak has moved to 50%–55%, but there is still some yield to 75%–80%. In part c, the low-yielding slices have all disappeared, and the peak yield is now at 75%–80%.

11.7 DETERMINING CAUSES OF CHIP YIELD LOSS

Before one starts to search for the underlying physical reasons for the "mathematical" defects discussed up to this point, the yield Y of Eq. 11.27 needs to be defined in terms of the types of defects causing the yield loss. It can, for example, be written as

$$Y = Y_D Y_P = Y_{DV} Y_{DNV} Y_{PV} Y_{PP} \qquad 11.29$$

where $1 - Y_{DV}$ is the catastrophic yield loss due to visual defects, $1 - Y_{DNV}$ is the catastrophic yield loss due to nonvisual defects, $1 - Y_{PV}$ is the parametric failure due to visual defects, and $1 - Y_{PNV}$ is the parametric failure due to nonvisual defects.

The significance of the visual defects (the $Y_{DV} - Y_{PV}$ combination) can be judged from Table 11.4, which shows the percentage of total multiprobe failures that in three separate evaluations were ascribed to visual defects. Thus, it is apparent that with the current state-of-the-art processing, visual defects are a major indicator of

TABLE 11.4

Incidence of Visual Defects Deemed to Cause Failures

Evaluation	*Percentage of Total Failures*
Single-Level Metal Gate CMOS	
Aluminum shorts	40
Aluminum opens*	20
Scratches	10
Irregular oxide pattern	5
Alignment	5
Total	80
Single-Level Bipolar Logic	
Metal shorts	15
Metal opens*	15
Irregular oxide pattern	10
Oxide undercutting	25
Alignment	10
Total	75
Silicon Gate MOS Memory	
Metal shorts	30
Metal opens*	30
Polysilicon shorts	15
Oxide holes (not pinholes)	5
Total	80

*Metal opens due to lithography are easy to see. Those due to breaks in the metal where it goes over oxide steps may be difficult to see optically. SEM views may be required.

chip failures. However, the problem of relating a particular visual flaw to an electrical failure exists since usually more visual flaws occur than do electrical failures. It should be remembered that visual defects that do not cause multiprobe (electrical) failures may, in some product lines, still be cause for rejection during a subsequent visual inspection. Even though a visual defect, such as a lead partially etched away or poor step coverage, may not cause an immediate electrical failure, it can cause a reduction in the time to failure of an integrated circuit. Since visual defects cause such a large fraction of multiprobe failures, both optical and scanning electron microscope examinations must be an integral part of any study of yield losses.

Another approach to studying the reasons for chip failures is by electrical testing. Unlike visual examinations, where there is a problem of relating a particular visual defect to a specific electrical failure, this approach has the problem of relating the electrical failure to a specific processing error. Test masks with special patterns can be substituted for regular masks during some point in the process and used to produce electrical error signals easily identifiable when specific process problems arise. These test patterns may be placed on separate test wafers, incorporated into the main body of the chip, or, in some cases, put in the scribe line. Otherwise, it is often difficult to determine what defects are causing specific device failures. The design of these kinds of masks is dependent on the specific integrated circuit to be studied and generally must be prepared by a circuit design engineer. Performing straight electrical test diagnostics without benefit of test patterns is widely used and is a multiple-step process. First, the particular transistor, diode, resistor, capacitor, or lead whose failure is most likely to cause the observed integrated circuit failure must be determined. Specially written multiprobe programs are often used for this determination. These programs, in addition to the standard test program, are required since multiprobe tests are used primarily to ensure that the chip meets customer requirements rather than for diagnostic purposes. After the faulty element is isolated, curve tracer analysis is generally the most fruitful technique for determining why the component failed, and the next two sections of this chapter are devoted to that approach. Finally, based on the results of the curve tracer analysis, the processing errors can be determined.

11.8 JUNCTION TESTING

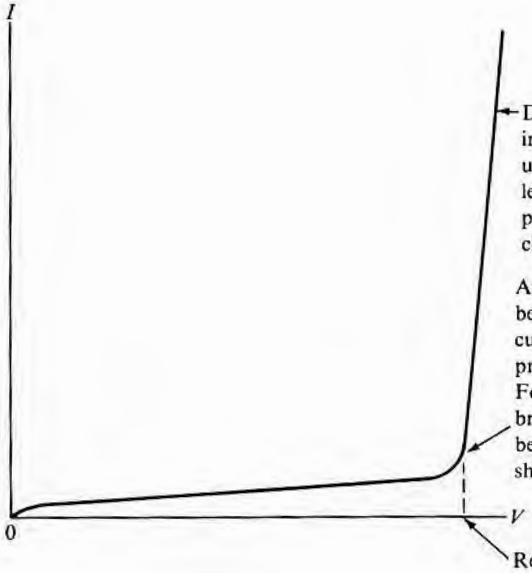
The pn and Schottky barrier junctions are basic building blocks of current semiconductor technology, and when they fail to perform properly, yield will suffer. The properties of a junction that are examined are the forward and reverse currents, the reverse breakdown voltage, and, occasionally, the series resistance. The current, forward or reverse, at a given voltage will almost never be less than projected and, if it is, will almost never be considered a reject. Higher than expected reverse voltage breakdown almost never occurs, and lower breakdown is usually because of excess reverse leakage current. Thus, junction analysis consists primarily of determining the origin of excess current.

11.8.1 Reverse-Biased pn Junctions

Current of a reverse-biased pn junction can be broken into three regions with quite separate behaviors as shown in Fig. 11.10. Before breakdown, which is where most junctions are intended to operate, the current is relatively low. This current is ordinarily referred to as "reverse current" and is discussed in more detail in the next paragraph. Above breakdown, the current (avalanche current) is quite large and is generally limited by external series resistance. In the region from near breakdown into breakdown, the current increases very rapidly (avalanche multiplication) and is proportional to $1/[1 - (V/V_B)^m]$ where V_B is the breakdown voltage and m is ~4 for n-type silicon and ~2 for p-type silicon.

FIGURE 11.10

Junction breakdown characteristics.



Differential resistance in this region can be used to estimate collector concentration profile beyond space charge width at breakdown.

Avalanche or tunneling beginning. Avalanche current in this region is proportional to $1/[1 - (V/V_B)^m]$. For avalanche, the break may actually be much sharper than shown here.

Reverse breakdown voltage can be used to estimate collector concentration.

The reverse current of an ideal diode is diffusion current, given by

$$I_d = \left(\frac{n_i^2}{N_A}\right)qA\sqrt{\frac{D_n}{\tau_n}} \qquad \text{for n}^+\text{p-diodes} \qquad 11.30\text{a}$$

$$I_d = \left(\frac{n_i^2}{N_D}\right)qA\sqrt{\frac{D_p}{\tau_p}} \qquad \text{for p}^+\text{n-diodes} \qquad 11.30\text{b}$$

where q is the electronic charge, k is Boltzmann's constant, A is the area of the junction, n_i is the intrinsic carrier concentration, $D_{p,n}$ is the carrier diffusion constant, $\tau_{p,n}$ is the carrier lifetime outside the space charge region, and $N_{A,D}$ is the doping level of the lightly doped side of the diode. This current is independent of voltage. That is,

$$I_d :: V^0 \qquad 11.31$$

It is exceedingly small and seldom seen at room temperature, being overshadowed by the generation current, which is described next. The temperature dependence of the diffusion current is that of n_i^2 and is given by

$$I_d(T) = BT^3e^{-1.21/kT} \qquad \text{for Si} \qquad 11.32\text{a}$$

$$I_d(T) = CT^3e^{-1.52/kT} \qquad \text{for GaAs} \qquad 11.32\text{b}$$

where B and C are constants and T is the temperature in K.

When generation–recombination centers are present in the space charge region, there will be a generation current component, given by

$$I_g = \frac{qn_iwA}{\tau_g} \qquad 11.33$$

where w is the width of the space charge and τ_g is the carrier generation lifetime in it. The lifetime τ_g depends on the number of recombination centers in the space charge region and in some cases is deliberately reduced by the intentional addition of lifetime killers such as Au or Pt. When this situation occurs, the reverse current will go up commensurately. If I_g is excessive and there is no intentional lifetime reduction, then gettering may be needed (see Chapter 8). Since the lifetime is relatively insensitive to temperature, I_g increases as n_i and thus not nearly as fast as I_d. Even though generation current almost always dominates the reverse current of silicon diodes at room temperature, diffusion current will be dominant at the high end of the temperature operating range. The smaller the bandgap, the more likely diffusion current will be the most important term, and, in the case of germanium, generation current is seldom seen. A GaAs diode would need to be above perhaps 200°C before I_d becomes an appreciable part of the total current.

Eq. 11.33 does not directly indicate a dependence of I_g on the applied voltage V_a, but it does show that I_g increases as the volume wA of the space charge region increases. A remains constant, but w will generally increase with V_a. For the specific cases of abrupt and linearly graded junctions,

$$w = \left[\frac{2\varepsilon\varepsilon_0(V_a + V_b)}{qN_{A,D}}\right]^{1/2} \qquad \text{for abrupt junctions} \qquad 11.34$$

$$w = \left[\frac{3\varepsilon\varepsilon_0(V_a + V_b)}{qa}\right]^{1/3} \qquad \text{for graded junctions} \qquad 11.35$$

where ε is the relative dielectric constant, ε_0 is the permittivity of free space, V_b is the built-in voltage, and a is the impurity gradient on the lightly doped side. Since V_b is less than a volt, and V_a is generally several volts,

$$I_g :: (V_a)^{1/2} \quad \text{or} \quad (V_a)^{1/3} \qquad 11.36$$

In the event of a resistive shunt R_s across the junction, the current component I_s will be given by

$$I_s = \left(\frac{1}{R_s}\right) V_a^1 \qquad 11.37$$

This problem is external to the junction and could, for example, be due to an extraneous diffusion or extra circuit components not detached from the junction being examined. A thorough microscopic examination of the circuit should locate the shunt.

If metallic precipitates or other defects that can cause tunneling are present in the space charge region, the current component I_t will be present. I_t will be a much stronger function of applied voltage than any of the others, increasing as V_a^n where n can be in the 5–8 range (18):

$$I_t :: V_a^n \qquad 11.38$$

For such precipitates to occur, a combination of a high concentration of elements such as Ni, Cu, or Fe must be present, along with crystallographic defects and a slow cooldown.

These various currents are depicted in Figs. 11.11 and 11.12 and are plotted on both linear scales and log–log scales. Linear scales

FIGURE 11.11

Curve tracer *I–V* plots of reverse pn junction current showing behavior of different current components (linear scales).

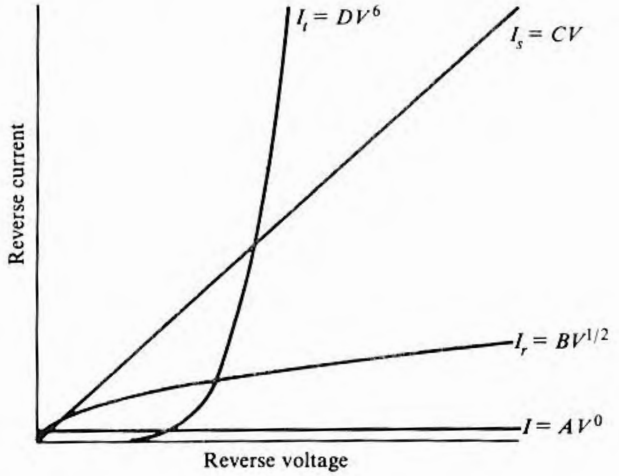


FIGURE 11.12

Functional relation between current and voltage for several modes of current generation in reverse-biased pn junction diodes.

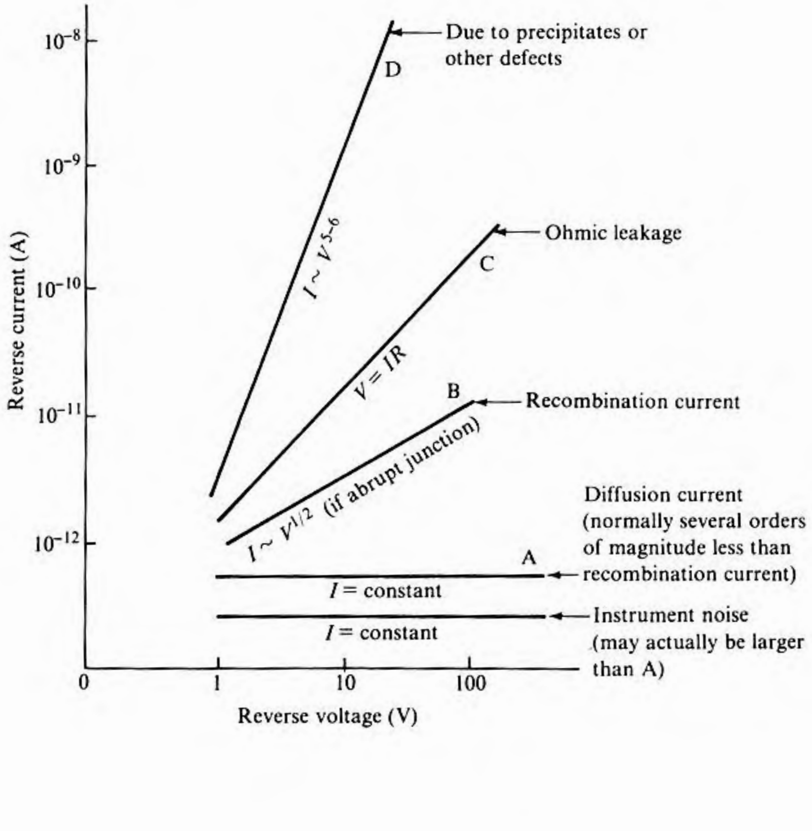
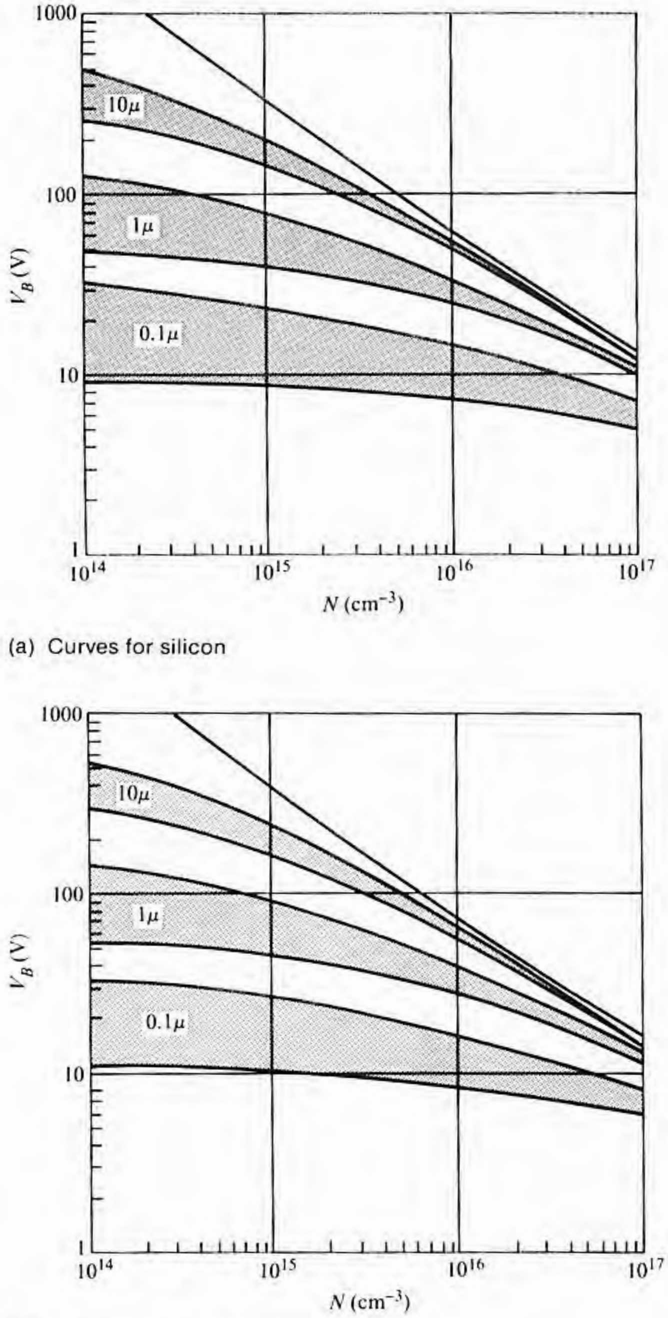


FIGURE 11.13

Calculated breakdown voltage versus background doping for an abrupt junction having various radii of curvature. (The upper line is for a flat junction. The upper bound of each shaded area is for a cylindrical junction with radius as indicated. The lower bound is for a spherical junction of the same radius.) (*Source:* Reprinted with permission from *Solid State Electronics 9*, S.M. Sze and G. Gibbons, Copyright 1966, Pergamon Press plc.)



(a) Curves for silicon

(b) Curves for gallium arsenide

are used on curve tracers and are appropriate for a quick overview of potential problems, but log–log scales are better for analytical interpretation. While the total current is the sum of all the components, if there is a problem of excess current, one will usually dominate. Then, since there is a different voltage dependency on each, the offending mechanism can generally be determined.

Voltage breakdown occurs when the voltage across the space charge region becomes large enough for carrier multiplication to occur (avalanche breakdown) or for tunneling to begin. The latter occurs only when both sides of the junction are heavily doped, which in ICs would only be because of a processing difficulty. If the junction is abrupt and one side is heavily doped, the breakdown voltage V_B depends on the impurity concentration of the lightly doped side as shown in each top curve of Fig. 11.13. If the space charge region moves out enough to touch a low-resistivity region (or an ohmic contact), breakdown will occur at that point, and V_B will depend on the width of the lightly doped side. If the junction is graded instead of abrupt, breakdown will depend on the gradient a and will be greater than for an abrupt junction. When a junction is physically curved, as, for example, at the periphery of a planar junction, the electric field increases as the radius of curvature decreases so that V_B depends on the curvature. This effect for varying degrees of curvature is also shown in Fig. 11.13. As junctions become shallower and their area smaller, curvature becomes more pronounced, and V_B for a given doping level decreases. It can also be seen from Fig. 11.13 that as the radius of curvature decreases, the V_B dependence on doping decreases.

11.8.2 Reverse Current Curve Tracer Diagnostics

Fig. 11.14 shows examples of the different kinds of curves that may occur when reverse-biased pn junctions (diodes) are examined with a curve tracer. Fig. 11.14a shows a good diode, assuming that the current gain is properly set. It also has the "textbook" shape. The other curves in Fig. 11.14 are all of good diodes. The loop in Fig. 11.14b is due to capacitance in the leads or junction and may be at least partially balanced out in most tracers. Stray pickup can cause the trace to be quite distorted. Fig. 11.14c shows the trace from drain to gate of a junction FET with the source floating. Pickup caused the loop, which occurs below 10 V. Above 10 V, the channel pinched off, and there was no longer any pickup. The trace in Fig. 11.14d is caused by charging of the emitter–base capacitance of a bipolar transistor (this curve was obtained by measuring between collector and emitter). These distorted traces of good junctions illustrate the point that care must be taken in interpreting curve tracer data.

FIGURE 11.14

Various I–V plots of good reverse-biased pn junctions as they appear on a curve tracer.

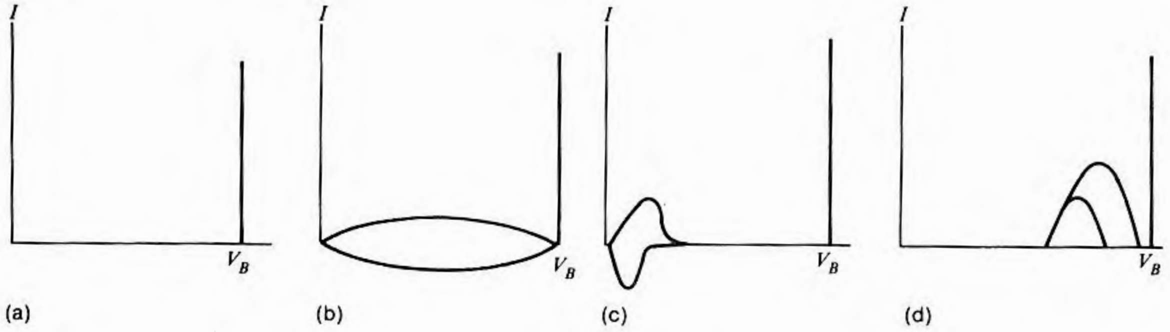


FIGURE 11.15

I–V curve tracer plots of reverse-biased pn junctions in parallel with various kinds of resistive paths.

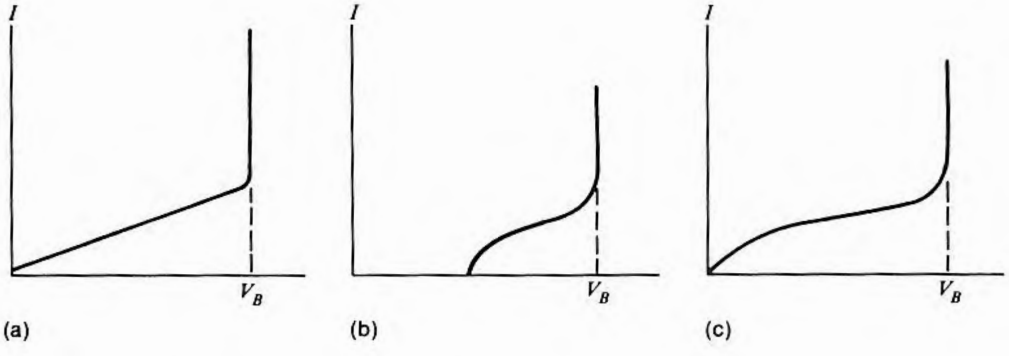


Fig. 11.15a has excessive leakage current that varies linearly with voltage and hence is due to a resistive shunt that does not pinch off.[11] Fig. 11.15b is characteristic of a low breakdown region connected to a higher breakdown region with a path that can pinch off.

[11] Pinchoff is the narrowing of the resistive path with increasing voltage and, as in a FET, occurs because of a space charge region moving into the path.

