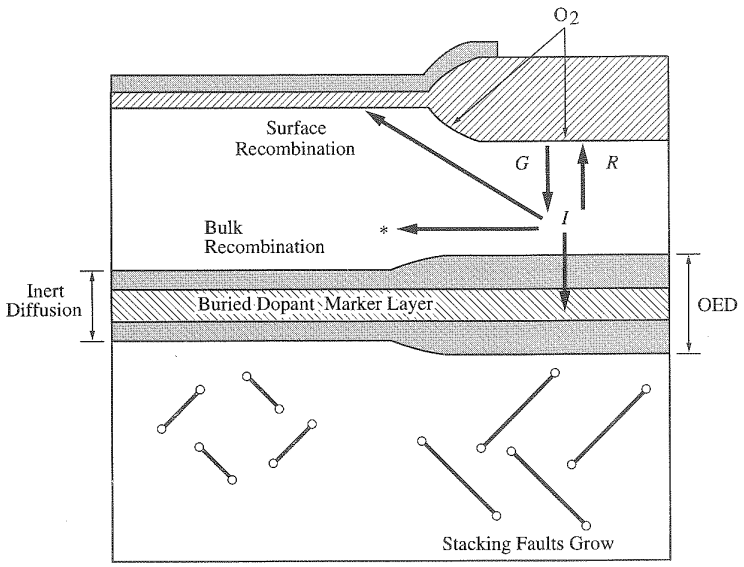


Figure 7-37 Generalized representation of point-defect generation (G), recombination (R), and diffusion processes in silicon. In this example, local oxidation (right side) generates interstitials which diffuse away from the Si/SiO_2 interface, locally enhancing the diffusion of dopants like B or P, while causing interstitial type stacking faults to grow.

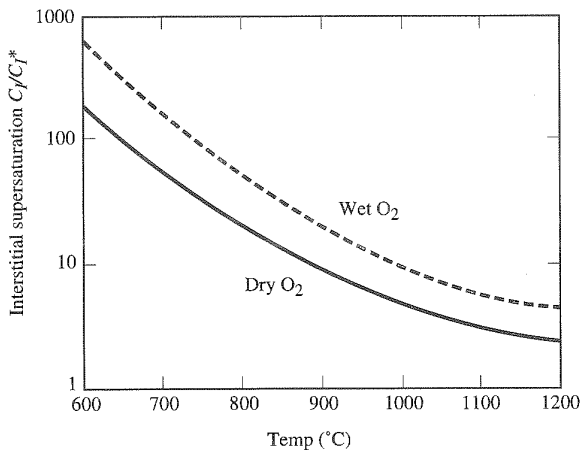


Figure 7-38 TSUPREM IV [7.14] simulations of the interstitial supersaturation generated by oxidation in wet and dry O_2 at various temperatures. The oxidation time is chosen to be at the transition between the linear and parabolic regimes.

This equation is a specific version of the more general forms given in Eqs. (6.48) and (6.49). The interstitial supersaturation results from the balance between the generation rate and the recombination rate of interstitials at the oxidizing interface. The interstitial excess above the equilibrium value (or supersaturation) depends primarily on temper-

TSMC-1015
U.S. Patent 8,796,779

ature and is larger at lower oxidation temperatures, as shown in Figure 7–38. The enhancement in diffusion coefficient is small at high temperatures but can be very large at low temperatures.

7.5.9 Dopant Diffusion Occurs by Both I and V

The idea that both I and V contribute to dopant diffusion in silicon is now generally accepted based on both experimental observations and theoretical calculations. The theoretical calculations have only recently become possible based on a direct quantum mechanical description of solid-state systems on powerful computers. The experimental results that dramatically capture the existence of a dual diffusion mechanism are based on the observation that under certain identical conditions such as oxidation of the silicon surface, one dopant has its diffusion enhanced while another dopant has its diffusion retarded.

For example Figure 7–39 shows the diffusion of arsenic and antimony (both n-type dopants) during oxidation of the silicon surface. The diffusion of the arsenic is enhanced while, simultaneously, the diffusion of the antimony is retarded below the normal values that would be observed if the anneal were carried out in an inert ambient such as argon. In this particular experiment, the retarded diffusion of the antimony is observed even inside the arsenic layer which is undergoing enhanced diffusion. Thus, by perturbing the point-defect populations by oxidation, the existence of two different fundamental mechanisms of diffusion is demonstrated.

It is important to note that this information could not be obtained by observing the diffusion of the dopants in an inert (argon or nitrogen) atmosphere. In that case, the only observation that could be made is that one dopant diffused faster or slower than another one. Instead, it is because the point-defect levels are perturbed from their equilibrium values by oxidation that some inference can be made about dopant diffusion mechanisms.