The most common example is a diffusion pipe connecting the collector and emitter of a bipolar transistor. This behavior is observed in looking at the collector–base junction. Fig. 11.15c has excessive current that appears to start at a zero voltage. Such behavior can be due either to generation current as shown in curve B of Fig. 11.12 or to a resistive path pinching off. Superficially, they appear to have much the same shape, but recombination current should continue to rise until breakdown with an approximate I–$\sim V^{0.5}$ dependency, whereas resistive pinchoff current will eventually saturate. Breakdown may occur before saturation, however. Larger-than-expected recombination current can occur because of excessive lifetime killers or because a surface inversion layer has extended the diode area.

Various inversion layer possibilities are shown in Fig. 11.16. If the inversion terminates as in Fig. 11.16a, excess current commensurate with the extra area will result. Depending on the value of the surface recombination velocity, the generation current in thin inversion layers can be substantial (19). Should the inversion layer extend

FIGURE 11.16

Extension of junction area through inversion layer formation.

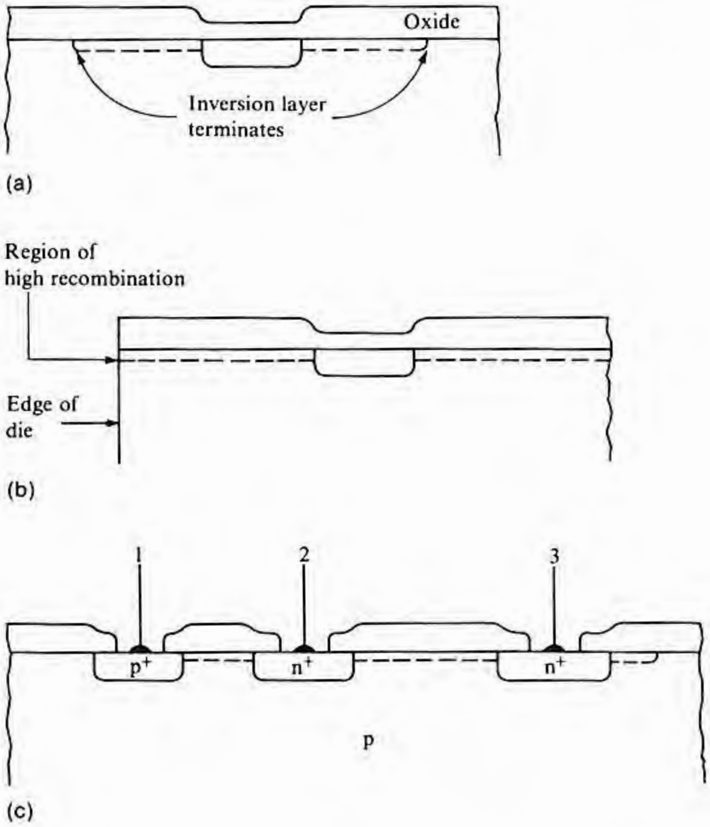


FIGURE 11.17

I–V curve tracer plots of leaky reverse-biased pn junctions.

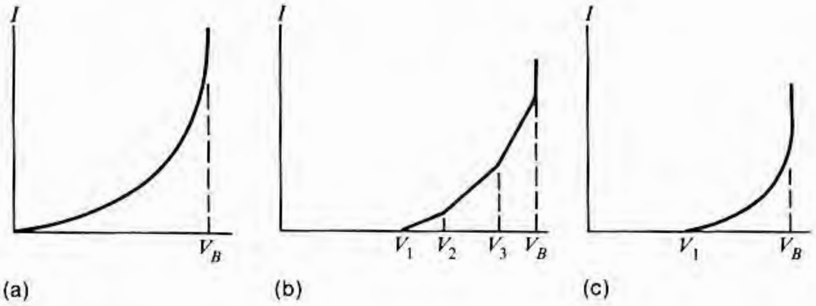
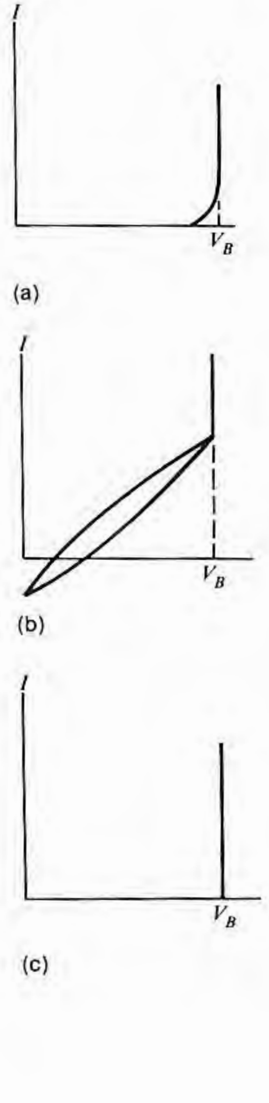


FIGURE 11.18

Miscellaneous I–V curve tracer plots of reverse-biased junctions.



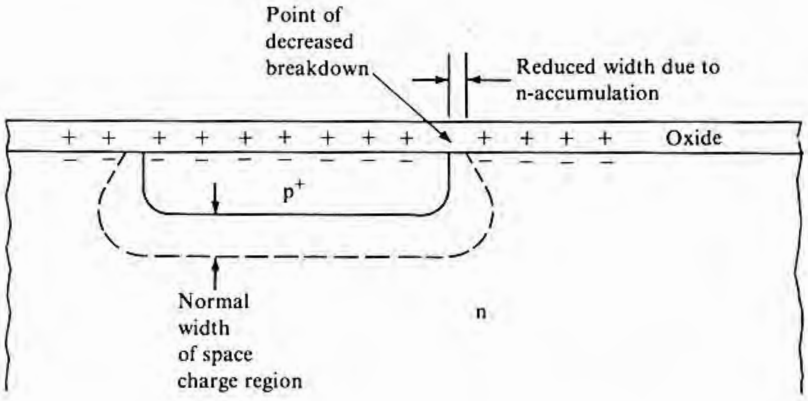
to a region of high recombination density such as the edge of a chip (Fig. 11.16b), orders of magnitude higher current can flow (20). If an inversion layer connects two n^+-regions as in Fig. 11.16c, a resistive path that will pinch off will be placed in parallel with the two back-to-back diodes.

Fig. 11.17a has excessive leakage current that begins near zero voltage, increases more rapidly than V, and hence is most likely due to precipitates (curve D of Fig 11.12). Because of the rapid current rise with voltage, such a curve could under some circumstances look like a forward diode ($I = Ae^{qV/kT}$). A log–log plot of I versus V over a few decades of current will allow separation. Fig. 11.17b is a case of premature breakdown. An initial low breakdown occurs in a small area of the junction at voltage V_1. Following it is a series of regions breaking down in succession at progressively higher voltages. When closely spaced, they give the illusion of the smooth curve of Fig. 11.17c. Possible causes are localized surface accumulation layers, localized high-concentration diffusion sources, or high fields due to high junction curvature at diffusion spikes. It is often possible to see light emitted from these small avalanching regions. The color will appear reddish if the light originates below the surface and white if it originates on the surface. If the voltage is held just above where the first breakdown shows on the curve tracer, one spot will usually be seen. Then, as the voltage is slowly raised, additional spots will light up in unison with the additional segments of current increases seen on the curve tracer. Since the spots are small and the light is dim, a microscope must be used either covered with a black cloth or placed in a darkened room. Another possibility for the curve of Fig. 11.17c is that the curve really looks like the one of Fig. 11.17a but only at point V_1 did the current exceed the noise level of the curve tracer and thus become noticeable.

Fig. 11.18a could be the trace of a good diode. If V_B is less than about 5 V, the trace could be that of a zener diode, which has a much softer breakdown (more gentle rounding of the knee of the curve) than avalanche breakdown. The trace might also be that of a Schottky diode. Schottky diode reverse current near breakdown appears excessive compared to that of a pn junction, and the break is not quite as sharp as expected. A log–log current voltage plot for Schottky diodes looks much different from that of a pn junction and is discussed in the next section. The trace in Fig. 11.18b is that of an otherwise good junction that has polarizable surface ions such as might be present in a wet package. While the trace in Fig. 11.18c looks like that of Fig. 11.18a, the breakdown, while sharp, is not at the correct voltage V_B. If the breakdown voltage is high, then almost certainly a material of higher-than-planned resistivity was used. If it is low, it could be because of a material with lower-than-planned resistivity; because the space charge reached through to a low-resistivity region and ceased widening before breakdown was reached; because excessive junction curvature produced high fields and premature breakdown; or because a surface accumulation layer[12] caused a narrowing of the space charge region at the surface as shown in Fig. 11.19. It is also possible that the final V_B is correct but that the low breakdown as observed on the curve tracer is because the junction is leaky and the current gain is set very high. In this case, a behavior like that shown in Fig. 11.20a will appear as in Fig. 11.20b.

FIGURE 11.19

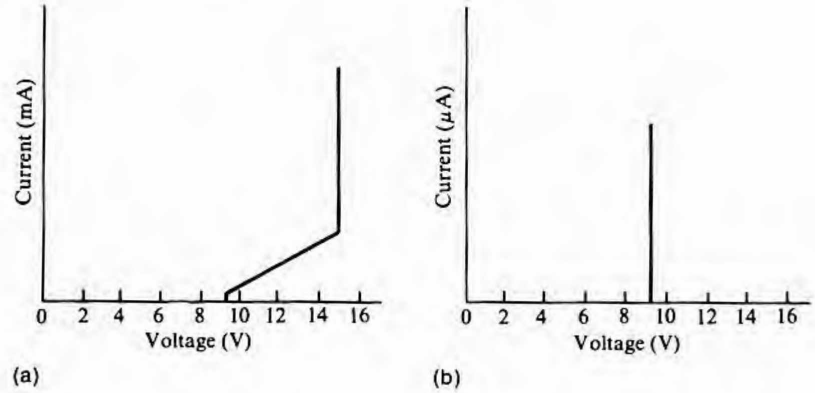
Method by which accumulation reduces breakdown voltage. (If p-type accumulation occurred rather than n, this effect would be observed in n^+p diodes.)


Point of decreased breakdown
Reduced width due to n-accumulation
+ + + + + + + + + + + + +
Oxide
p^+
n
Normal width of space charge region

[12] For a discussion of the possible causes of accumulation and inversion layers, see Chapter 3.

FIGURE 11.20

Effect of scale change on appearance of breakdown characteristics.



11.8.3 Reverse-Biased Schottky Junctions

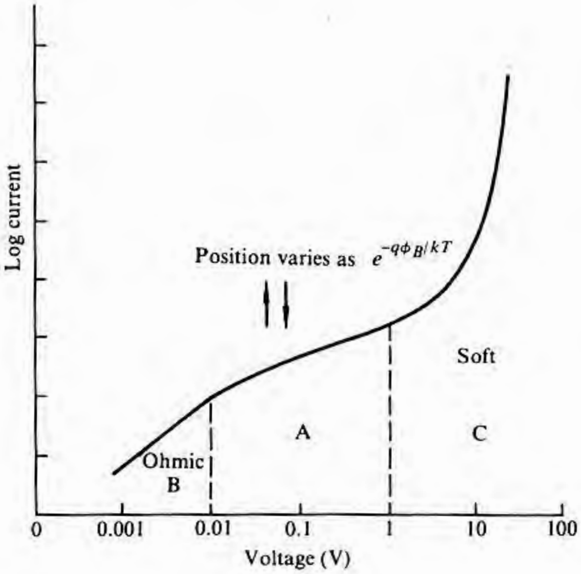
The reverse current of an ideal Schottky diode is given by

$$I_s = AT^2A^{**}e^{-q\phi_B/kT} \qquad 11.39$$

where A is the diode area, A^{**} is the effective Richardson constant, and ϕ_B is the potential barrier. Since ϕ_B is voltage dependent, unlike a pn junction with only diffusion current, a theoretical Schottky diode reverse current never saturates but increases steadily with voltage as shown in Fig. 11.21. In addition to this current, there can also be (as in pn junctions) recombination current and tunnel cur-

FIGURE 11.21

Schottky diode reverse behavior plotted on a log–log scale.



rents associated with the space charge region. Because of the high field associated with the very sharp curvature at the edge of the metal, additional potential barrier reduction occurs at the edge. Thus, Schottky diodes usually have either pn junction guard rings or field plates. Excess reverse current or low breakdowns can then be due to either pn junction or Schottky problems. The most likely cause is a lowering of the potential barrier due to a Schottky metal contamination. In this case, the current will be higher than expected, but the shape of the I–V curve should still be like that of Fig. 11.21. If a pn junction guard ring is suspected, the Schottky metal can be stripped away and the remaining current examined. To fully determine the various excess current sources, it may be necessary to use a special test structure with an isolated guard ring that can be independently biased. However, such a structure usually requires deviations from the standard process flow and is thus of limited usefulness.

11.8.4 Forward-Biased pn Junctions

Like reverse current, forward current I_f of a diode has several components whose relative magnitudes may change with design, processing deviations, and stress. For a nondegenerate pn junction, the components are the diffusion current I_{fd}, composed of current diffusing from n to p and from p to n; bulk recombination current I_{fr}, which arises because some of the injected minority carriers recombine in the space charge region; surface recombination current I_{fs}, from carriers that recombine at the surface; and a shunt component current I_{fsh}, which arises if a shunting resistor is present. Schottky diode forward current was discussed in Chapter 10. Note here that because Schottky diodes are primarily majority carrier devices, the only current component that they might have in common with pn junction diodes is I_{fsh}.

I_f is given by

$$I_f = I_{fd} + I_{fr} + I_{fs} + I_{fsh} \qquad 11.40$$

The forward diffusion current has the form

$$I_{fd} = I_D\,(e^{qV/kT} - 1) \qquad 11.41$$

where I_D is the sum of the two I_d values of Eq. 11.30. When V is greater than a few kT/q (0.026 V at room temperature), Eq. 11.41 becomes

$$I_{fd} = I_d e^{qV/kT} \qquad 11.42$$

The forward current due to recombination in the space charge region is given by

$$I_{fr} = I_r e^{qV/nkT} \qquad 11.43$$

where

$$I_r = \frac{qn_iwA}{\tau_g} \qquad 11.44$$

Eq. 11.44 is the same expression as that for generation current given in Eq. 11.33. n is an exponent that varies with the source of centers, but for space charge recombination it is usually about 2. If it originates in an inversion layer or channel,[13] it will be greater than 2 and normally less than 4 (21).

The surface recombination current has the form

$$I_{fs} = I_s e^{qV/2kT} \qquad 11.45$$

where I_s depends, not on the lifetime in the bulk, but on the surface recombination velocity s_0. That is,

$$I_s = \frac{n_i s_0}{2} \qquad 11.46$$

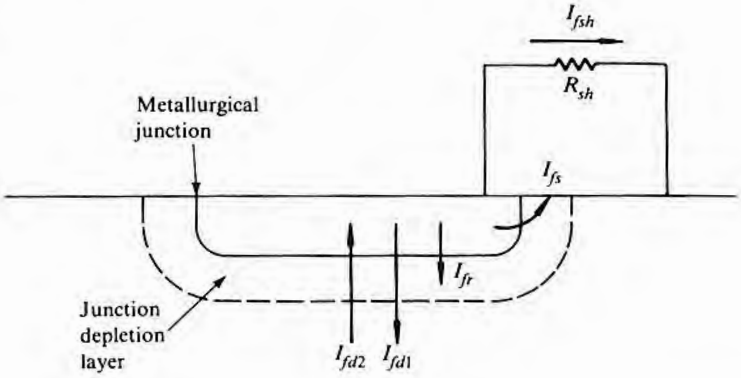
If a shunt resistance R_{sh} is present, then

$$I_{sh} = \frac{V}{R_{sh}} \qquad 11.47$$

These various currents are shown pictorially in Fig. 11.22. If a series resistance R_s is present, then the voltage across the junction is less than the applied voltage V_a by an amount I_fR_s. When the

FIGURE 11.22

Components of current in a forward-biased pn junction.



[13] In this context, channeling occurs when the surface potential on each side of the junction is the same. Channeling can also refer to an inversion layer connecting two otherwise isolated regions of the same type.

current is high enough for the injected carrier concentration to be comparable to the doping concentration, the form of the diffusion current of Eq. 11.41 changes to

$$I_{fd} = I_{d'} e^{qV/2kT} \qquad 11.48$$

and thus might be mistaken for recombination current. Note that $I_{d'}$ does not equal I_d of Eq. 11.30 and that V is the voltage across the junction and not V_a across the terminals. This is an important consideration since the effect only occurs at very high currents, and that is when a series resistance voltage drop is most noticeable. The onset of the $qV/2kT$ regime begins approximately when (22)

$$V = \left(\frac{2kT}{q}\right) \ln\left(\frac{2N}{n_i}\right) \qquad 11.49$$

where N is the doping concentration of the lightly doped side of the junction. Although reverse current I–V curves are most conveniently plotted on a log–log scale, forward characteristics are better displayed on log-I–linear-V coordinates. Fig. 11.23 shows a com-

FIGURE 11.23

Forward pn junction I–V curve showing various components. (Values are typical for silicon.)

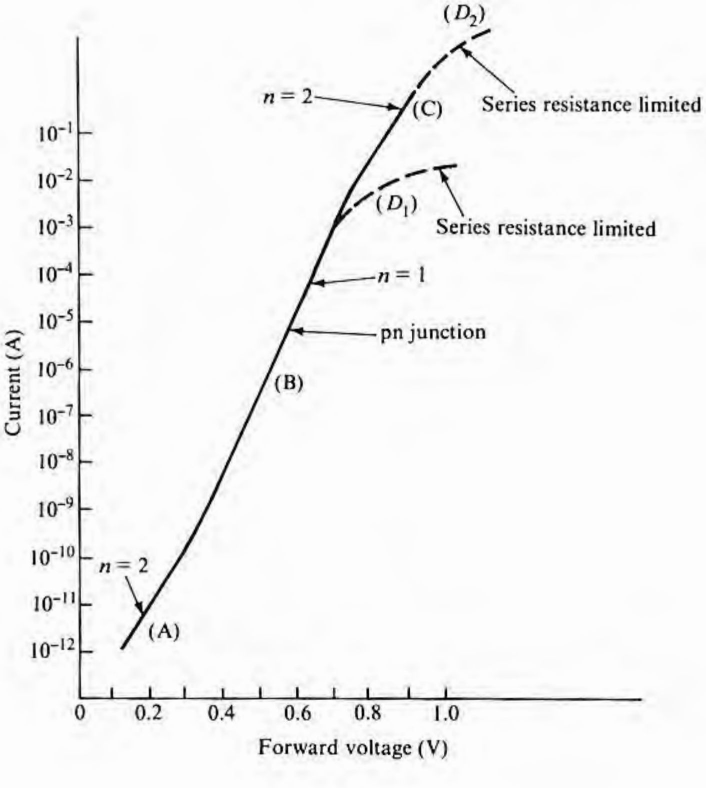
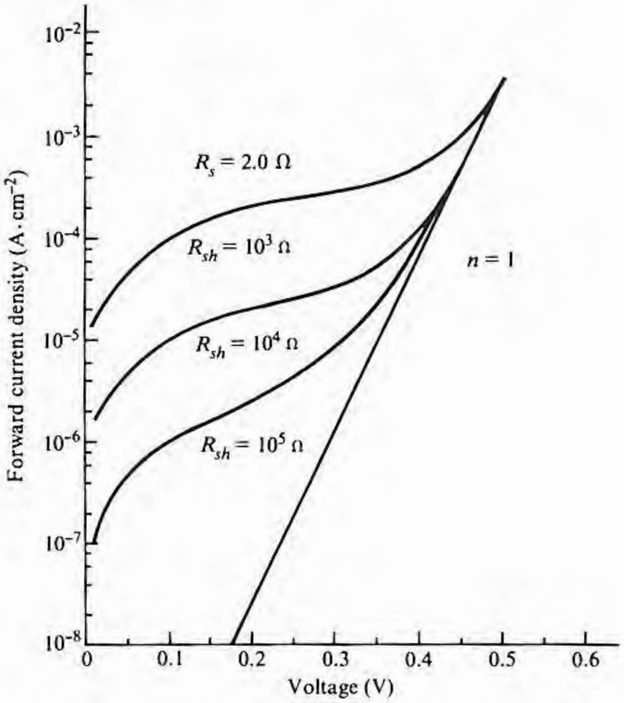


FIGURE 11.24

Calculated effect of shunt resistance on forward pn junction characteristics. (*Source:* Adapted from Richard J. Stirn, National Workshop on Low Cost Polycrystalline Silicon Solar Cells, sponsored by the National Science Foundation and the Energy Research and Development Administration, Southern Methodist University, Dallas, 1976.)



posite curve for a pn junction exhibiting most of the components just described and the regions in which they are ordinarily seen. At very low voltages, recombination current with a $2kT/q$ dependence usually dominates (region A). For higher voltages, diffusion current ($1kT/q$) will become larger as shown in region B. If series resistance is low enough, high injection level diffusion current ($2kT/q$) of region C will be seen before the resistance limit of D_2 is reached. Otherwise, the *I–V* curve will be series resistance limited sooner as shown in region D_1, and the current increase is directly proportional to the voltage increase. Shunt resistance will affect the *I–V* curve in the low current region, distort the curve as shown in Fig. 11.24, and may be mistaken for recombination current with a high *n* value.[14]

[14] *n* can be quickly determined by measuring the voltage differential ΔV in millivolts required to increase the current by a factor of 10. Under these conditions, $n = \Delta V/60$. Note also that if the area of the junction increases by a factor of 2 and the diode is operating in the $n = 1$ region, the voltage for a fixed current will decrease by 18 mV.

Not listed in Eq. 11.40 is tunnel current, which might be present if both sides of the junction were to be very heavily doped. It appears as an excess current superimposed on the normal current, peaks at about 100 mV for silicon, and will have virtually disappeared by the time V reaches 300 mV. It is also possible for adjacent junction interactions to cause enhanced forward current.

11.9 MOS TRANSISTOR TESTING

Fig. 11.25 shows cross sections, circuit diagrams, polarities, and terminology for MOS transistors. Both enhancement-mode and depletion-mode transistors are made, although the former are more common. The testing is somewhat different for the two types, but in either case, much of it can be done with a curve tracer. Three small-signal transistor parameters often measured are as follows:

1. Channel transconductance g_m

$$\left.\frac{\partial I_{DS}}{\partial V_{GS}}\right|_{V_{DS}=\text{constant}}$$

Usually, V_{DS} is made large enough for I_{DS} to be saturated, in which case $g_m = g_{m(\text{sat})}$.

2. Channel (drain) conductance g_D

$$\left.\frac{\partial I_{DS}}{\partial V_{DS}}\right|_{V_{GS}=\text{constant}}$$

3. Gain μ

$$\left.\frac{\partial V_{DS}}{\partial V_{GS}}\right|_{I_{DS}=\text{constant}}$$

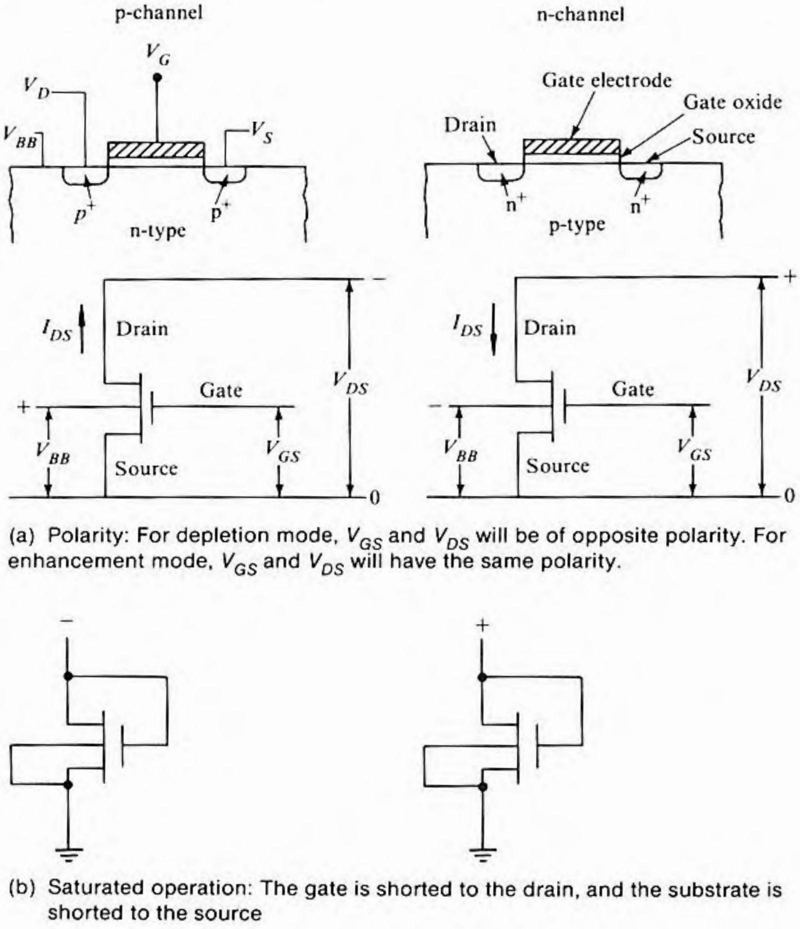
11.9.1 Threshold Voltage

The most commonly checked parameter in diagnosing problems with enhancement[15] transistors is the threshold voltage V_T. The threshold voltage is the gate voltage required to just form the conductive channel between source and drain. V_T can be determined with the transistor operating either in the unsaturated mode or in the

[15] For zero gate voltage, there is no channel and, except for reverse current leakage, no current flow from source to drain of an enhancement transistor. Depletion transistors have a channel and current flow at zero gate voltage. That current can be either increased or decreased with a change of polarity of the gate voltage.

FIGURE 11.25

MOS transistor structure, polarities, and terminology.



(a) Polarity: For depletion mode, V_{GS} and V_{DS} will be of opposite polarity. For enhancement mode, V_{GS} and V_{DS} will have the same polarity.

(b) Saturated operation: The gate is shorted to the drain, and the substrate is shorted to the source

Symbol	Meaning	Symbol	Meaning
V_D	Drain voltage	V_{GS}	Gate–source voltage
V_S	Source voltage	V_{BB}	Substrate back bias voltage (substrate to source)
V_G	Gate voltage		
V_{DS}	Drain–source voltage	I_{DS}	Drain–source current

(c) Terminology

saturated mode.[16] The saturation-mode method is somewhat simpler and is more often used. For unsaturation, when $V_{GS} - V_T >> V_{DS}$,

$$I_{DS} = k\left[V_{DS}(V_{GS} - V_T) - \frac{V_{DS}^2}{2}\right] \qquad 11.50$$

where k is the conduction factor. For a small fixed V_{DS}, an extrapolation to $I_{DS}/V_{DS} = 0$ of a plot of I_{DS}/V_{DS} versus V_{GS} will give V_T. For operation in the saturated mode, when $V_{GS} - V_T \leqslant V_{DS}$,

$$I_{DS} = \frac{k}{2}(V_{GS} - V_T)^2 \qquad 11.51$$

A plot of V_{GS} versus $\sqrt{I_{DS}}$ extrapolated to $I_{DS} = 0$ will give V_T. A procedure often used is to pick two points from a curve tracer plot of I_{DS} versus V_{GS} and, by using Eq. 11.51, calculate V_T. For example, if V_{GS1} is the voltage at one I_{DS} and V_{GS2} is the voltage for an I_{DS} of 10 times the first,

$$V_T = 1.46V_{GS1} - 0.46V_{GS2} \qquad 11.52$$

Sometimes, in order to simplify measurement, V_{TX} for a given transistor may be specified. It designates the gate voltage required to cause some preselected small current to flow and thus requires only one measurement. In V_T measurements, the source and substrate must be tied together; otherwise, V_T or V_{TX} will be displaced by approximately $(1/2)\sqrt{V_{BB}}$. Note that to keep from forward biasing the substrate source junction, the substrate must be positive with respect to the source for p-channel transistors and negative for n-channels. The gate and drain can be tied together to ensure saturation operation and also to give a two-terminal device that is easy to connect to a curve tracer. The V_T measured on a long-channel transistor will not be directly applicable to short-channel devices on the same wafer. In general, short-channel V_T's will be less.

When viewed on a curve tracer, a threshold of a few tenths of a volt might be confused with a forward-biased junction, and one above 5–6 V might be mistaken for a reverse-biased junction. To decide whether the trace belongs to a pn junction or a MOS transistor, plot V_{GS} versus $\sqrt{I_{DS}}$ for a range of voltages above where the break appears. If the trace is indeed for a MOS device, the curve will be a straight line. When the threshold voltage is not as expected,

[16] For more details of device behavior in these regions, see, for example, S.M. Sze, *Physics of Semiconductor Devices*, 2d ed., John Wiley & Sons, New York, 1981.

a change in trapped charge is generally the cause. However, a change in the interface charge density, oxide thickness, wafer doping level, or gate electrode work function could also be the trouble. Thresholds are adjusted via ion implantation, and it is those transistors whose V_T's are usually measured. However, it is also very helpful to know what the V_T would have been without adjustment. Therefore, one or more test transistors may have the threshold adjust omitted. Their V_T's are often referred to as "natural" thresholds.

If the MOS transistor to be studied consists of a lead, the field oxide, and two adjacent diffused regions, the V_T is referred to as the "thick-field turn-on." When this voltage is too low, surface inversion and unwanted shorting of components can occur when voltage is applied to the lead. The causes of a low thick-field turn-on are the same as those causing low V_T's of transistors.

11.9.2 Subthreshold Leakage Current

When the gate voltage is less than V_T, I_{DS} does not go to zero. Rather, a small current, the subthreshold leakage current, remains (23). It has the general character shown in Fig. 11.26. The lower current limit is the reverse current of the drain–substrate junction. The slope $\partial(\log I_{DS})\partial V_{GS}$ of the curve below threshold is proportional to the gate oxide thickness and the square root of substrate doping (24).

11.9.3 Pinchoff Voltage

For depletion-type transistors, the pinchoff voltage V_P is equivalent to V_T for enhancement transistors and is also given by Eqs. 11.50 and 11.51. For these transistors, saturation is not obtained by tying gate to drain. The drain voltage must be substantially greater than the gate voltage. I_{D0} is defined as the current at $V_{GS} = 0$. By measuring it and the current I_{D2} at some other gate voltage V_{GS2}, V_P from Eq. 11.51 is

$$V_P = \frac{V_{GS2}}{1 - \sqrt{I_{D2}/I_{D0}}} \qquad 11.53$$

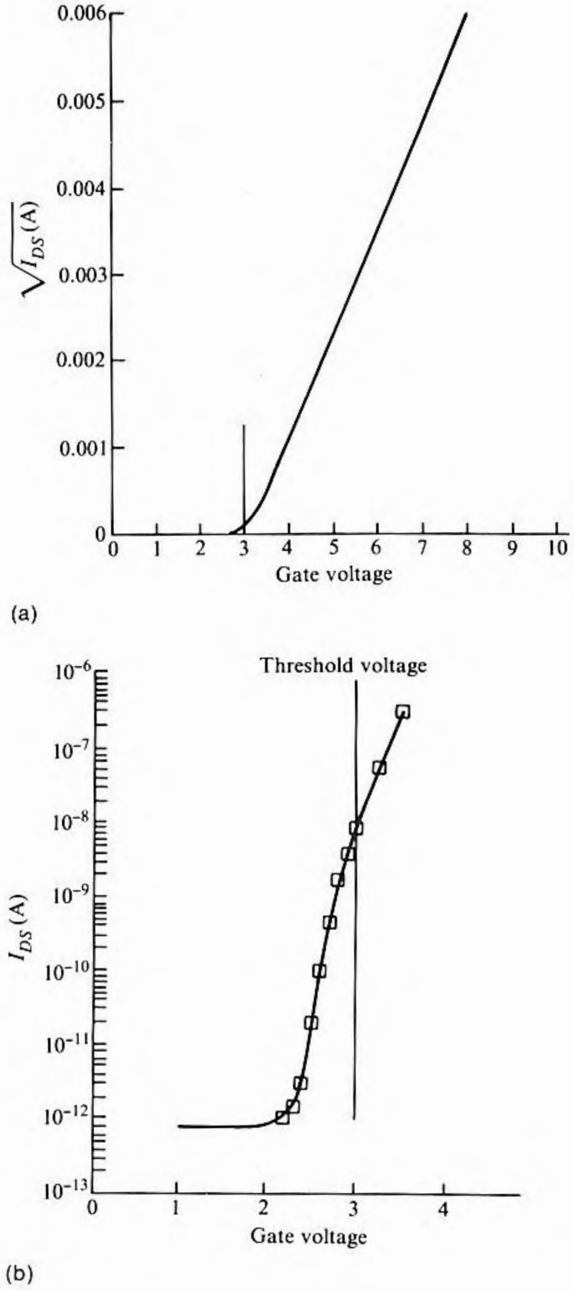
and is somewhat analogous to Eq. 11.52 for enhancement-mode transistors. Alternatively, V_{GS} versus $\sqrt{I_{DS}}$ can be plotted, and the curve can be extrapolated back to $\sqrt{I_{DS}} = 0$. The depletion transistor counterpart to subthreshold current is subpinchoff current (25).

11.9.4 Conduction Factors *k* and *k*′

The conduction factor k is defined through Eq. 11.50 or Eq. 11.51. To remove the effect of channel length and width, k', defined as $k' = (l/w)k$, is sometimes used instead of k. w is the width of the transistor channel, and l is its length. From Eq. 11.51, when the transistor is operating in the saturated mode, a plot of $\sqrt{I_{DS}}$ versus

FIGURE 11.26

Subthreshold current as shown on (a) $\sqrt{I_{DS}}$ versus V_{GS} plot used to determine threshold voltage and (b) log I_{DS} versus log V_{GS} plot. (*Source:* Subthreshold data from W. Milton Gosney, *IEEE Trans. on Electron Dev. ED-19*, p. 213, 1972.)



V_{GS} will have a slope of $\sqrt{k}$. The channel width is usually considerably greater than the length, and, as a consequence, the effect of lateral diffusion under the masking oxide on the width can be neglected in making k' calculations. Thus, the width can generally be determined closely enough by a measurement of the mask. The length, however, must be corrected by the amount of diffusion under the gate. For short-channel transistors, the length obtained in this manner may still be in substantial error, and special long-channel transistors are usually used.

11.9.5 Channel Mobility

To a first approximation,

$$k' = \frac{\mu\varepsilon\varepsilon_0}{2t} \qquad 11.54$$

where μ is the carrier mobility in the MOS channel, not the MOS small-signal gain discussed earlier. ε is the dielectric constant of the gate insulator, ε_0 is the permittivity of free space, and t is the gate insulator thickness. Thus, in principle, μ can be calculated from Eq. 11.54. However, for better accuracy, the drain conductance as $V_{DS} \rightarrow 0$ is normally used. For $V_{DS} << (V_{GS} - V_T)$,

$$I_{DS} = \left(\frac{w}{lt}\right)\mu\varepsilon\varepsilon_0(V_{GS} - V_T)V_{DS} \qquad 11.55$$

Thus,

$$g_D = \frac{\partial I_{DS}}{\partial V_{DS}} = \left(\frac{w}{lt}\right)\mu\varepsilon\varepsilon_0(V_{GS} - V_T) \qquad 11.56$$

Often, g_D near V_{DS} for several values of V_{GS} will be determined in order to obtain a better value. It should be remembered that μ is a tensor quantity and even in cubic crystals will depend on the direction of the current flow when the current is restricted to very thin layers, such as MOS inversion layers (26).

11.9.6 Drain Breakdown Voltage

The drain breakdown voltage BV_{DSS} is measured by tying gate, source, and substrate to ground and measuring the drain voltage required for some small preselected current. The limiting factor in this breakdown is generally the space charge region from the drain punching through to the source. Thus, the shorter the channel length, the lower the BV_{DSS}.

11.10 BIPOLAR TRANSISTOR DIAGNOSTICS

The bipolar transistor DC parameters that may require checking can be divided into the four categories of current gain, breakdown voltages, reverse bias currents, and resistance-dependent properties such as saturation voltage (the last of which is not discussed here).

11.10.1 Current Gain

Since the common emitter transistor configuration is most commonly used in bipolar ICs, h_{fe} is the current gain of most interest. Table 11.5 lists and defines it and several other gain parameters of interest. h_{fe} and α can be measured in a straightforward manner on a curve tracer. Fig. 11.27 shows typical traces when the curve tracer is connected to show common emitter and common base characteristics. In each case, the horizontal axis represents collector–emitter voltage, and the vertical axis represents collector current. In Fig. 11.27a, the family of curves comes from a discrete set of base currents (2, 4, 6, 8, and 10 μA). In Fig. 11.27b, the emitter current is increased in steps of 500 μA, and since α is very close to unity for this transistor, I_C very closely follows the input emitter current.

When h_{fe} is too high, a narrower-than-expected base is the probable problem. When it is too low over the whole current range, the problem is usually a base that has become too wide, thus reducing the base transport efficiency β. However, if the emitter doping level were to be substantially reduced, the emitter efficiency γ would decrease because of a decrease in the ratio of minority to majority carriers injected into the base. When the problem is only a reduction

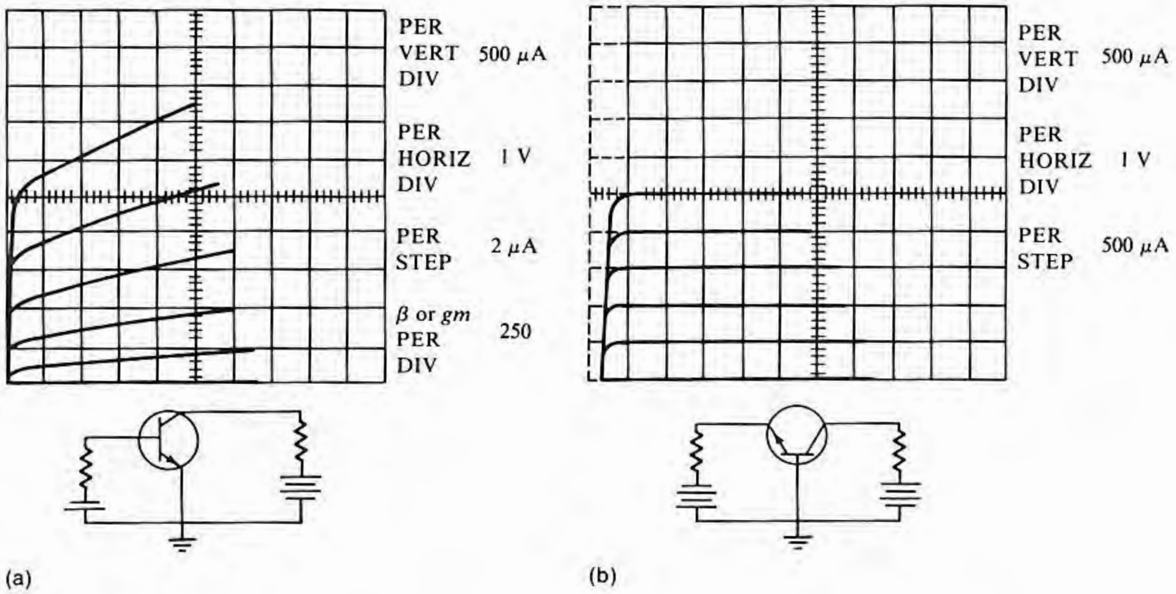
TABLE 11.5

Current Gain Parameters

Symbol	*Meaning*	*Definition*
h_{fe}	Small-signal current gain (common emitter configuration)	$\partial I_C/\partial I_B$ where I_C = collector current and I_B = base current.
Inv. h_{fe}	Inverse small-signal current gain	Measured with collector and emitter terminals interchanged.
h_{FE}	Large-signal current gain (common emitter configuration)	I_C/I_B (see Fig. 11.27a to judge likely value difference from h_{fe}).
α	Small-signal current gain (common base configuration)	$\partial I_C/\partial I_E$ where I_E = emitter current; $h_{fe} = \alpha/(1 - \alpha)$ where $\alpha = \gamma\beta$.
γ	Emitter efficiency	Ratio of minority carrier current crossing into base to total emitter current; $\gamma = \gamma_1\gamma_2$.
γ_1		Ratio of injected emitter minority current to injected majority current.
γ_2		Fraction of injected emitter current lost to recombination in emitter–base space charge region.
β	Base transport efficiency	Fraction of minority current surviving passage across base.
β	Archaic symbol for h_{fe}	

FIGURE 11.27

Typical curve tracer display of (a) common emitter and (b) common base characteristics.



in low-current h_{fe}, then the difficulty is almost certainly emitter efficiency. However, the emitter efficiency reduction in this case is due to excessive carrier recombination in the emitter–base space charge region (bulk or surface).

To see whether β or the emitter efficiency γ is limiting h_{fe}, the common emitter output admittance h_{oe} can be examined. As normally defined,

$$h_{oe} = \left.\frac{\partial I_C}{\partial V_{CE}}\right|_{I_B} \qquad 11.57$$

which is just the slope of a curve in Fig. 11.27a. Many curve tracers allow curves similar to those of Fig. 11.27a to be displayed, but with constant emitter–base voltage (V_{BE}) steps. Generally, either set of curves can be chosen by a simple switch change so that the two can be alternately observed on the screen. If the slope of the two curves are very close, then emitter efficiency determines current gain. If base transport is limiting h_{fe}, the slope of the constant I_B curve will be approximately twice that of the constant V_{BE} curve (27).

The emitter–base current is of the form $I_B = I_0 e^{qV_{BE}/nkT}$ *where* $n = 1$ implies predominately diffusion current and $n = 2$ implies predominately recombination current (see section 11.8.4) and a much lower h_{fe}. When $n = 1$, $\gamma_2 = 1$ and γ_1 determines γ. When $n = 2$, the reverse is true. The V_{BE} region where each mechanism dominates can be seen from either a log I_B–V_{BE} or a $1/h_{FE}$–log I_C plot. For each region, n can be determined directly from a log I_B–V_{BE} curve or from the slope S of the log $(1/h_{FE})$–log I_C curve through the relation $n = 1/(1 + S)$ (28). Fig. 11.28 shows log I_B–V_{BE} plots for two transistors. Unit 1 has very poor low-current h_{FE}, which is reduced to 1 for base current of only 0.2 μA. Unit 2 is considered normal, and although the curve does not extend that low, it can be seen that base current will be less than 0.001 μA before h_{FE} is reduced to 1. Fig. 11.29 shows a $1/h_{FE}$ plot for unit 1. While not shown on either of the figures, at somewhat higher currents, h_{FE} will start decreasing with increasing current because of high current density effects. Should it be desired to determine whether bulk or surface recombination is dominant, an I_B–T plot may allow a determination. Bulk recombination processes usually have an activation energy of about $E_G/2$, where E_G is the bandgap. Surface recombination tends to be nearly temperature independent (29).