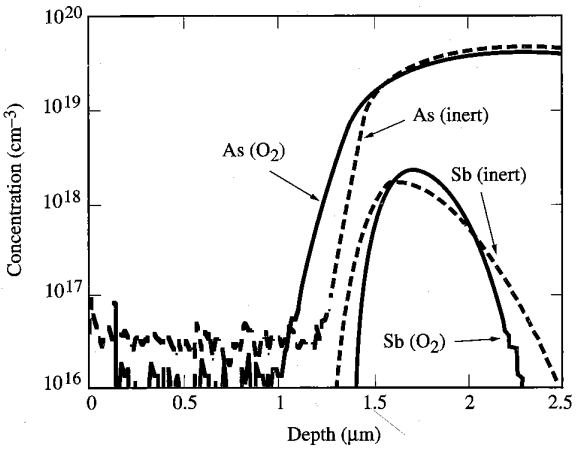


Figure 7–39 SIMS plot showing arsenic diffusion enhanced during oxidation at the same time that antimony diffusion is retarded below normal values. This is a strong indication that two different diffusion mechanisms are taking place. (After [7.23].)

It has also been found that nitridation of the silicon surface in an ammonia ambient has exactly the opposite effects to oxidation. In other words, B and P exhibit retarded diffusion, while the diffusion of Sb is enhanced. Preexisting stacking faults also shrink in a nitriding ambient. Thus the conclusion is that nitridation injects vacancies. This complementary set of observations "weaves a seamless logic" for the dual roles of interstitials and vacancies in assisting dopant diffusion in silicon [7.24, 7.25].

These observations can be formalized by saying that dopants diffuse with a fraction f_I interstitial-type diffusion mechanism and with a fraction $f_V = 1 - f_I$ vacancy-type mechanism. Under perturbed conditions, when the interstitial (C_I) or vacancy (C_V) concentrations are different from their equilibrium concentrations (C_I^*, C_V^*), the diffusivity of a dopant can be written

$$D_A^{\text{eff}} = D_A^* \left(f_I \frac{C_I}{C_I^*} + f_V \frac{C_V}{C_V^*} \right) \tag{7.65}$$

where D_A^{eff} is the effective diffusivity of the dopant measured under conditions where the point-defect populations are perturbed and D_A^* is the normal, equilibrium diffusivity measured under inert conditions and by definition $f_I + f_V = 1$.

It is interesting to note that if two different diffusion mechanisms are operative (i.e., interstitial- and vacancy-assisted mechanisms) and if they have very different activation energies (Q_A), then a plot of D_A versus $1/T$ would show distinct regions with different slopes. For dopants in silicon, there has never been such an observation in the intrinsic Q_A versus $1/T$ plot (Figure 7–15). We might expect that this slope change or curvature would be most obvious for arsenic, which has approximately equal I and V components of diffusion. This indicates that the activation energies of diffusion via both I and V mechanisms are comparable at processing temperatures.

It is the observation of retarded diffusion that allows the extent of diffusion by an I or V mechanism to be quantified. What is measured experimentally is how much the dopant is retarded compared with its normal diffusion in an inert ambient, that is, the extent of retardation for a dopant whose diffusion is retarded (superscript R) is given by

$$\left(\frac{D}{D^*} \right)^R = f_I^R \frac{C_I}{C_I^*} + f_V^R \frac{C_V}{C_V^*} = (1 - f_V^R) \frac{C_I}{C_I^*} + f_V^R \frac{C_V}{C_V^*} \tag{7.66}$$

A very conservative assumption under conditions of interstitial excess and a resulting vacancy deficit is that the vacancy component of diffusion is reduced to zero ($C_V/C_V^* = 0$) so that

$$\left(\frac{D}{D^*} \right)^R \geq (1 - f_V^R) \frac{C_I}{C_I^*} \tag{7.67}$$

An even more conservative assumption under conditions of interstitial excess is that $C_I/C_I^* = 1$ so that

$$f_V^R > 1 - \left(\frac{D}{D^*} \right)^R \tag{7.68}$$

This provides a mathematical justification for the statement that a dopant whose diffusivity is reduced to $^1/_{10}$ of normal when there is an interstitial excess present must diffuse at least $^9/_{10}$ by a vacancy mechanism. This is an absolute lower bound on the extent of diffusion by a V-type mechanism.

If another dopant is enhanced by some factor under the same conditions of an interstitial excess, then for the dopant that is enhanced (superscript E)

$$\left(\frac{D^{\text{eff}}}{D^*}\right)^E = f_I^E \frac{C_I}{C_I^*} + f_V^E \frac{C_V}{C_V^*} \tag{7.69}$$

Since the interstitial fraction of the dopant diffusion $f_I^E \leq 1$ and vacancy-assisted component of diffusion $f_V^E\,(C_V/C_V^*) \geq 0$, a conservative estimate is that

$$\left(\frac{D}{D^*}\right)^E \leq \frac{C_I}{C_I^*} \tag{7.70}$$

Simply stated, the measured enhancement in the diffusivity of a dopant won't exceed the interstitial supersaturation value. By substituting for C_I/C_I^*, Eq. (7.67) can now be used to show that the vacancy component of the dopant whose diffusivity is retarded can be better approximated by

$$f_V^R \geq 1 - \frac{\left(\frac{D}{D^*}\right)^R}{\left(\frac{D}{D^*}\right)^E} \tag{7.71}$$

An even stricter bound on the preference for a particular microscopic diffusion mechanism is possible if the interstitials and vacancies recombine to reach a steady-state condition given by

$$C_I C_V = C_I^* C_V^* \tag{7.72}$$

In this case, a strict lower bound on f_V can be derived as (see problem 7.19)

$$f_V^R \geq 0.5 + 0.5\sqrt{1 - \left[\left(\frac{D}{D^*}\right)^R\right]^2} \tag{7.73}$$

Thus by observing how much a dopant is enhanced or retarded under different conditions where the point defects are perturbed, an estimate of their propensity to diffuse via a vacancy- or interstitial-assisted mechanism can be made.