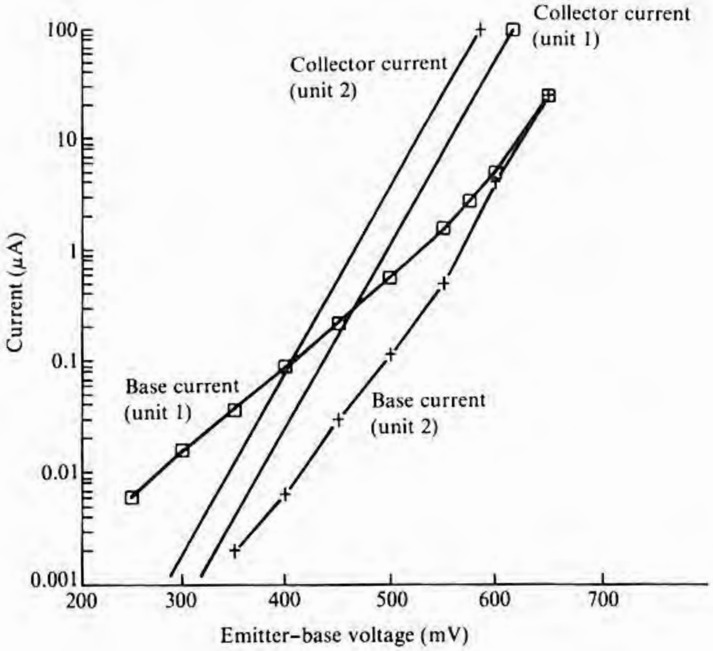
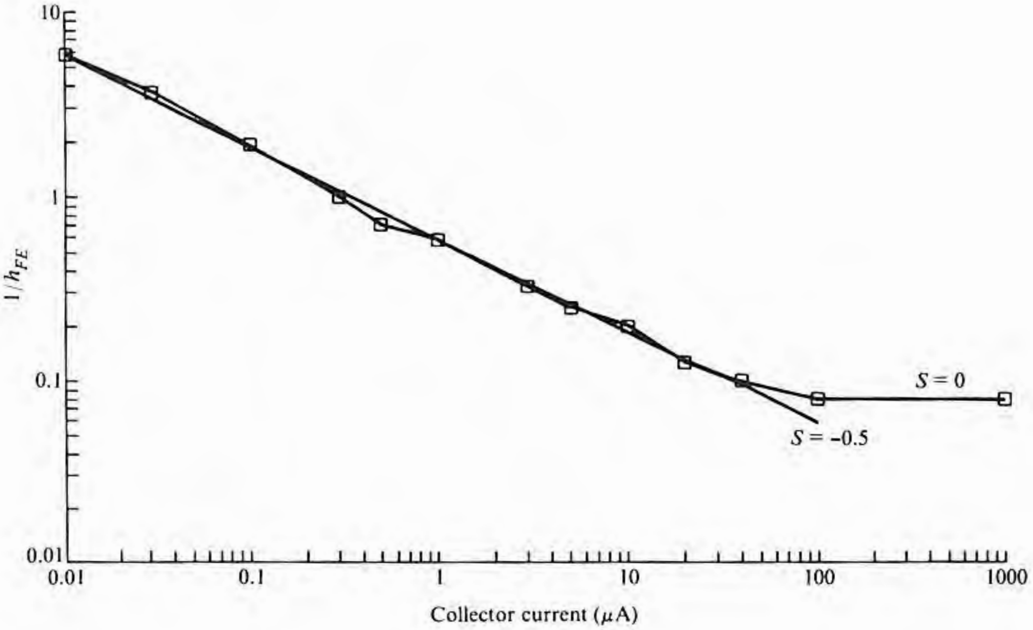


FIGURE 11.28

Base and collector currents versus emitter–base voltage for two units. (Unit 1 has very poor low-current h_{FE} because of excessive emitter–base recombination.)

FIGURE 11.29

Plot of $1/h_{FE}$ versus current showing a region of high emitter–base recombination for unit 1.



11.10.2 Reverse Breakdown Voltages

When a device with multiple junctions, such as a bipolar transistor, is examined, a variety of interactions can change breakdown voltages. Depending on the manner in which the junctions are electrically connected, several different breakdown voltages may be observed. By studying their relative magnitudes, considerable insight into some varieties of processing problems can be gained. The bipolar transistor breakdown voltages are listed in Table 11.6.

$V_{(BR)CEO}$ is the voltage most different from isolated junction breakdown. The difference arises because when a bipolar transistor is operated with the base floating, avalanche will occur when the multiplication factor (section 11.8.1) approaches $h_{FE}/(1 + h_{FE})$ rather than when it becomes very large. Simple theory[17] predicts that $V_{(BR)CEO}/V_{(BR)CBO} \cong (1/h_{FE})^{1/P}$ where P is a number experimentally observed to be about 4 for silicon. As an example of what this will predict, consider a transistor with an h_{FE} of 100 and a $V_{(BR)CBO}$ of 60

[17] For a discussion of this effect, see any standard text on semiconductor devices.

TABLE 11.6

Summary of Active Bipolar Transistor Device Breakdown Voltages and Their Abbreviations

IEEE Symbol	*Common Symbol*	*Definition*
$V_{(BR)CBO}$	BV_{CBO}	Reverse breakdown voltage, collector to base, with emitter open: This voltage is the same as that which would be observed if there were no emitter–base junction—that is, if $V_{(BR)CBO} = V_{(BR)}$.
$V_{(BR)CEO}$	BV_{CEO}	Reverse breakdown voltage, collector to emitter, with base open: This voltage is ordinarily much less than $V_{(BR)CBO}$ and is dependent on the current gain of the transistor.
$V_{(BR)CES}$	BV_{CES}	Reverse breakdown voltage, collector to emitter, with base shorted to emitter: This voltage lies between $V_{(BR)CEO}$ and $V_{(BR)CBO}$ and should be close to $V_{(BR)CBO}$.
$V_{(BR)CER}$	BV_{CER}	Reverse breakdown voltage, collector to emitter, with a specified resistance between base and emitter: For $R = \infty$, $V_{(BR)CER} = V_{(BR)CEO}$, and for $R = 0$, $V_{(BR)CER} = V_{(BR)CES}$. Thus, $V_{(BR)CER}$ will always lie between $V_{(BR)CES}$ and $V_{(BR)CEO}$.
$V_{(BR)EBO}$	BV_{EBO}	Reverse breakdown voltage, emitter to base, with collector open.

V. Then, $(1/100)^{1/4} = 0.32$, so the collector–emitter breakdown will be just over 20 V. If a resistor is connected from base to emitter, the collector–emitter *I–V* trace will snap down from $V_{(BR)CBO}$ to $V_{(BR)CEO}$ or some intermediate value, depending on the resistor value, as the current is increased. In the case of silicon, the transistor itself behaves as though it has a built-in resistor. Germanium, however, does not show the effect.

When the observed CEO/CBO ratio is larger than expected, it is generally because the $V_{(BR)CBO}$ value being used is not the true value, but one reduced because of, for example, a collector region that is too thin. If the ratio is smaller than predicted, punchthrough may be limiting $V_{(BR)CEO}$. Punchthrough occurs when the collector–base junction space charge region widens into the base enough to reach the emitter–base junction. $V_{(BR)CEO}$ will be reduced, but $V_{(BR)CBO}$ will be unaffected. It can be detected by noting when (or if) the emitter floating potential[18] begins to rapidly increase as V_{CB} is

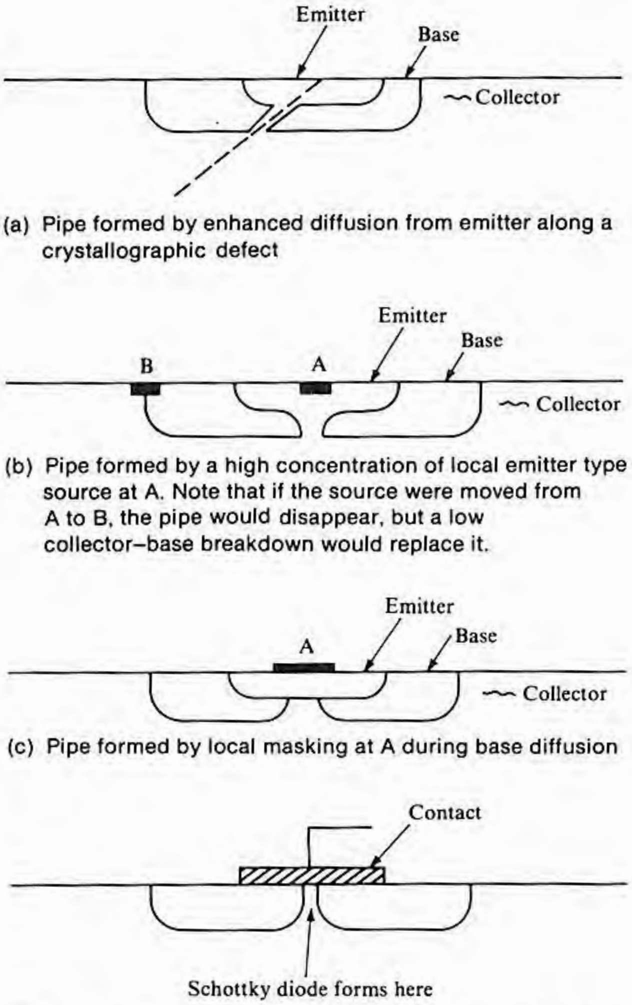
[18] Emitter floating potential is the voltage between emitter and ground with the emitter floating and a reverse voltage applied from collector to base. With no punchthrough, the value for silicon will be about 200 mV.

varied over the operating range. Punchthrough[19] can be caused either by a thinner-than-expected base (perhaps only locally due to a nonuniform diffusion front) or by a higher-than-normal base resistivity.

The ratio $V_{(BR)CES}/V_{(BR)CBO}$ is normally close to 1. The exact value depends on inverse current gain, but it should be greater than 0.8

FIGURE 11.30

Examples of pipe formation in pn structures.



(a) Pipe formed by enhanced diffusion from emitter along a crystallographic defect

(b) Pipe formed by a high concentration of local emitter type source at A. Note that if the source were moved from A to B, the pipe would disappear, but a low collector–base breakdown would replace it.

(c) Pipe formed by local masking at A during base diffusion

(d) Pipe similar to the one in part c, but the diode diffusion was partially blocked in a region later covered by a contact

[19] Depending on the specific design, punchthrough may occur even on good transistors, but outside the normal operating range.

for an inverse gain of less than 0.5. If it is appreciably less, punchthrough is probably limiting it.

When pipes are present, they will radically change the relations among the various breakdown voltages. Pipes are small cross-sectional electrical paths formed by diffusion flaws and are more likely to be a problem in bipolar transistors. Examples of pipes and some of their causes are shown in Fig. 11.30. If a pipe like the one of Fig. 11.30a, 11.30b, or 11.30c occurs, the emitter–base breakdown voltage $V_{(BR)EBO}$ will appear normal. However, the collector–base breakdown will also initially equal $V_{(BR)EBO}$ since the emitter and collector are tied together by the pipe. However, as the voltage is increased, the current will not increase as rapidly as normally observed during avalanche since the resistance of the pipe will act as a current limiter. The space charge movement into the pipe will eventually pinch off additional current flow so that the classical pipe I–V curve looks like the curve of Fig. 11.15b. (In the rare event that the pipe is caused by a metallic path, or if the pipe has a very large cross section, the path will appear resistive and will not pinch off.) When a pipe is present, the application of an emitter–collector voltage will cause excess current immediately. If the pipe is small enough, it will pinch off and allow $V_{(BR)CEO}$ to be observed and will look like the curve of Fig. 11.15c. Inversion layers sometimes form paths across device surfaces, and their effect may superficially resemble pipe behavior. With voltage between collector and emitter, both pipe and surface inversion will have the same characteristics. However, in looking from collector to base, only an inversion layer will cause current flow to begin immediately. Fig. 11.30d shows a pipe-like defect that will cause a Schottky diode to form over a small area and be in parallel with the pn junction. Since no care will have been exercised in forming the Schottky diode, it will be very leaky in the reverse direction and hence cause the whole junction to appear leaky. This defect can be found in MOS as well as in bipolar structures.

11.10.3 Reverse and Forward Currents

The bipolar transistor reverse current definitions are given in Table 11.7. When behavior is normal, I_{CBO} is just that of a reverse-biased pn junction, and the others can be calculated[20] from it and forward and reverse h_{FE}. For most cases, $I_{CBO} < I_{CES} < I_{CEO}$. Note that while these currents have been defined as "saturation currents," they are seldom completely independent of voltage. Thus, when a numerical value for I_{XYZ} is given, the voltage at which it was measured must

[20] See a standard text on transistor theory.

TABLE 11.7

Bipolar Transistor Reverse Current Definitions

Symbol	*Meaning*
I_{EBO}	Emitter saturation cuirrent, with collector open
I_{ECO}	Emitter saturation current, with base open
I_{CBO}	Collector saturation current, with emitter open
I_{CEO}	Collector saturation current, with collector open
I_{CES}	Collector saturation current, with base shorted to emitter
I_{CER}	Collector saturation current, with base connected to emitter through a resistor R

also be given. Excess reverse currents that may flow in the various configurations were discussed in the previous section.

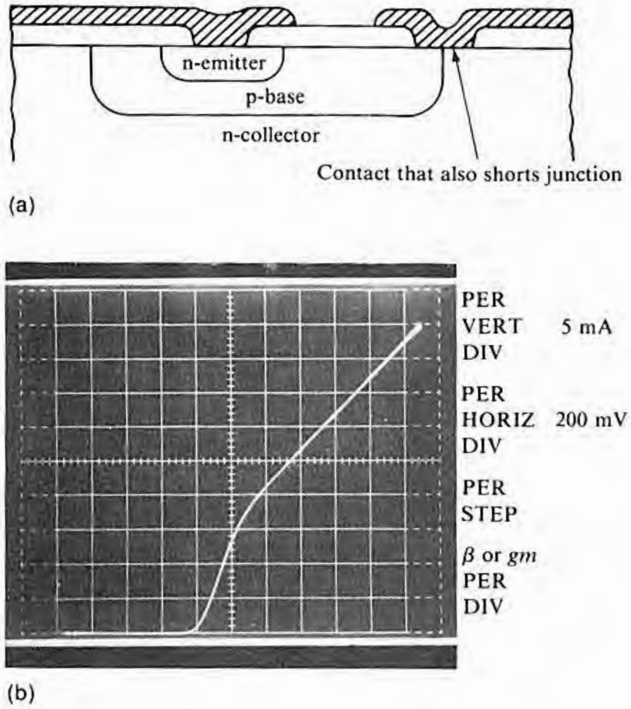
Forward transistor junction currents also depend on the manner in which the junctions are connected. Since IC diodes are often made by shorting out one junction, the resulting forward current–voltage characteristics may differ from those expected from an isolated junction (30). In particular, the emitter–base current with the collector–base junction shorted is substantially different, as shown in Fig. 11.31. When such a curve is encountered unexpectedly, at lower current levels it might be identified as a Schottky forward, but as the current and voltage are increased (as shown in the figure), it will switch back toward normal isolated pn junction behavior. The point of switchback is dependent on the transistor characteristics and collector resistivity and occurs when enough current flows to saturate the transistor.

11.11 YIELD ECONOMICS

Since the processes, yields, and economics of a semiconductor operation are interrelated, it is necessary to understand the impact of the various yields and processes on overall IC manufacturing costs in order to plan efficient yield improvement programs. As will be seen in the next section, for the case of silicon ICs, the cost of the silicon itself is generally a very small part of the overall cost. The cost of manufacturing the chip will range from about 1/5 the assembled unit cost for a very small device such as a TTL gate to perhaps 9/10 the assembled unit cost for a large chip device like a 1 megabit DRAM. Assembly/test yields are generally above 95%, while multiprobe yields probably average no more than 50% (although some chip yields may exceed 90%). Thus, because of the relatively low chip yields and the relatively high percentage of the total finished-unit cost represented by chips, most IC manufacturing emphasis is on chip yield improvement.

FIGURE 11.31

Effect of a shorted collector–base junction on emitter–base current.



11.11.1 Silicon Usage

The overall yield for the silicon semiconductor manufacturing operation, from polycrystalline semiconductor-grade silicon to the outgoing packaged unit (IC or discrete transistor), is sometimes calculated in terms of grams of polysilicon per device. Such yields are useful in determining the industry requirements for polysilicon, given a device unit market projection. These requirements, in turn, are examined by IC manufacturers interested in knowing whether or not future polysilicon shortages that could limit their growth are likely; by polysilicon manufacturers concerned about whether they are over- or underbuilding poly plant capacity; and by government agencies concerned about the health of the semiconductor industry in their respective countries.

The actual quantity of single crystal in a packaged unit is quite small. A silicon 256K DRAM has about 0.05 gram of silicon per unit; a TTL gate IC, about 0.5 milligram. Table 11.8 summarizes the silicon losses at the yield points discussed earlier and shows that only about 10% of the starting polysilicon actually finds its way into a finished IC. Thus, if 0.05 gram of silicon is in a package, then no

TABLE 11.8

Overall Silicon Yield

Step	*Yield (% by weight)*
Polysilicon to single crystal	50
Single crystal to polished slice	50
Slice to finished wafer	70
Finished wafer to chip*	40–80
Chip to finished IC	93
Overall yield	6.5–13

*If the wafer is thinned before being broken into individual chips, the yield at this point will be commensurately less.

TABLE 11.9

Elements of Cost for Producing a Finished Good Wafer

Item	*Normalized Cost**
Starting slice	20
Other direct material	12
Masks/slice	
Chemicals/slice	
Labor	12
Labor-associated overhead	8
Overhead materials	8
DI water	
Gases	
Other	
Process control slice	
Depreciation	40
Facility	
Equipment	
Total	100

*Estimated values. Actual numbers will vary substantially, depending on size and age of facility, capacity utilization, kind of product being built, and so on.

more than 0.5 gram of polysilicon would have been used, and at a price of $60/kilogram, the cost of the polysilicon would contribute only 3¢ to the cost of the chip.

11.11.2 Wafer Cost

The actual cost of producing a wafer in a given IC factory is generally considered to be a trade secret, but elements that make up the cost are given in Table 11.9. Tables such as this one can be used to help estimate whether, for example, the yield or productivity improvement to be derived from a new piece of equipment will more than offset the added depreciation.

11.11.3 Wafer Diameter

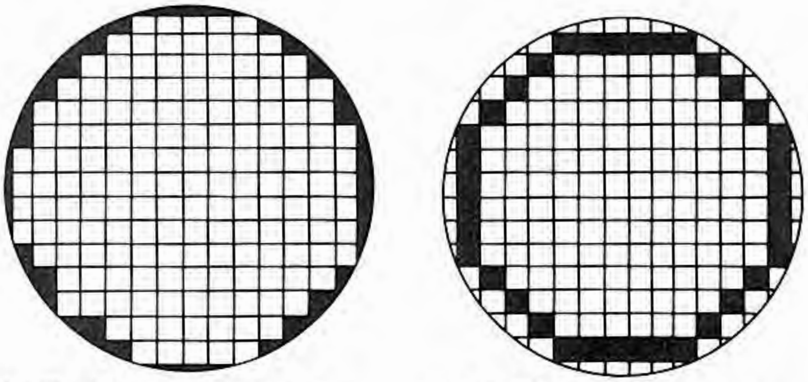
The purpose of the original drive to change from 1.5 inch to 2 inch diameter slices was to reduce labor. The reasoning was that no more effort should be involved in handling 2 inch wafers than in handling 1.5 inch wafers. However, the move to larger slices also enabled more wafer area to be processed in the same floor space and thus could delay the construction of new facilities. In many cases, larger wafers could be processed in the same equipment and thus give increased capacity with minimal capital expenditure.

Multiprobe yield has been observed to increase along with diameter. This increase is due to two related effects. The first is that as diameter increases, the area on the periphery containing partial chips becomes a smaller part of the total wafer area. The second is that the outer ring of whole chips is generally yield depressed because of such things as wafer–edge rounding, resist-edge beading, shadowing of the periphery by reference pins of front-side referencing equipment, and damage from wafer-handling equipment. Both of these effects are shown in Fig. 11.32. The larger the chip is with respect to the wafer, the more pronounced is the improvement in going to larger diameters.

As the wafer diameter has increased in 25 mm increments from 50 mm to 150 mm, the chip cost has decreased. The maximum economical diameter at a given time has been dictated by various economic or technological factors such as labor costs or the availability of suitable equipment. However, at some point, a limit is almost certain to be set by the semiconductor's high-temperature yield point and thermal conductivity.

FIGURE 11.32

Effect of wafer diameter on multiprobe yield. (The example shown would scale to 1 × 1 cm^2 chips on a 150 mm wafer, and 23% of the total chips are in this 150 mm row. Thus, if they have a reduced yield, the yield of the whole wafer is severely impacted.)



(a) Shaded portion shows area lost due to incomplete chips

(b) Shaded portion shows area occupied by outer row of whole chips

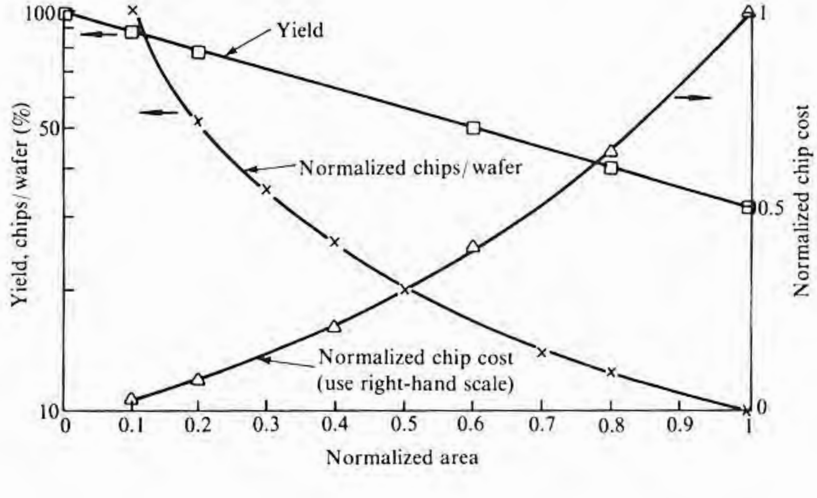
11.11.4 Effect of Chip Size on Yield and Cost

Since for a fixed process, the chip multiprobe yield decreases as the chip size increases, the number of good chips per wafer decreases with chip size both because of decreasing yield and because of fewer potential chips per wafer. The cost of producing a wafer is essentially independent of chip size as long as the process remains the same, so the cost per chip will increase with increasing chip size in the manner shown in Fig. 11.33. The designer must therefore be as frugal with area as possible in the initial design. After a new chip is introduced and the manufacturing and design kinks are eliminated, a size reduction (shrinkage) of the chip can be undertaken in order to move down the chip cost curve. Initially, the mask image size is reduced by a small amount (perhaps 10% in linear dimensions). Such a reduction reduces all geometries and may make the yield drop instead of increase. However, experience has shown that generally some reduction is possible before this happens. Even when a marked yield drop does occur, it is likely that there are a few localized spacing problems that can be corrected with a simple mask revision.

As just mentioned, the simplest method of reducing chip size (other than by clever designs) is by geometry reduction. However, as complexity passes some point, the area required for interconnections becomes a limitation. Then, a move is generally made to multilevel metallization, which can only be done if the extra steps required for the additional levels of metallization do not reduce the yield as much as the chip area shrinkage helps it. Indeed, not until

FIGURE 11.33

Effect of chip area on cost.



the late 1970s was the process improved to the point that a yield crossover could be achieved.

11.11.5 Revenue from Wafers

The discussion thus far has been related to the costs accrued in producing a wafer. In a competitive world, these costs must be covered by the revenue derived from the wafer.[21] Such revenue is generally described as the net sales billed per wafer, or NSB/wafer. The NSB depends on what value the market places on the specific circuits built on the wafer and on the number of good chips on the wafer. A wafer containing a very large number of small simple devices—a TTL gate, for example—can have a higher value than a mature product requiring a much larger chip, such as a 64K MOS DRAM. The reason is that the selling price of small-chip devices is more influenced by packaging costs than by chip complexity. At the other end of the spectrum, the relatively new, more complex, large-chip circuits command enough higher price to more than offset the reduced number of chips and thus cause the NSB/wafer to again rise. Table 11.10 gives details of the makeup of NSB/wafer for some representative ICs, and Fig. 11.34 shows the general trend of NSB/wafer versus chip area. The values for the low and middle portions of the curve are relatively stable with time (neglecting inflation) since the products and packaging techniques are quite mature. The upper end of the curve keeps shifting as new products move down the experience curve.[22]

TABLE 11.10

Elements Needed To Calculate NSB/Wafer

Product	*Chip Size (cm^2)*	*Number of Chips/Wafer*	*Overall Yield (%)*	*Price/Unit ($)*	*NSB ($)*
TTL gate	0.0064	18,875	75	0.12	1700
64K DRAM	0.15	776	66	0.75	384
256K DRAM	0.39	276	60	2.00	331
Microprocessor	0.5	220	40	20.00	1760

Note: Estimated numbers are for a 125 mm wafer diameter for the mid-1980s. Actual yields and costs will vary from manufacturer to manufacturer and are generally considered to be trade secrets.

[21] A "merchant" semiconductor operation is assumed. Captive wafer-fab areas may have more leeway since their products are typically custom ICs with high leverage at the system level.

[22] The experience curve is an expansion of the "learning curve" concept that was apparently introduced by the aircraft industry during World War II.

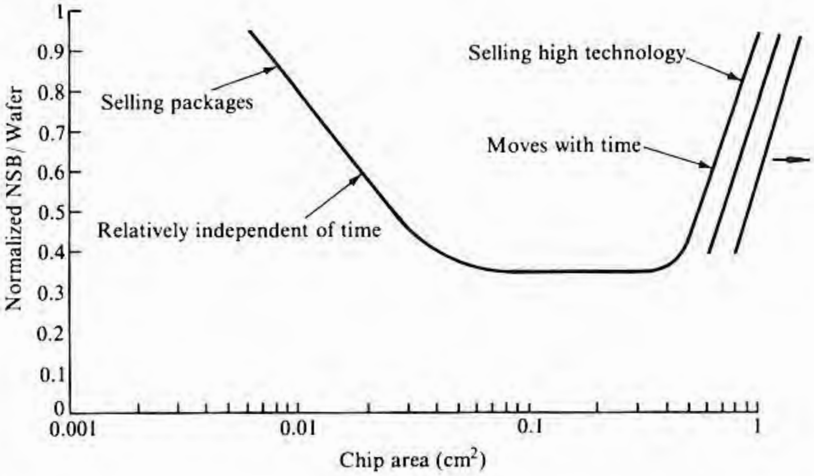


FIGURE 11.34

General trend of impact of time and technology on NSB/wafer.

11.12 FACTORS AFFECTING YIELD AND YIELD IMPROVEMENT

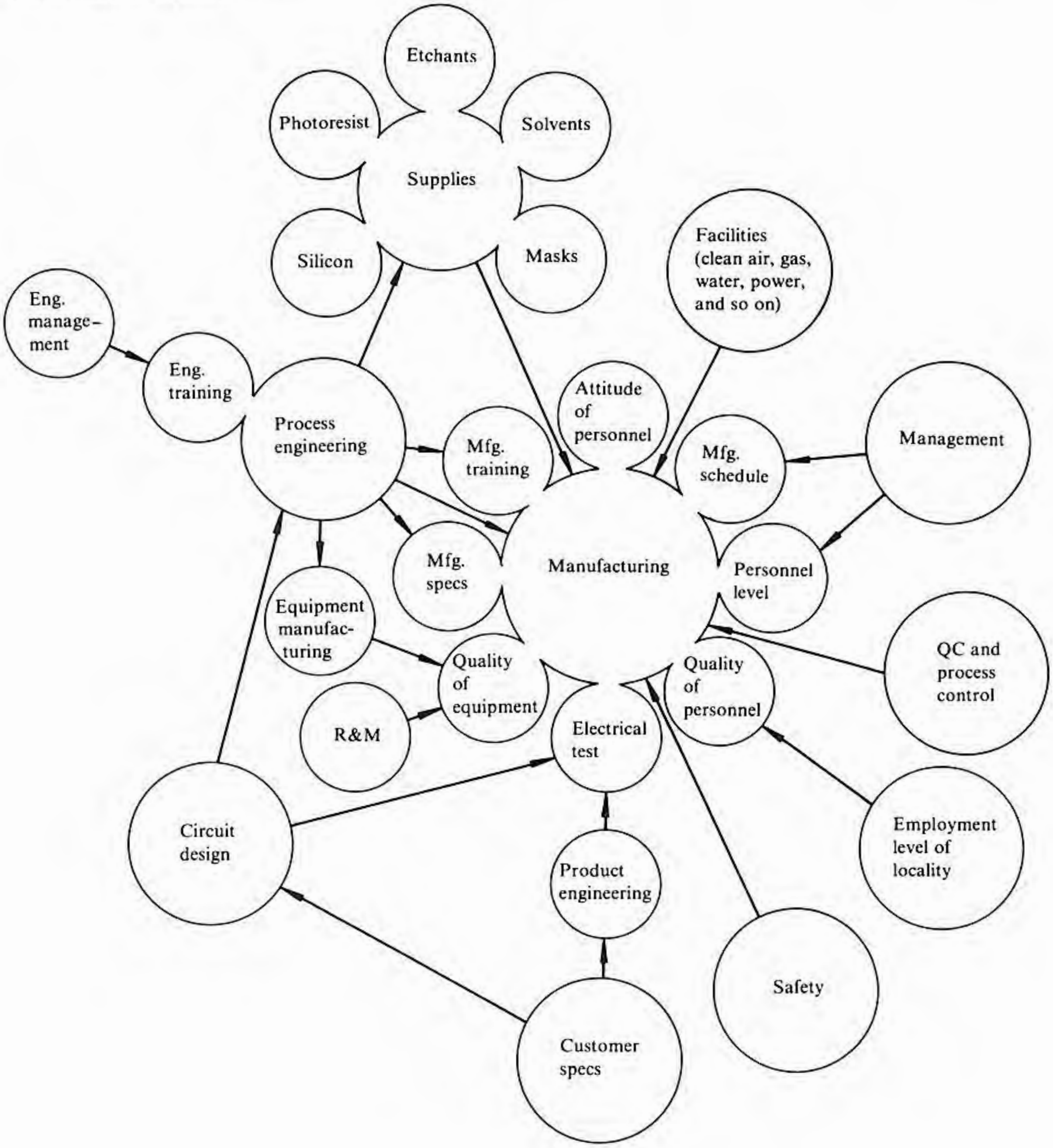
The discussion thus far has been directed primarily at ways of evaluating and tracking yield, with some discussion of how various yields affect the final economic performance of a wafer-fab facility. In order to improve the yield, it is of course first necessary to determine the specific physical defect causing a problem, as was briefly discussed in the previous section. However, to determine the root cause and to effect a permanent cure may be a much more difficult task. Fig. 11.35 depicts factors, many rather obscure, that can affect yields. Despite the complexity of this figure, the factors can be broadly grouped as follows:

People
Equipment
Incoming materials
Device design
Semiconductor processing

Only the last item, semiconductor processing, is dependent on items historically considered as the domain of process engineering. It is clear, however, that the whole span of possibilities, and not just the technical aspects of semiconductor processing, must be considered before instituting changes designed to permanently improve yields. To this end, Fig. 11.35 can be used as a checklist as each factor is considered in the context of the specific manufacturing environment.

FIGURE 11.35

Factors affecting yields.



CHAPTER 11 KEY IDEAS

- □ Simple theory predicts that device yield Y versus area A should vary as $Y = e^{-DA}$ where D is the defect density.
- □ The simple yield expression $Y = e^{-DA}$ gives a pessimistic prediction for large DA products.
- □ Multiprobe tests designed to test chips to customer specifications are usually not satisfactory for yield diagnostics.
- □ Unless special analytical procedures are used, long-term yield trend lines are generally too insensitive to changes to be very useful.
- □ Because of the normal variability in wafer yield, split lot results must be carefully analyzed to determine whether or not any observed yield differences are significant.
- □ Before an electrical study of a malfunctioning circuit is begun, the circuit should be given a careful visual (microscopic) examination.
- □ In both MOS and bipolar circuits, the study of pn junction leakage current and breakdown voltage is a key part of low-yield analysis.
- □ The dominant component of forward or reverse junction current can generally be determined from, respectively, a log I–V or a log-I–log-V plot.
- □ A common cause of junction failure is an excessive concentration of heavy metals in or near the junction space charge regions.
- □ Low threshold voltage is a common cause of MOS device failure.
- □ In most cases, the cost of the silicon in a silicon IC is a small part of the total IC cost.
- □ In addition to the technical issues involved in yield improvement, a large number of economic and people-related factors must also be considered.

CHAPTER 11 PROBLEMS

1. If the multiprobe yield of a group of similar devices with 5 significant levels is as follows (yield in percent/area in mm^2), determine their defect density. Which equation was used? Why?

55	6
39	9.4
38	10
30	12
21	16
15	19

2. If a chip of area 1 mm^2 has a yield of 83%, what would the expected yield be for a chip of area 1.5 mm^2? Explain your choice of equations.

3. If an incoming 125 mm slice has 8 randomly spaced point defects on it, how much will the yield of a 0.5 cm^2 chip be reduced by these defects if there are 8 processing levels, each of which introduces 0.1 defect/cm^2?

4. What, if anything, is wrong with the diode whose reverse I–V characteristics are as follows? (Units are A/V.)

4×10^{-12}	0.5
7×10^{-12}	1
2×10^{-10}	4
10^{-8}	20
10^{-7}	20.5
10^{-3}	23

5. What device characteristic is most likely being measured when the following *I–V* data are obtained? (Units are A/V.) Plot the data in a manner that demonstrates your answer.

10^{-12}	0.5
1.2×10^{-12}	2
2×10^{-11}	2.5
9×10^{-9}	3
1.4×10^{-6}	4
2.5×10^{-5}	7

6. Sketch the impurity profile for a diode most likely to have the following *I–V* curve. (Units are A/V.)

10^{-12}	-1
1.25×10^{-12}	2
1.6×10^{-12}	5
2×10^{-12}	10

7. When a large number of 1 cm^2 silicon IC chips is produced, estimate the amount of semiconductor-grade polysilicon required for each one if the multiprobe yield is 55%. Assume that the saw used in slicing the crystal into slices cuts a 500 μm wide slot. List all other assumptions made.

CHAPTER 11 REFERENCES

1. William H. Woodall, "The Design of CUMSUM Quality Control Charts," *J. Quality Technol. 18*, pp. 99–102, 1986. (For older references, see H. Mack Truax, "Cumulative Sum Charts," *Industrial Quality Control*, pp. 18–25 and references therein, December 1961.
2. The application of the Cumulative Sum Approach to yield tracking was suggested by John Conroy of Texas Instruments Incorporated.
3. Frederick E. Croxton and Dudley J. Cowden, *Applied General Statistics*, Prentice-Hall, Englewood Cliffs, N.J., 1955.
4. J.T. Wallmark, "Design Considerations for Integrated Electronic Devices," *Proc. IRE 48*, pp. 293–300, 1960.
5. S.R. Hofstein and F.P. Heiman, "The Silicon Insulated-Gate Field-Effect Transistor," *Proc. IEEE 51*, pp. 1190–1202, 1963.
6. C.H. Stapper, "Comments on 'Some Considerations in the Formulation of IC Yield Statistics'," *Solid-State Electronics 24*, pp. 127–132, 1981.
7. B.T. Murphy, "Cost–Size Optima of Monolithic Integrated Circuits," *Proc. IEEE 52*, pp. 1537–1545, 1964.
8. R.B. Seeds, "Yield and Cost Analysis of Bipolar LSI," *Proc. IEEE Int. Elec. Dev. Meeting*, October 1967.
9. Anil Gupta and Jay W. Lathrop, "Yield Analysis of Large Integrated-Circuit Chips," *IEEE J. Solid-State Circuits*, pp. 389–395, 1972.
10. R.M. Warner, Jr., "Applying a Composite Model to the IC Yield Problem," *IEEE J. Solid-State Circuits*, pp. 86–95, 1974.
11. Charles H. Stapper, "On a Composite Model to the IC Yield Problem," *IEEE J. Solid-State Circuits SC-10*, pp. 537–539, 1975.
12. John E. Price, "A New Look at Yield of Integrated Circuits," *Proc. IEEE 58*, pp. 1290–1291, 1970.
13. This derivation is courtesy of Fred Strieter.
14. This derivation is courtesy of Steve Cowdrey.
15. Albert V. Ferris-Prabhu, "Role of Defect Size Distribution in Yield Modeling," *IEEE Trans. on Electron Dev. ED-32*, pp. 1727–1736, 1985.
16. C.H. Stapper, A.N. McLaren, and M. Dreckmann, "Yield Model for Productivity Optimization of VLSI Memory Chips with Redundancy and Partially Good Product," *IBM J. Res. Develop. 24*, pp. 398–409, 1980.
17. Paul Cox et al, "Statistical Device Characterization and Parametric Yield Estimation," *Solid State Technology*, pp. 154–160, and references therein, August 1985.
18. William Shockley, "Problems Related to p-n Junctions in Silicon," *Solid-State Electronics 2*, pp. 35–67, 1961.
19. J.C. Inkson, "An Investigation of Inversion Layer Induced Leakage Current in Abrupt p-n Junctions," *Solid-State Electronics 13*, pp. 1167–1174, 1970.

20. D.J. Fitzgerald and A.S. Grove, "Mechanism of Channel Current Formations in Silicon p-n Junctions," in M.E. Goldberg and Joseph Vaccaro, eds., *Physics of Failure in Electronics*, Vol. 4, Rome Air Development Center, 1966.
21. Chih-Tang Sah, "Effect of Surface Recombination and Channel on p-n Junction and Transistor Characteristics," *IRE Trans. on Electron Dev. ED-9*, pp. 94–108, 1962.
22. Sorab K. Ghandhi, *The Theory and Practice of Microelectronics*, p. 282, John Wiley & Sons, New York, 1968.
23. W. Milton Gosney, "Subthreshold Drain Leakage Currents in MOS Field Effect Transistors," *IEEE Trans. on Electron Dev. ED-19*, pp. 213–219, 1972.
24. R.R. Troutman, "Subthreshold Slope for Insulated Gate Field-Effect Transistors," *IEEE Trans. on Electron Dev. ED-22*, pp. 1049–1051, 1975.
25. Thomas E. Hendrickson, "Subpinchoff Conduction in Depletion-Mode IGFETS," *IEEE Trans. on Electron Dev. ED-25*, pp. 425–431, 1978.
26. D. Coleman, R.T. Bate, and J.P. Mize, "Mobility, Anisotropy and Piezoresistance in Silicon p-Type Inversion Layers," *J. Appl. Phys. 39*, pp. 1923–1931, 1968.
27. Lowell E. Clark, "High Current-Density Beta Diminution," *IEEE Trans. on Electron Dev. ED-17*, pp. 661–666, 1970.
28. W.H. Schroen, J.G. Aiken, and G.A. Brown, "Reliability Improvement by Process Control," *Proc. 10th Annual Reliability Physics Symposium*, pp. 42–48, 1972.
29. A.A. Bergh and C.Y. Bartholomew, "The Effect of Heat-Treatment on Transistor Low Current Gain with Various Ambients and Contamination," *J. Electrochem. Soc. 115*, pp. 1282–1286, 1968.
30. David K. Lynn et al., eds., *Analysis and Design of Integrated Circuits*, p. 253, McGraw-Hill Book Co., New York, 1967.

APPENDIX A

Crystallography

A.1 Silicon and Gallium Arsenide Structure

At atmospheric pressure, silicon (along with diamond, germanium, and gray tin) has a diamond cubic structure. Its point group is m3m, and its space group is Fd3m. Gallium arsenide and other III–V compounds have a zinc blende structure. The diamond lattice can be described as two interpenetrating face-centered cubes displaced $(a_0/4, a_0/4, a_0/4)$ from each other along the right-handed x–y–z coordinate system as shown in Figs. A.1a and A.2. a_0 is the lattice spacing (the length of the unit cell). Ordinarily, distances are normalized to a_0 so that the position is written as (1/4,1/4,1/4). When one of the face-centered cubes is populated solely by one element and the other cube is populated by a different element, the structure is zinc blende. The two cubes are shown isometrically in Fig. A.1a, while Fig. A.1b shows a top view of the two cubes plus an additional cube. Fig. A.2 shows the unit cell, which is made up of atoms from each of the two face-centered arrays. Shown in this figure are 18 atoms, but only 8 of them belong to a single unit cell. The others are parts of adjacent cells that are not completely shown.

The atoms are bonded together as shown in Fig. A.3. Each atom is surrounded by four others in tetrahedral fashion. In the case of the zinc blende structure, the four bonds from one constituent atom all go to atoms of the other constituent. The bond length is twice the covalent radius for silicon and is the sum of the covalent radii of the two constituents for zinc blende. The angle between any two adjacent bonds is 109° 28′.

When they are viewed in certain directions, the atoms and their bonds appear to lie in sheets. To describe the position of these sheets (planes) and the direction of the bonds, it is convenient to use Miller indices,[1] which were originally developed to map the external

[1] Named after William H. Miller, professor of mineralogy at the University of Cambridge from 1832 to 1880.

FIGURE A.1

(a) View of two interpenetrating face-centered cubes that comprise diamond lattice. (b) Top view of position of atoms with respect to plane passing through atoms 1, 2, 3, and 4.

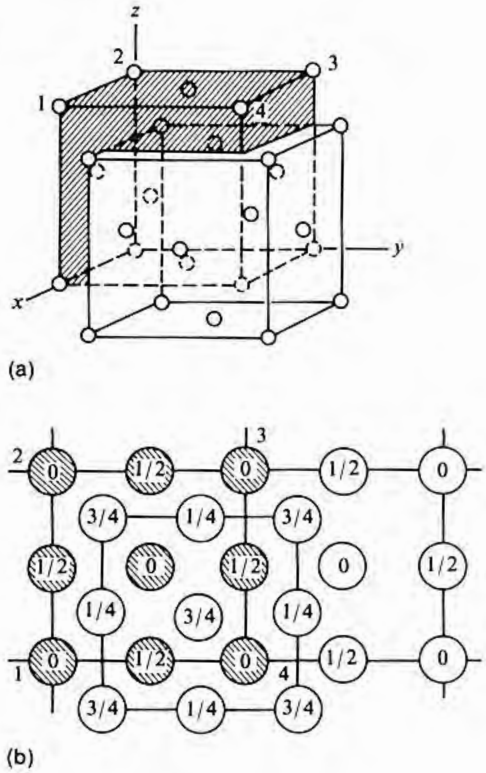
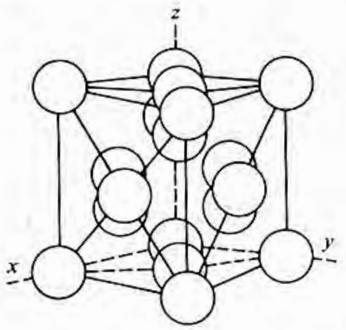


FIGURE A.2

Diamond lattice. [The atoms actually belonging to the unit cell are those with coordinates (0,0,0); (0, ½, ½); (½, 0, ½); (½, ½, 0); (¼, ¼, ¼); (¼, ¾, ¾); (¾, ¼, ¾); and (¾, ¾, ¼).] (*Source:* Adapted from R.W.G. Wyckoff, *Crystal Structures*, Interscience Publishers, New York, 1960.)



faces of crystals. The Miller indices of a plane intersecting the x–y–z axes at distances a, b, and c, respectively, from the origin are the smallest set of integers in the ratio of $1/a$ to $1/b$ to $1/c$. For describing the plane containing a specific set of atoms within a unit cell, the reciprocals are used directly. Miller indices are written as $(hk\ell)$. Negative numbers are written with a bar over the number, as, for example, $(1\bar{1}1)$, which is read as "one bar-one one." Fig. A.4 shows the low indices planes (111), (110), and (010). Note that when a plane is parallel to an axis, it intersects it at infinity, and the reciprocal is zero. Thus, a (010) plane is parallel with both the x and z axes. There are eight planes that make up the 111 family: (111), $(\bar{1}\bar{1}\bar{1})$, $(\bar{1}11)$, $(1\bar{1}\bar{1})$, $(\bar{1}\bar{1}1)$, $(11\bar{1})$, $(1\bar{1}1)$, and $(\bar{1}1\bar{1})$. They are referred to collectively by using braces instead of parentheses—for example, {111}.

A crystallographic direction is written as $[hk\ell]$. In the cubic system, the $[hk\ell]$ direction is perpendicular to the $(hk\ell)$ plane. A whole family of $hk\ell$ directions is designated as $<hk\ell>$. A direction is a vector, and if it goes from the origin to $x = u$, $y = v$, $z = w$, it has the direction coordinates $[uvw]$. The direction indices $[hk\ell]$ are the

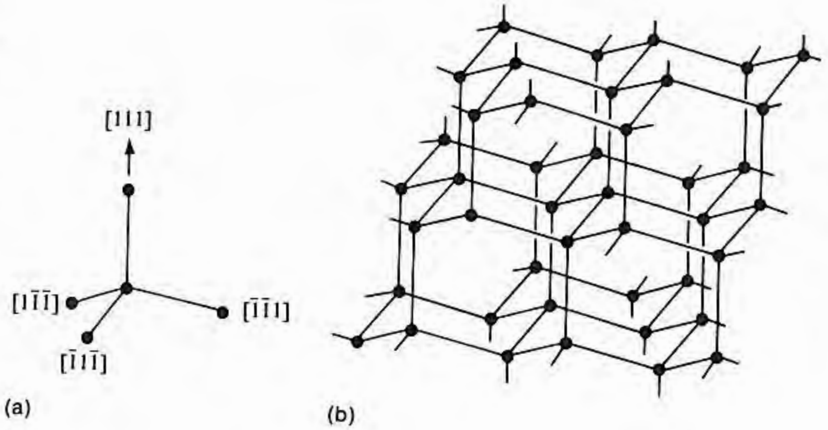


FIGURE A.3

(a) Single atom tetrahedrally bonded to four other atoms. (b) Several atoms joined together to form diamond lattice. (*Source:* Adapted from J. Hornstra, *J. Phys. Chem. Solids* 5, pp. 129–141, 1958.)

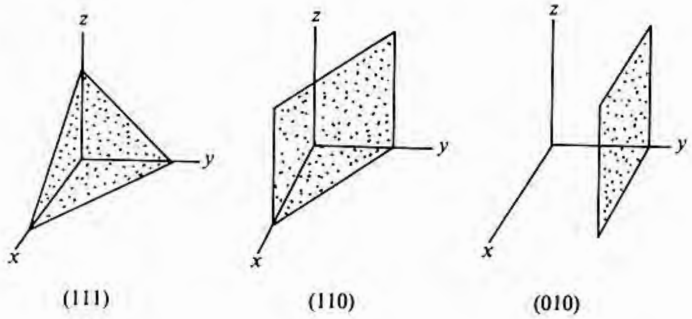


FIGURE A.4

Examples of low indices planes.

smallest set of integers having the ratio of u:v:w. Just as all parallel planes have the same Miller indices, all parallel directions have the same direction indices. Table A.1 provides helpful formulas to use in working with planes and directions and also gives a short table of angles.

All of the planes that intersect at a common line (an infinite number) are said to belong to the same zone, and the common line is called the zone axis. Two planes (HKL) and $(hk\ell)$ will have the same zone axis $[uvw]$ if $[uvw]$ is perpendicular to the normal of each plane—that is, when

$$Hu + Kv + Lw = hu + kv + \ell w = 0$$

Fig. A.5 shows a series of planes belonging to the [110] zone. One way of looking at this series of planes is to consider that they resulted from rotating the (001) plane about a line in the [110] direction. Another is that the (001) plane was tilted a given number of

TABLE A.1

Formulas for Calculations Involving Planes and Directions for Cubic Crystals

To Calculate:	*Use the Formula:*
Angle ϕ between (HKL) and $(hk\ell)$ planes	$\cos\phi = \dfrac{Hh + Kk + L\ell}{[(H^2 + K^2 + L^2)(h^2 + k^2 + \ell^2)]^{1/2}}$
Angle θ between $[HKL]$ and $[hk\ell]$ directions	Use expression above for $\cos\phi$.
Direction $[uvw]$ of line of intersection of (HKL) and $(hk\ell)$ planes (this line is sometimes called trace of one plane on the other)	$u = K\ell - kL$ $v = Lh - H\ell$ $w = Hk - hK$
Spacing between adjacent $(hk\ell)$ planes	$d = \dfrac{a_0}{(h^2 + k^2 + \ell^2)^{1/2}}$

Angles between Planes or Directions for Some Selected Low Indices Planes

HKL	*hkℓ*	*Angle*			
100	100	90			
	110	45	90		
	111	54.74			
	311	25.24	72.45		
110	110	60	90		
	111	35.26	90		
	311	31.48	64.76	90	
111	111	70.53			
	311	29.50	58.52	79.98	
311	311	35.10	50.48	62.96	84.78

For a more extensive table (up to 554/554), see R.J. Peavler and J.L. Lenusky, "Angles between Planes in Cubic Crystals," *IMD Spec. Rpt. No. 8*, American Institute of Mining, Metallurgical and Petroleum Engineers.

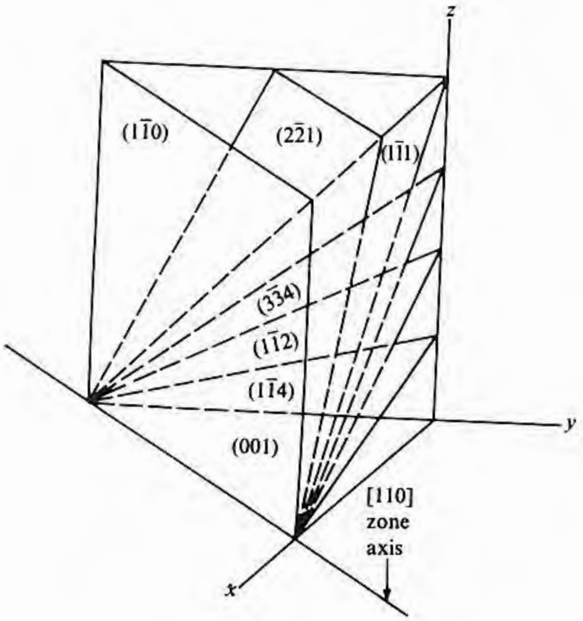
degrees toward the nearest (111) plane. This latter terminology is often used in describing the manner in which wafers are slightly misoriented from a low index plane and is best understood by looking at a stereographic projection.[2]

Examples of stereographic projections for the three common low indices planes are shown in Figs. A.6 and A.7. Dots on the periphery represent planes that are perpendicular to the projection plane (plane of the paper) and oriented so that they are tangent to the peripheral circle at the point of the dot (Fig. A.6b). An interior dot represents a plane making an angle other than 90° with the pro-

[2] For a complete discussion of the mechanics of stereographic projections, see a standard textbook on X-ray crystallography.

FIGURE A.5

Examples of planes belonging to [110] zone.

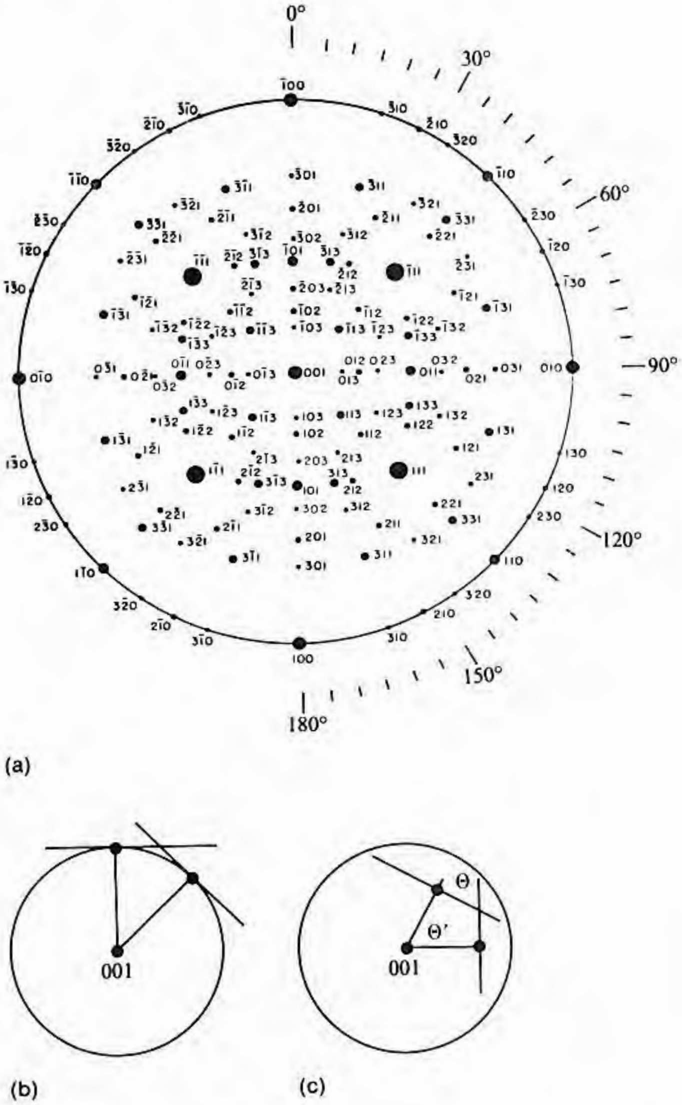


jection plane. The angle between the traces of any two planes intersecting the stereographic plane, as, for example, angle θ in Fig. A.6c, is the same as the angle between the two lines joining the dots with the center of the projection (angle θ′ in Fig. A.6c). Angles between planes represented by interior dots and the projection plane can be determined graphically with a transparent overlay stereographic net, such as a Wulff net. Such a net is not shown, but the closer the dot is to the center of the projection, the smaller the angle between that plane and the projection plane. Thus, using the terminology of tilting toward the nearest {111} plane, it can be seen from Fig. A.6 that any of the four {111} planes are equally close and that to tilt to any of them requires rotating about a <110> zone axis. If the intent was to tilt away from the (111) plane toward the nearest {110}, then reference to Fig. A.7a shows that the zone axis should be a <110> so that the first {110} encountered would be either a (101), a (011), or a (110). Rotation about a <121> zone axis would lead to the (110) planes on the periphery being the first ones reached and would require a greater tilt angle (90°).

Fig. A.8 shows how the atoms and their bonds appear when projected onto (100), (110), and (111) planes—that is, how they appear if viewed by looking into the lattice along [100], [110], and [111] directions, respectively. The scale is the same for all three views, so it is clear why the [110] direction is referred to as the "open" direc-

FIGURE A.6

(a) 001 stereographic projection. (b) Planes represented by peripheral dots are perpendicular to projection plane. (c) Planes represented by interior dots intersect projection plane at an angle other than 90°.



tion. Fig. A.8a shows that three atomic layers exist between the (010) planes at $x = 0$ and at $x = a_0$ [there are also three between the (100)'s and (001)'s]. These are often referred to as (040) planes. Similarly, the spacing for the (110)'s is such that to properly locate them with respect to the individual atoms requires that they be referred to as (220)'s. Table A.2 lists a_0 for several common semiconductors as well as the separation of low indices planes for silicon and gallium arsenide.

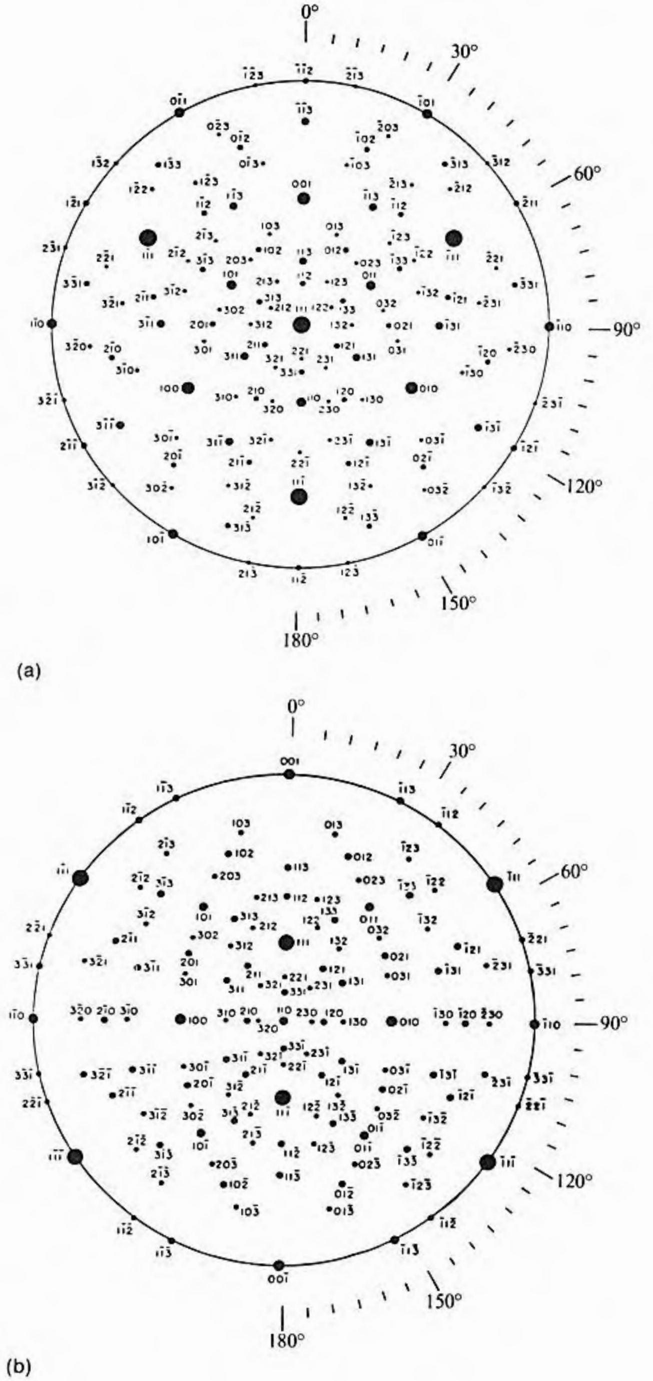


FIGURE A.7

111 and 110 stereographic projections.

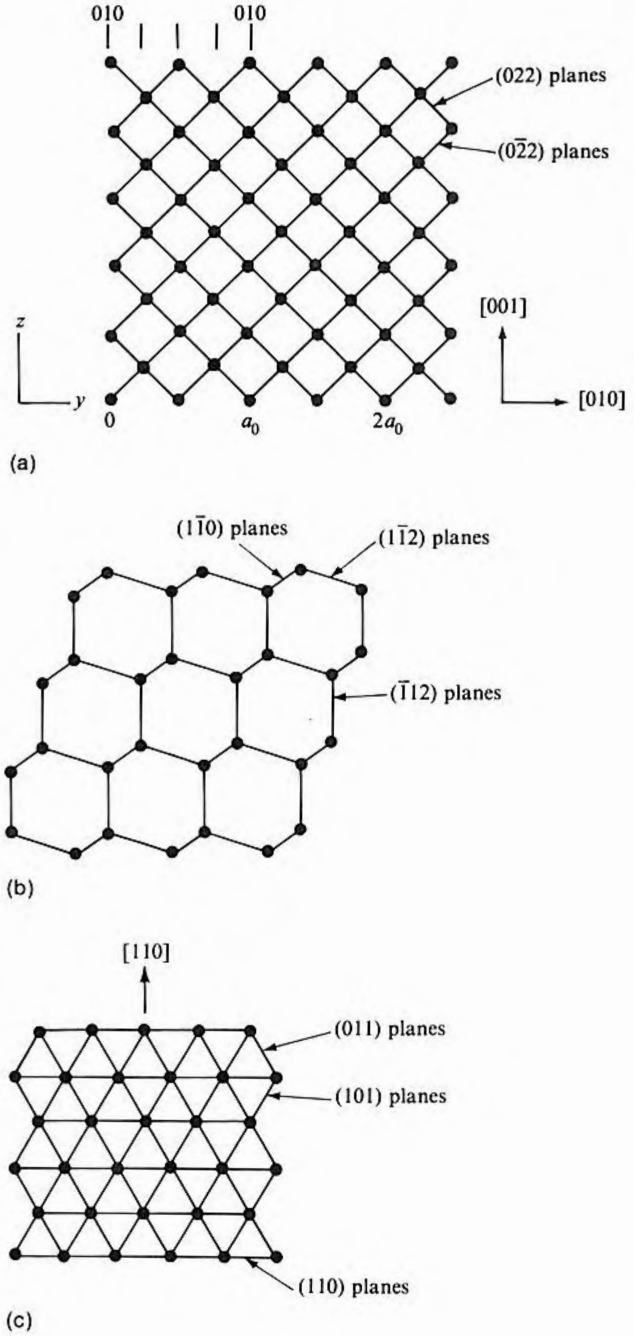


FIGURE A.8

Views of atoms and bonds when sighting into crystal in [100], [110], and [111] directions, respectively.

TABLE A.2

Lattice and Plane spacings (Å)

Material	a_0	*Plane*	*Spacing* *Si*	*Spacing* *GaAs*
Diamond	3.668	100	5.43	5.65
SiC	4.3596	400	1.36	1.41
Si	5.4307	110	3.84	4.00
GaP	5.4504	220	1.92	2.00
GaAs	5.6533	111	3.14	3.26
Ge	5.6575	444	0.784	0.815

Fig. A.8b is reproduced in Fig. A.9a, but dashed lines have been added to indicate the traces of the parallel sheets of atoms that make up (111) planes. The two closely spaced layers are separated substantially from the next set of double layers. Normally, "the (111) plane" is considered to include both sheets, and the separation of (111) planes as calculated from Table A.1 gives the distance from the middle of one double layer to the middle of the next one. From Fig. A.9, it can be seen that the distance is also given by the length of the bond plus the separation of the close spaced layers. This separation is 1/4 the (111) spacing. Fig. A.9b is the same as Fig A.3b except that parts of two planes have been shaded to show the double sheet just described. These two sheets are multiply bonded together, with three of the four bonds of each atom being used for that purpose. The fourth bond goes to the next set of double layers. It is this paucity of bonds that allows slip and cleavage to rather easily occur between adjacent silicon (111) planes. Table A.3 gives the number

FIGURE A.9

(a) A 110 section of diamond lattice showing location of two sheets of atoms comprising a (111) plane. (b) A perspective view with a portion of two pairs of close-spaced (111) planes shaded. [These show two of the four orientations of (111) planes of the lattice; the other two are perpendicular to the $[1\bar{1}\bar{1}]$ and $[\bar{1}1\bar{1}]$ directions.]

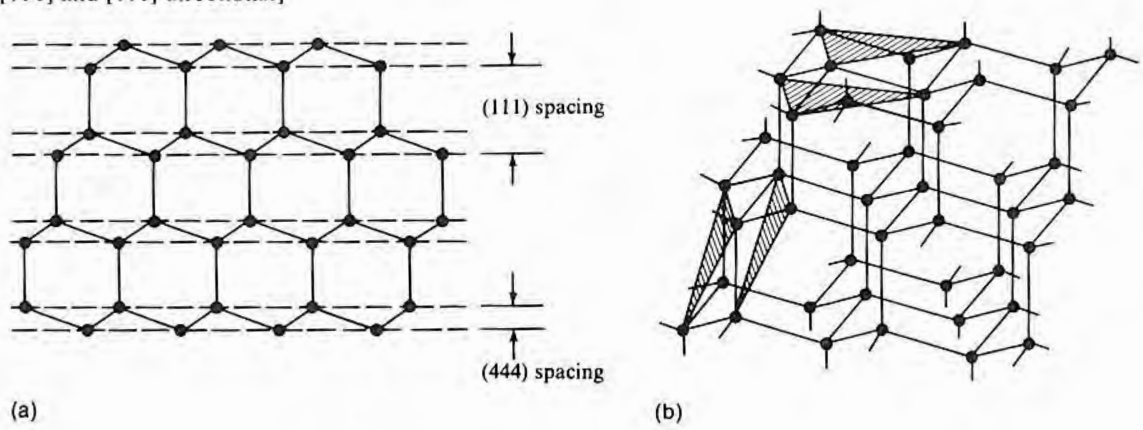


TABLE A.3

Plane-to-Plane Bond Density for Diamond and Zinc Blende Structures

Plane	*Bonds/Atom*	*Bonds/a_0^2*	*Bonds/cm²*	*Relative Bonds/ Unit Area*	*Bonds/cm² (Si)*
(111)*	1	$4/\sqrt{3}$	$4/(d^2\sqrt{3})$	1	7.9×10^{14}
(110)	1	$2\sqrt{2}$	$2\sqrt{2}/d^2$	1.22	9.6×10^{14}
(100)	2	4	$4/d^2$	1.73	1.4×10^{15}
(111)†	3	$12/\sqrt{3}$	$12/(d^2\sqrt{3})$	3	2.4×10^{15}

*Wide spacing.

†Close spacing.

Source: J.H. Brunton, "Fracture Propagation in Diamond," *Proc. 1st Int. Cong. on Diamonds in Industry*, Paris, 1962.

of bonds/atom that go from one plane to the next, as well as the number of bonds/cm².

The gallium arsenide double layer has one layer totally of gallium and the other totally of arsenic. Gallium arsenide shows some ionic bonding characteristics, and the attraction between the two layers provides enough strength to offset the low bond density and prevent easy cleavage between (111) planes. There is also only one bond per atom between (110) planes, in which case, with the adjacent planes each having equal numbers of gallium and arsenic atoms, ionic bonding is less, and cleavage occurs. The difference in composition of each layer of a gallium arsenide (111) double layer causes several properties to appear differently when this layer is approached from the [111] or the $[\bar{1}\bar{1}\bar{1}]$ direction. One example is etching, in which a (111) face behaves differently depending on whether the gallium or the arsenic layer is reached first.

A.2 Crystallographic Defects

The crystallographic defects that are likely to be encountered during wafer processing and that are readily observable are dislocations, slip, and stacking faults. In addition, twins are occasionally encountered. Defects that are not so apparent and whose presence must be inferred are vacancies, interstitials, and antistructure. An antistructural defect is an atom of the wrong kind at a lattice site and can only occur in compound semiconductors. Vacancies and interstitials, also sometimes referred to as either point or Schottky defects, are thermodynamically stable and will be present in all crystals. In addition, some process steps produce excess quantities of them. If a vacancy and an interstitial are adjacent (paired), they are called

Frenkel defects. Vacancies also sometimes pair with impurities in the crystal to form various complexes. Some of these will occur during high-temperature operations such as diffusion, but they are particularly likely to form during irradiation by high-energy neutrons, electrons, or protons.

A vacancy is an atom missing from a lattice site. Since the bonding is disrupted, vacancies can have a charge and, in silicon, are thought to be able to exist as neutrals (V^0), singly charged donors (V^+), singly charged acceptors (V^-), and doubly charged acceptors ($V^=$) (1). The relative amounts of each type will depend on the temperature and doping level. For silicon, the equation for the neutral equilibrium concentration at a temperature T (K) is of the form

$$N_V^0 = Ne^{-E_V^0/kT}$$

where E_V^0 is ~2.3 eV and N is the number of silicon atoms per unit volume (2). In the case of compound semiconductors, the number of vacancies for each constituent may be different. For gallium arsenide, the calculated concentrations, assuming equilibrium vapor pressure, are (3)

$$N_V^0 = (3.3 \times 10^{18})e^{-0.4/kT} \qquad \text{for gallium}$$

$$N_V^0 = (2.2 \times 10^{20})e^{-0.7/kT} \qquad \text{for arsenic}$$

The concentration of charged vacancies is related to the neutral concentration through the difference between the location of the charged vacancy energy level E_V and the Fermi level E_F in the bandgap. For silicon (4),

$$N_{V^+} = N_V^0 e^{(E_{V^+} - E_F)/kT}$$

$$N_{V^-} = N_V^0 e^{(E_F - E_{V^-})/kT}$$

$$N_{V^=} = N_V^0 e^{(2E_F - E_{V^-} - E_{V^=})/kT}$$

An interstitial is an extra atom residing in the space between the normal lattice sites. In the diamond and zinc blende lattices, there are five spaces in the interior of each unit cell and three centered on the unit cell walls. In each case, the location is vertically down half a space from the atoms shown in the plan (top) view of Fig. A.1. The interior interstitials are referred to as hexagonal sites; the other three, as tetragonal. Energy considerations make it appear that only the hexagonal sites are stable locations (5). In semiconductors with zinc blende structure, there are two further splits of interstitial sites, those whose nearest atomic neighbors are either group III elements or group V elements.

The most commonly discussed dislocations are screw and edge.

FIGURE A.10

(a) Screw dislocation. (b) Edge dislocation.

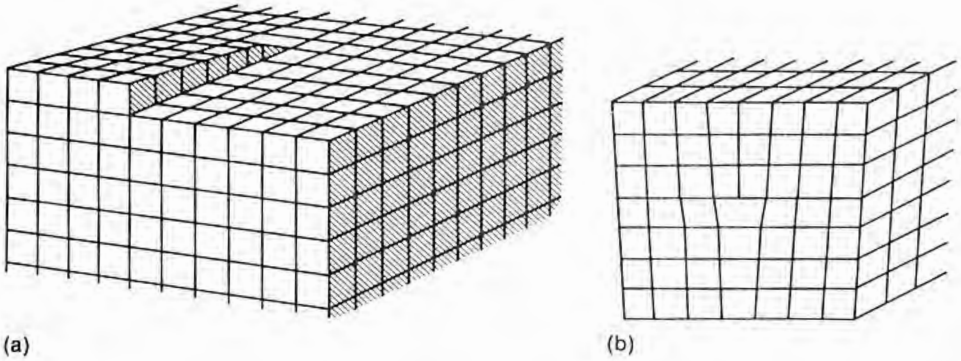


Fig. A.10a shows an idealized structure containing a screw dislocation. Edge dislocations are generally depicted as the disorder caused by an extra layer of atoms introduced into the crystal as shown in Fig. A.10b. This disorder will be at the edge of the extra plane, and crystal symmetry will dictate what orientation "extra" planes actually occur and the direction in which the edge of the plane will lie. In the case of silicon, various dislocation configurations are possible (6), some of which have both edge and screw characteristics. The one usually observed is the 60° dislocation. It is shown in Fig. A.11, is located at the edge of an extra (111) double layer, and runs in a [110] direction. Dislocations in semiconductors are generally formed, not by the physical introduction of an additional layer, but by a shearing deformation causing part of the lattice to slip relative to the rest, as indicated in Fig. A.12.

Misfit dislocations occur when two lattices that are not exactly the same size meet. Where only a small mismatch occurs, the bonds will distort and join; when the cumulative mismatch becomes on the order of the spacing of the planes, a whole plane on one side will fail to bond and will give the appearance of an extra half-plane as shown in Fig. A.13. The net result is a configuration that looks just like that of Fig. A.12b. The dislocations run in [110] directions, and it is (111) planes that appear to have been added, but the dislocations themselves are all confined to lie in the interface between the two lattices. Such dislocations are most commonly observed when a material is epitaxially overgrown onto a substrate with differing lattice constants. This occurs, for example, when high-resistivity silicon is grown onto a heavily doped substrate or when gallium arsenide is grown onto silicon.

FIGURE A.11

A 60° dislocation in a diamond lattice. [The extra half-plane is a (111) double layer. The dislocation at the end of the plane, indicated by the arrow, travels in a [110] direction.]
(*Source:* Reprinted with permission from R.G. Rhodes, *Imperfections and Active Centres in Semiconductors*, Copyright 1964, Pergamon Press plc.)

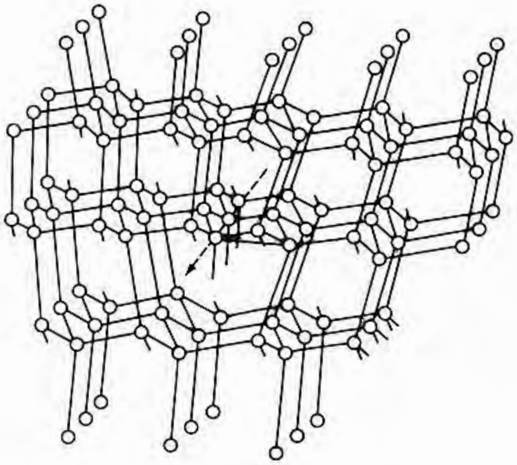


FIGURE A.12

Formation of an edge dislocation by mechanical deformation. (Application of stress causes bonds A–2, B–3, C–4, and so on to break and change to A–1, B–2, C–3, and so on.)

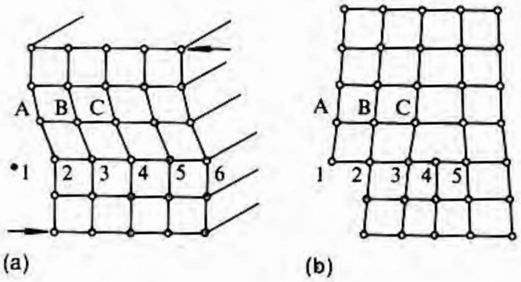
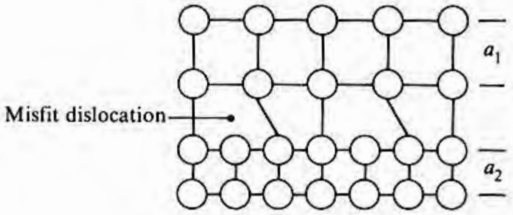


FIGURE A.13

Misfit dislocations caused by matching of two lattices with differing lattice spacings a_1 and a_2. (In an actual situation, the lattice spacings would be very close together, and the dislocations would be widely separated.)



Slip is the plastic deformation of a crystal by one part slipping across the other as shown in Fig. A.14. The plane along which the shear motion occurs is the slip plane and, for silicon, is the {111} family. If slip over the whole plane were to occur simultaneously and all of the motion were to be in exact atomic units, after slip, there would be no crystal defects. Unfortunately, this is not true; all

FIGURE A.14

Idealized slip.

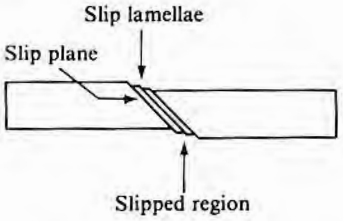
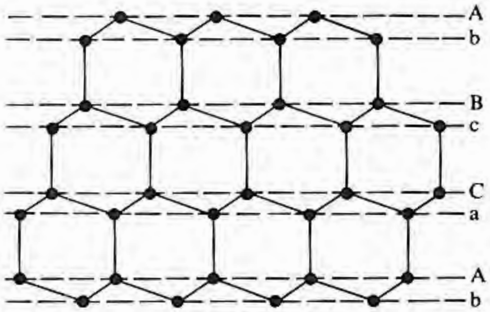


FIGURE A.15

A 110 section through a diamond lattice showing six different layers of atoms that are encountered in moving in a <111> direction.

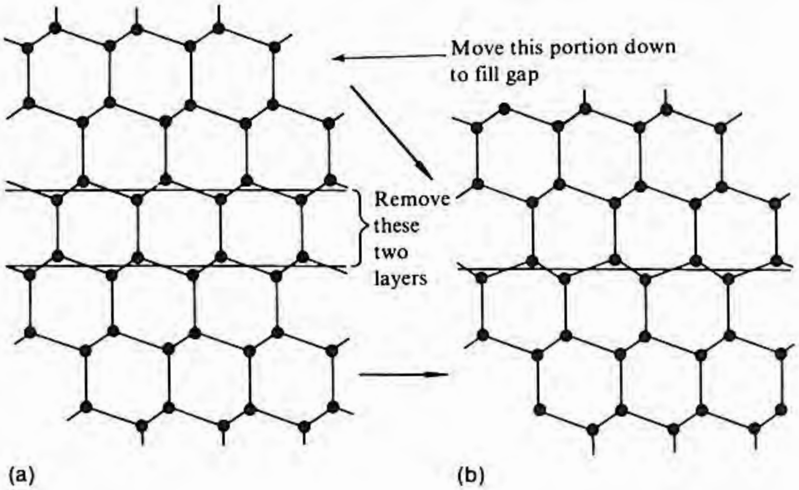


regions do not move simultaneously. The result is deformation in some places and slip in others. Thus, it appears that numerous extra planes, along with the dislocations at their boundaries, have been formed. Hence, slipped regions in a wafer usually cause rows of dislocations to intersect the wafer surface. It is the etch pits associated with these dislocations that are generally referred to as "slip lines." In some cases, slip steps can be detected at the wafer surface by microscopic observation with interference contrast or by scanning with a very sensitive surface profilometer.

The stacking sequence of layers in a <111> direction consists of three pairs of double layers (Ab, Bc, Ca) as shown in Fig. A.15. Each of the six layers in the sequence is different in either atomic placement or direction of bonds. Stacking faults are errors in the way in which these various layers are stacked. The result of the errors is that it appears that an additional layer either has been inserted or has been removed. However, unlike the dislocation, which has an additional double layer, a stacking fault involves the addition or removal of the bottom of one double layer and the top of the double layer below it—that is, bB, cC, or aA—which is necessary since these are the pairs that have atoms in the same positions. Fig. A.16a shows two planes marked for removal, and Fig. A.16b shows the result when the lattice collapses and closes the gap. One double layer is introduced with bonding that is different from the remainder of the crystal. The result is referred to as an intrinsic stacking fault. One way in which such a fault could occur is by the aggregation of

FIGURE A.16

A (110) section through a diamond lattice showing formation of an intrinsic stacking fault by removal of two adjacent layers of atoms.



a large number of vacancies and the subsequent collapse of the lattice. If an extra layer is inserted into the lattice, then two double layers are different, and the fault is extrinsic. They can grow by the capture of self-interstitials. When the extra or missing layers terminate within the crystal, dislocations will occur at the edges of the fault.

Twinning is a crystal defect characterized by a twin plane separating two crystal regions of different orientation. The atomic positions on each side of the plane are mirror images, and the nearest neighbors across the boundary are in the correct position. Crystal symmetry requirements restrict the possible twin-plane orientations as well as the orientation of the material on each side of the twin. In the case of a diamond lattice, (111) and (112) planes are allowed, but when detailed energy requirements are considered, only the (111) is expected (7) and is the only one that has been observed. Zinc blende, with its two components, can, in principle, have a twin with the nearest neighbors in the correct positions and the proper constituents at the locations, or else, while the positions are correct, the constituents can be reversed. Twinning most often occurs during crystal growth and will generally be screened out before reaching a wafer-fab facility.

APPENDIX REFERENCES A

1. G.D. Watkins, *Radiation Damage and Defects in Semiconductors*, p. 228, Institute of Physics, London, 1973.

2. R.G. Rhodes, *Imperfections and Active Centres in Semiconductors*, Pergamon Press, New York, 1964.

3. Sorab K. Ghandhi, *VLSI Fabrication Principles*, John Wiley & Sons, New York, 1983.

4. C.P. Ho and J.D. Plummer, "Si/SiO$_2$ Interface Oxidation Kinetics: A Physical Model for the Influence of High Substrate Doping Levels," *J. Electrochem. Soc. 126*, pp. 1516–1522, 1979.

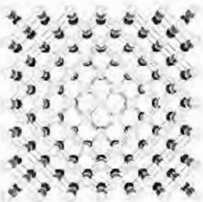
5. K. Weiser, "Theory of Diffusion and Equilibrium Position of Interstitial Impurities in the Diamond Lattice," *Phys. Rev. 126*, pp. 1427–1436, 1962.

6. J. Hornstra, "Dislocations in the Diamond Lattice," *J. Phys. Chem. Solids 5*, pp. 129–141, 1958.

7. E.I. Salkouitz and F.W. Von Batchelder, "Twinning in Silicon," *J. Metals 4*, p. 165, 1952.

APPENDIX

B



Phase Diagrams

B.1 BINARY PHASE DIAGRAMS

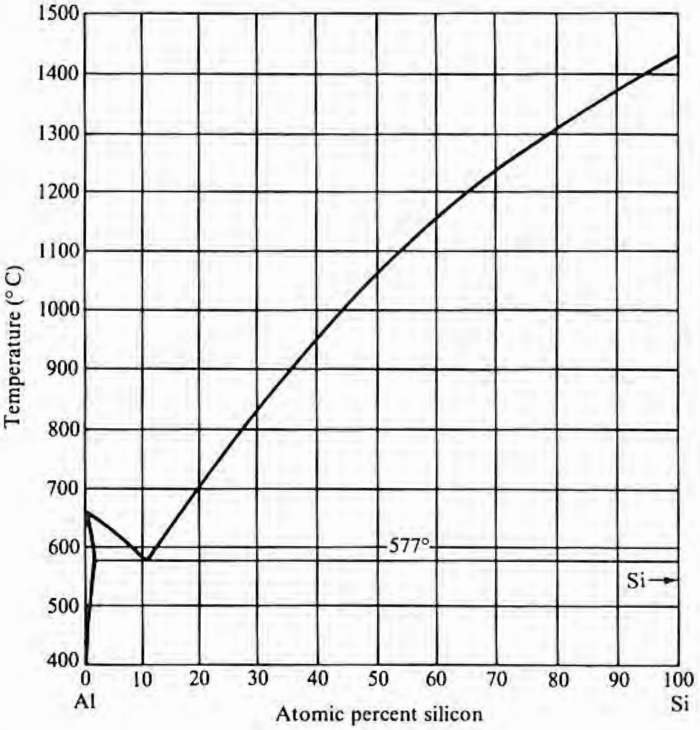
Phase diagrams are very helpful in compactly showing the melting and freezing behavior of mixtures of two or more components. They are a series of curves, generally at constant pressure, that map out the solid and liquid regions as a function of temperature and composition.[1] All of the phase diagrams must be experimentally determined, and the details for a large portion of the possible two-element combinations have now been published (1). An example of a binary phase diagram is shown in Fig. B.1. It is for the Al–Si mixture and shows composition along the *x* axis, ranging from 100% Al on the left to 100% Si on the right. Temperature is on the *y* axis, and for this diagram (and most others), the pressure is 1 atm. This curve is one of the relatively simple binary eutectic diagrams, which include combinations such as Sb–Si, Be–Si, Ga–Si, Au–Si, and Ag–Si. The three main classes of binary phase diagrams are isomorphous, eutectic, and peritectic.

The simplest of the binary systems is the isomorphous, which only occurs when the two components *A* and *B* are mutually soluble in all proportions, both as a liquid and as a solid. Not many combinations meet these requirements, but Ge–Si and Se–Te are two examples that do. The isomorphous curve of components *A* and *B* is shown in Fig. B.2. Above the top line (liquidus line), all of the material is melted. Below the bottom line (solidus line), all of the material is frozen. In the region between the two curves, there will be both liquid and solid material. The behavior of a frozen (solid) isomorphous alloy of composition *X* is as follows: As the temperature is increased, melting will begin when the temperature reaches T_1.

[1] Despite the fact that the data must be obtained experimentally, the phase diagram concept is based on thermodynamical principles, and some general rules are available to help in shaping the curves. For more details, the reader should consult a standard text on phase diagrams.

FIGURE B.1

Al–Si phase diagram. (*Source:* Max Hansen and Kurt Anderko, *Constitution of Binary Alloys*, © 1958 by the McGraw-Hill Book Co., New York. Used with permission.)



The composition of the melt (composition Z) is given by the point at which the horizontal constant temperature line (tie line) intersects the liquidus curve. (The tie line gives the composition of the two phases that can co-exist in equilibrium at a given temperature and pressure.) As the temperature is raised the tie line moves upward, indicating a change in the composition of both the melt and the remaining solid. Finally, at T_3, the solid disappears, and the alloy is completely melted. In cooling, the reverse occurs. All of the material remains molten until temperature T_3, at which point material with composition Y begins to freeze out. As the temperature decreases, the composition of the solid gradually changes[2] until temperature T_1 and composition X is reached, at which point the melt is gone, and the material is completely frozen.

Fig. B.3 shows the classical eutectic phase diagram. The region

[2] It must be noted that under equilibrium conditions, it is not just the material melting or just frozen that has a composition determined by the intersection of the tie line with the solidus curve. *All* of the solid material must have that composition. That is, the melting or freezing process must proceed slowly enough for solid-state diffusion to maintain the equilibrium composition of the solid.

FIGURE B.2

Isomorphous phase diagram.

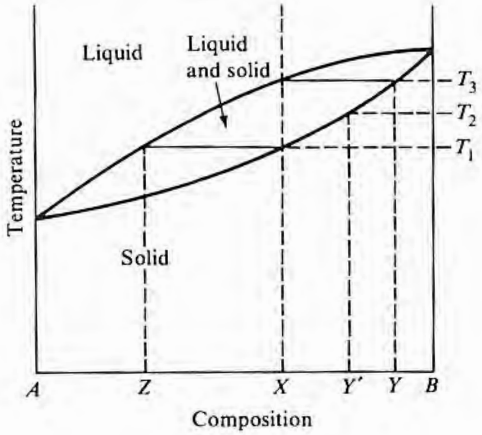
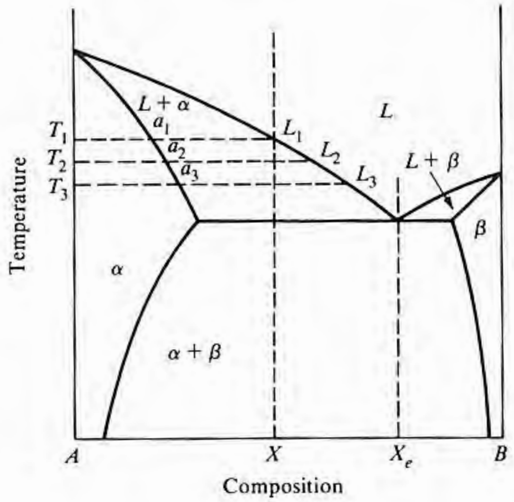


FIGURE B.3

Eutectic phase diagram.



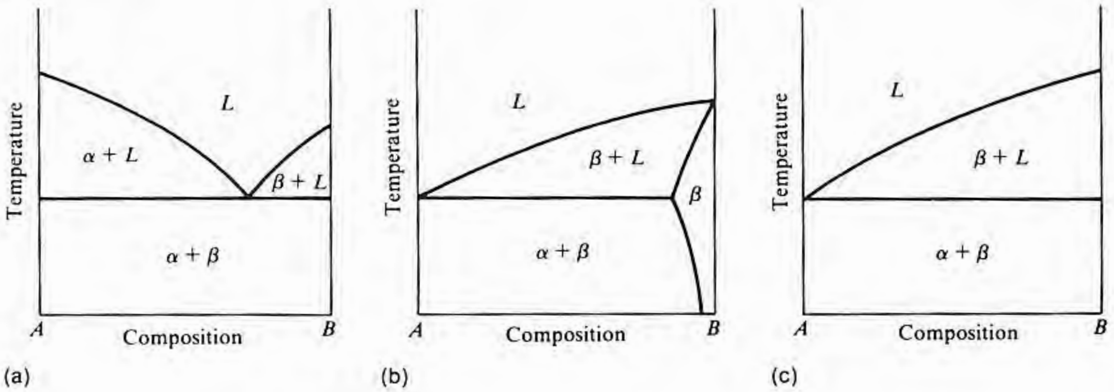
labeled α encloses the α solid phase, which is comprised of component A saturated with component B. The maximum amount of B at any temperature that can be in the α phase is read from the curve that separates the α region from the liquid or the α + β region. Similarly, the β region is composed of component B saturated with component A. If the temperature for a given composition is above the top curve, all material will be molten. The minimum in the liquidus line is the eutectic point, and the composition X_e will have the lowest melting point for the alloy A–B. Further, all material of this composition will melt and freeze isothermally at the eutectic temperature, producing a mixture of α and β. Thus, at the eutectic point, as the temperature is lowered, there is a change from a liquid phase to two solid phases. If a composition X is not at X_e, then the melting and freezing behavior is much like that of the isomorphous system. That is, for X located as in Fig. B.3, as the temperature drops to T_1, the α phase will begin freezing and will have a composition α_1. As the temperature continues to decrease, the tie line will move down, indicating that the composition of the α phase is changing to α_2 to α_3 while the remaining liquid composition changes from L_1 to L_2 to L_3. When the horizontal line (eutectic line) is reached, the remaining melt has the eutectic composition and freezes isothermally to produce a mixture of α and β. Were the initial composition point X to lie to the right of point X_e, then the behavior would be the same except that the β phase would freeze out first.

For many alloy constituents, there will be so little solubility of A in B and/or B in A that the α and β regions will appear only as vertical lines on the diagram, as shown in Fig. B.4a. The Al–Si diagram is an example of this behavior in that the aluminum solubility in silicon is so low that it does not show[3] (Fig. B.1), although there is substantial silicon solubility in aluminum. Another example is Au–Si. Cases also exist in which the eutectic composition is so close to the lower melting point component that the eutectic minimum does not show. Then, the curve looks like either Fig. B.4b or B.4c. An example is Sb–Si.

There are also two other phase diagrams in the eutectic class, the eutectoid and the monotectic. The eutectoid diagram, shown in Fig. B.5a, looks like the eutectic except that all phases are solid. At higher temperatures, some sort of transition from the γ phase to liquid will occur. There do not appear to be any examples with silicon as a constituent. Cr–Ni is thought to show such behavior. The

FIGURE B.4

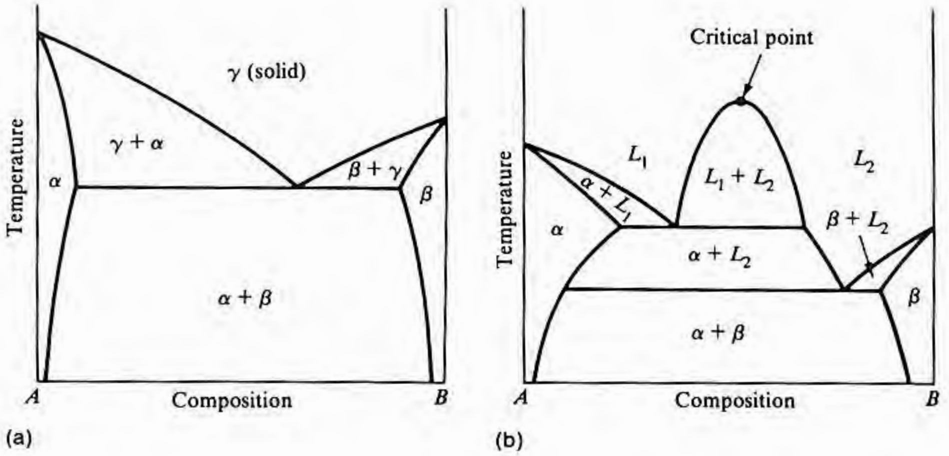
Effect of terminal solid solubility range on eutectic phase diagram.



[3] Curves for the solubility of various elements in silicon are given in Chapter 8.

FIGURE B.5

(A) Eutectoid and (b) monotectic phase diagrams.



monotectic diagram is shown in Fig. B.5b. At the monotectic composition, if the temperature is high enough, one liquid phase exists, but as the temperature is decreased, it separates into two liquid phases. As the temperature is dropped further, a solid phase forms, and one liquid phase remains. Further cooling gives a mixture of two solid phases. Bi–Si and Pb–Si behave in this manner.

A peritectic transformation occurs if, as the temperature rises, a solid phase transforms into a liquid and a new solid phase. Fig. B.6a shows a peritectic phase diagram. If the composition is at the peritectic value, decreasing temperature causes the α phase to freeze out until the peritectic line is reached (T_p), at which time the liquid and α combine to form the β phase. Such reactions are generally extremely slow since the first β to form will create a diffusion barrier that partially isolates the α phase from the remaining liquid. Often, the phase formed during the peritectic reaction will not involve one of the terminal phases but rather a compound of the two terminal phases. In this case, the curve might appear as in Fig. B.6b. This figure also illustrates a feature exhibited by many phase diagrams, which is that the overall phase diagram may be a composite incorporating different reactions in different compositional ranges. In this particular case, for compositions with *B* less than *Y*, the system behaves as peritectic. However, if there is more *B* than *Y*, the system is eutectic. *X* and *Z* are, respectively, the peritectic and the eutectic compositions. In the event that there is little compositional range for each solid phase (α, β, γ), the diagram reduces to that of

Fig. B.6c, which is more typical of those with silicon as a constituent. If the peritectic reaction occurs very close to one of the terminal phases, the curve can look like that of Fig. B.4b or B.4c. In this circumstance, without further information, the difference between a peritectic and a eutectic diagram cannot be ascertained. Silicon alloys that show peritectic behavior are, for example, As-Si, Ca–Si, and Mo–Si.

The peritectoid and syntectic systems of the peritectic class are analogous to the eutectoid and monotectic systems of the eutectic class. The peritectoid diagram looks like a peritectic diagram except that all phases are solid. In a syntectic reaction, as the temperature increases, instead of a single solid phase separating into a liquid phase and a different solid phase, it separates into two liquid phases.

Sometimes, a combination of the two terminal components will form a single solid phase that will melt to give only one phase. This behavior is distinctly different from either the eutectic system (two

FIGURE B.6

(a) Peritectic phase diagram. (b) Phase diagram showing both peritectic and eutectic behavior. (c) Effect of narrow terminal phases.

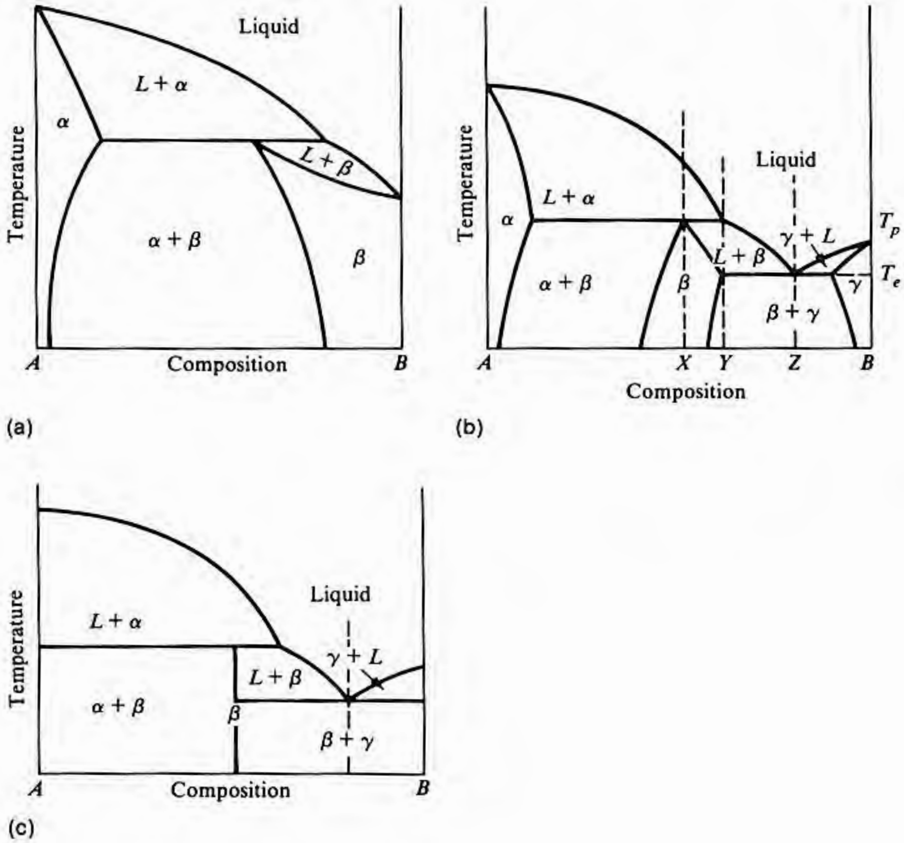
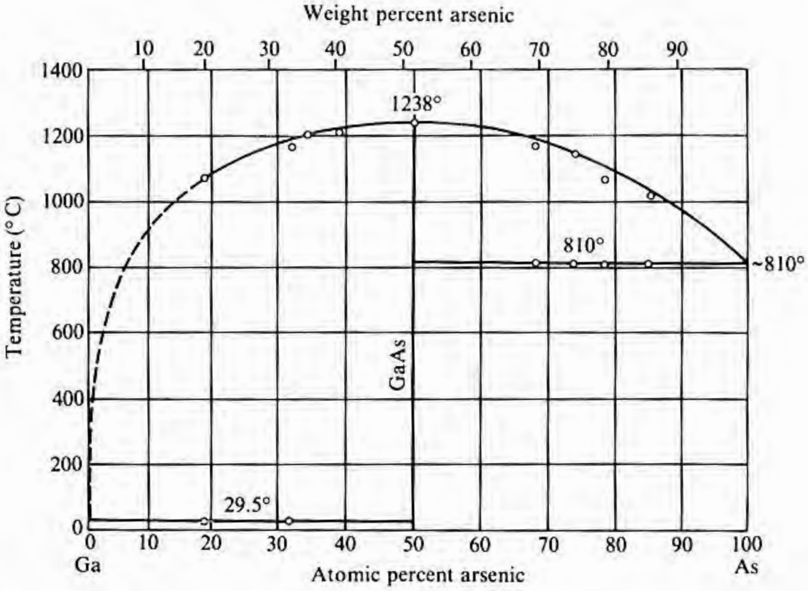


FIGURE B.7

Ga–As phase diagram. (*Source:* Max Hansen and Kurt Anderko, *Constitution of Binary Alloys*, © 1958 by the McGraw-Hill Book Co., New York. Used with permission.)



solid phases melt to give a liquid) or the peritectic system (one solid phase melts to give another solid phase plus a liquid) and is referred to as congruent melting. Such phases are sometimes called intermetallics and include compound semiconductors such as gallium arsenide, gallium phosphide, and indium antimonide. Also included are many of the silicides discussed in Chapter 10, although some of them apparently undergo peritectic (incongruent) melting, which includes, for example, Pt_5Si_2. When a congruently melting phase is formed, it divides the phase diagram into independent sections as is shown, for example, in the As–Ga diagram of Fig. B.7. The diagram is divided into two sections, each of which looks like Fig. B.4c.

B.2 TERNARY PHASE DIAGRAMS

When three elements are involved, such as gallium, arsenic, and phosphorus, the phase diagrams are referred to as ternary and become much more complex. An equilateral triangle is used as the base of a three-dimensional composition–temperature diagram as shown in Fig. B.8. The three corners of the triangle (Gibbs triangle) are labeled, respectively, as component *A*, component *B*, and component *C*. Along each leg, the percent composition scale goes from 0 to 100%. Each of the three sides of the triangular cross-sectional prism of Fig. B.8 depicts a conventional binary phase diagram of the components *A–B*, *B–C*, and *C–A*. In this example, each pair of components forms a simple eutectic of the form shown in Fig. B.4a. By taking a section parallel to the base at a given temperature, an iso-

FIGURE B.8

Three-dimensional view of a ternary phase diagram.

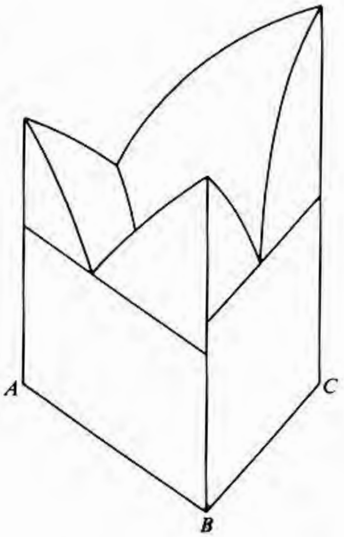
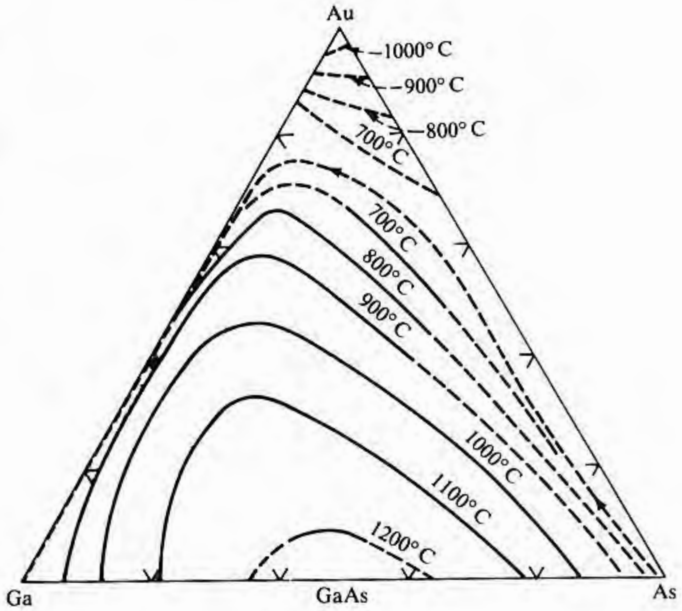


FIGURE B.9

Isothermal section of Ga–As–Au ternary phase diagram. (Because this diagram is a particularly simple one, there are only one or two curves per temperature section. Thus, curves for several temperatures can be combined on one section.) (*Source:* Adapted from M.B. Panish, *J. Electrochem. Soc. 114*, p. 516, 1967.)



thermal section involving all three components is obtained as shown in Fig. B.9. This particular section is not from the diagram of Fig. B.8, but rather is of Ga–As–Au (2). To follow the behavior as a function of temperature, it is then necessary to have a series of sections covering the temperature range of interest. In this figure, the intersection of the solidus lines with the plane is shown for the plane

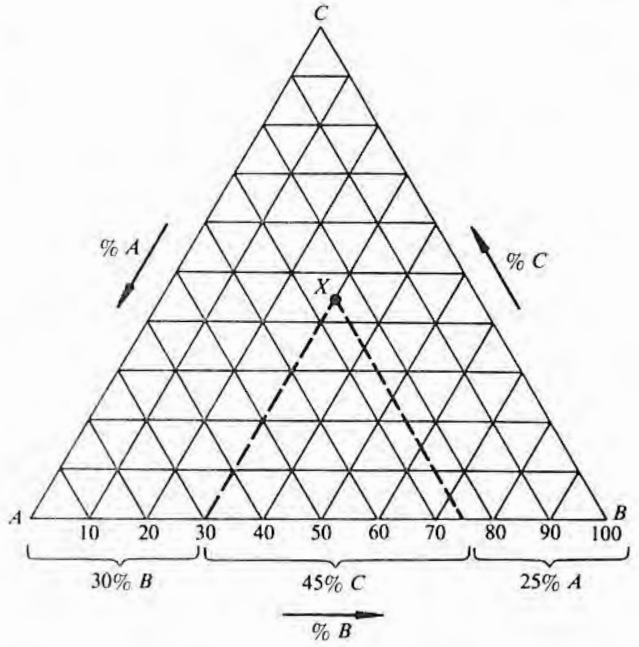


FIGURE B.10

Determination of composition from triangular scale. (Normally, the interior lines will not be drawn on the diagram. This scale is shown moving counterclockwise. Some diagrams use a clockwise scale.)

located at several different temperatures. Thus, one figure suffices to show several different temperature sections. The composition of the material at any point X in the triangle is read from the scales as shown in Fig. B.10. Draw a line through X parallel to the triangle leg opposite apex A. The intersection of the line with either of the other two legs gives the amount of A. Use a similar construction to find the amount of B. C can either be read as shown in the figure or calculated from $C = 100 - A - B$.

Since, as was discussed earlier, a congruently melting compound can be used as a terminal phase of a phase diagram, it is not unusual, particularly in the case of oxides, to find, for example, an SiO_2–P_2O_5 diagram. This concept can also be extended to ternary phase diagrams; SiO_2–P_2O_5–B_2O_3 is an example (3). The result is that some four-component, such as Si–P–B–O, systems can be described with the procedures normally used for ternary materials.

APPENDIX B REFERENCES

1. Max Hansen and Kurt Anderko, *Constitution of Binary Alloys*, 2d ed., McGraw-Hill Book Co., New York, 1958.
2. Ernest M. Levin, Carl R. Robbins, and Howard F. McMurdie, *Phase Diagrams for Ceramists*, The American Ceramic Society, Columbus, Ohio, 1964.
3. M.B. Panish, "Condensed Phase Systems of Gallium and Arsenic with Group IB Elements," *J. Electrochem Soc. 114*, p. 516, 1967.

APPENDIX C

Reference Tables

TABLE C.1

Numerical Values of Useful Constants

Quantity	*Symbol*	*Value*
Avogadro's number	n	6.023×10^{23} atoms per gram molecular weight
Universal gas constant	R	8.31×10^{7} ergs/K·mol 1.99×10^{-3} kcal/K·mol
Standard volume of perfect gas	V	22.41 liters/mol
Boltzmann's constant (R/n)	k	8.62×10^{-5} eV/K 1.38×10^{-16} ergs/K 1.38×10^{-23} J/K
Electron charge	q	1.6×10^{-19} C
Electron rest mass	m_o	9.108×10^{-28} gram
Planck's constant	h	4.14×10^{-15} eV·s 6.62×10^{-27} erg·s
Permittivity of free space	ε_0	8.86×10^{-14} F/cm
Permeability of free space	μ_0	1.26×10^{-8} H/cm
Velocity of light	c	3×10^{10} cm/s

TABLE C.2

Length, Energy, and Pressure Conversions

Quantity		*Value*
Length		
1 Å	=	10^{-8} cm
1 nm	=	10^{-7} cm
1 μm (micron)	=	10^{-4} cm
Energy		
1 eV	=	1.602×10^{-19} J
	=	1.602×10^{-12} ergs
	=	3.832×10^{-20} cal

TABLE C.2 (*continued*)

Length, Energy, and Pressure Conversions

Quantity		*Value*
1 e*V*	=	23.05 kcal/mol
1 J	=	10^7 ergs
1 gram-cal	=	4.181 J (at 20°C)
1 W·s	=	1 J = 0.24 cal
Pressure		
1 atm	=	1.013×10^6 dynes/cm^2
	=	14.7 lb/in.^{2}
	=	760 mm Hg
1 mm Hg	=	1333 dynes/cm^2
	=	133.3 Pa
1 dyne/cm^2	=	7.502×10^{-4} mm Hg
1 Pa (pascal)	=	10 dynes/cm^2
	=	1 N/m^2
	=	0.0075 mm Hg
	=	0.0075 torr
1 torr	=	1 mm Hg
	=	133.3 Pa
1 bar	=	10^6 dynes/cm^2
1 kPa	=	10^4 dynes/cm^2
1 MPa	=	10^7 dynes/cm^2
1 GPa	=	10^{10} dynes/cm^2
1 kg/mm^2	=	9.8×10^7 dynes/cm^2
1 lb/in.^{2}	=	6.89×10^4 dynes/cm^2

TABLE C.3

Greek Alphabet

	Lowercase Letter	*Uppercase Letter*		*Lowercase Letter*	*Uppercase Letter*
Alpha	α	A	Nu	ν	N
Beta	β	B	Xi	ξ	Ξ
Gamma	γ	Γ	Omicron	ο	O
Delta	δ	Δ	Pi	π	Π
Epsilon	ε	E	Rho	ρ	P
Zeta	ζ	Z	Sigma	σ	Σ
Eta	η	H	Tau	τ	T
Theta	θ	Θ	Upsilon	υ	Υ
Iota	ι	I	Phi	φ	Φ
Kappa	κ	K	Chi	χ	X
Lambda	λ	Λ	Psi	ψ	Ψ
Mu	μ	M	Omega	ω	Ω

TABLE C.4

International System of Units

Quantity	*Unit*	*Symbol*
Length	meter	m
Mass	kilogram	kg
Time	second	s
Temperature	kelvin	K
Current	ampere	A
Frequency	hertz	Hz (l/s)
Force	newton	N (kg·m/s^2)
Pressure	pascal	Pa (N/m^2)
Energy	joule	J (N·m)
Power	watt	W (J/s)
Electric charge	coulomb	C (A·s)
Potential	volt	V (J/C)
Conductance	siemens	S (A/V)
Resistance	ohm	Ω (V/A)
Capacitance	farad	F (C/V)
Magnetic flux	weber	Wb (V·s)
Magnetic induction	tesla	T (Wb/cm^2)
Inductance	henry	H (Wb/A)

Index

Semic
Integra
Proces

W. R. Run
Retired, fo

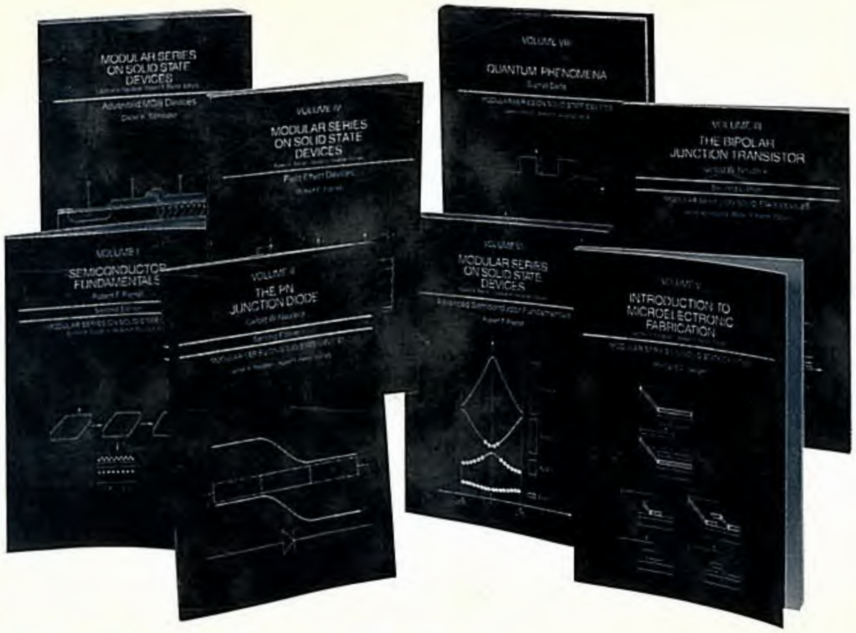
K. E. Bea
T.I. Fellow,

This boo
that portion
ing process
fabrication.'
depth offere
reference a

The boo
and advan
a broader
manufactu
industrial v
industry, m
lems, what
The autho
key materi
the proces
circuits for
circuit per

Walter F
familiar wi
circuits we
their own
personal
been acti
ment of n
of proces
The histo
included
an import
present p

The bo
detail tha
these top
in today's
in most o
their disc
defect as


MODULAR SERIES ON SOLID STATE DEVICES
Advanced MOS Devices
VOLUME VIII
QUANTUM PHENOMENA
VOLUME IV
MODULAR SERIES ON SOLID STATE DEVICES
Field Effect Devices
VOLUME III
THE BIPOLAR JUNCTION TRANSISTOR
VOLUME I
SEMICONDUCTOR FUNDAMENTALS
VOLUME II
THE PN JUNCTION DIODE
MODULAR SERIES ON SOLID STATE DEVICES
Advanced Semiconductor Fundamentals
VOLUME V
INTRODUCTION TO MICROELECTRONIC FABRICATION

Modular Series on Solid State Devices

R.F. Pierret and G.W. Neudeck, Editors, both of Purdue University

The Addison-Wesley Modular Series on Solid State Devices is a set of correlated but essentially self-contained volumes, each devoted to a specific device, a family of devices, or a coherent unit of device-related information. Emphasis in the first four volumes is on developing a fundamental understanding of the internal workings of the most basic solid state device structures. Subsequent volumes are intended to enhance the depth and breadth of understanding on specific topics.

Each volume contains numerous class-tested exercises that range in difficulty from easy to challenging. The modular nature of the series permits the volumes to be used sequentially as course texts or as tools of reference, review, or home study for full-time students, practicing engineers, and scientists.

90000
9 780201 108316
ISBN 0-201-10831-3

Electrical and Computer Engineering
Addison-Wesley Publishing Company