If an independent measure of the vacancy or interstitial supersaturation can be made, the fraction of diffusion by a vacancy or interstitial mechanism can be directly computed from Eq. (7.66). Such an estimate of the supersaturation might be made by observing the growth rate of stacking faults, for example. However, these methods tend to be less reliable than the self-consistent approach which relies on observing enhanced

and retarded dopant diffusion in the same experiment as described above. Remember that this problem of inferring the mechanism of dopant diffusion occurs because there is no reliable way to directly observe the interstitial or vacancy populations themselves at processing temperatures because of their low concentrations. In spite of this limitation, a self-consistent picture of the atomic-scale mechanisms of dopant diffusion in silicon can be deduced by following the trail of physical evidence. These kinds of experiments have been widely done, resulting in the f_I and f_V values for the common dopants in silicon shown in Table 7–6.

We have now "split" the overall dopant diffusivity into components dominated by interstitial- and vacancy-assisted mechanisms. Though this contains a lot more insight than a purely macroscopic description of the diffusion process, it is still a crude description of the atomic-level mechanisms. Consider that the dopant may diffuse much faster with the interstitial component than the vacancy component. Or that the activation energy for diffusion with interstitials might be different from the activation energy for diffusion with vacancies. How are these physical effects captured in our description of the diffusion process?

As described previously, the effective diffusivity of a dopant at low concentration in silicon is given by

$$D_A^{\text{eff}} = D_A^0 \exp\left(\frac{-E_A}{kT}\right) \tag{7.74}$$

where E_A is the activation energy of diffusion and D_A^0 is the preexponential factor. The description above showed that this effective, intrinsic diffusivity is given by the sum of the diffusivities by each mechanism:

$$D_A^{\text{eff}} = D_{AI} + D_{AV} \tag{7.75}$$

where D_{AI} and D_{AV} are the contributions due to the I and V components of diffusion, respectively. Each of these terms depends on both the diffusivity of the mobile species and the fraction of the dopant atoms that are mobile, so that

$$D_A^{\text{eff}} = d_{AI}\left[\frac{C_{AI}}{C_A}\right] + d_{AV}\left[\frac{C_{AV}}{C_A}\right] \tag{7.76}$$

Table 7–6 Approximate values for f_I and f_V for silicon self-diffusion and for the common dopants in silicon

	f_I	f_V
Silicon	0.6	0.4
Boron	1.0	0
Phosphorus	1.0	0
Arsenic	0.4	0.6
Antimony	0.02	0.98

where d_{AI} or d_{AV} is the diffusivity of the actual mobile species causing the migration and C_{AI}/C_A and C_{AV}/C_A are the fractions of the dopant in the particular migrating state.

This is a much more rigorous description of the diffusion process, as it breaks the overall diffusion down into the diffusivity of each individual migrating species and the microscopic concentration of those migrating species. In the following sections, we will consider how these mobile species form on an atomic level and how Fick's laws can be applied to these mobile diffusing species to give a detailed description of the diffusion process on an atomic scale.

7.5.10 Activation Energy for Self-Diffusion and Dopant Diffusion

It is found experimentally that the activation energy [E_A in Eq. (7.74)] is about 1 eV lower for dopant diffusion in silicon than it is for self-diffusion (3 – 4 eV versus 4 – 5 eV). This raises a fundamental question regarding this difference. Because dopants have a lower activation energy for either I or V or mixed mechanisms of diffusion, there must be a lowering of the activation barrier specifically related to group III and V elements that is not present for group IV elements (Si or Ge). Coulombic effects can contribute to some but not all of this barrier lowering. (The dopant atoms are charged on substitutional lattice sites and the Si atoms are not.) If Coulombic effects alone dominated dopant-defect interactions, then all n-type or p-type dopants would have the same activation energy. Since this not the case, there must be additional dopant-defect interactions to be considered. One such possibility is the relaxation of the silicon lattice around a dopant, related to the size of the dopant atom. For example, theoretical calculations indicate that a substitutional boron atom causes the lattice to relax inwards by about 12% [7.26]. It is easy to imagine that lattice distortions when both dopants and defects are in close proximity could significantly alter the activation energy for diffusion. Another possibility relates to the electronic interactions due to charge transfer between the dopant and the defect. These effects lead to a higher probability of a point-defect binding to a dopant atom or interacting with a dopant atom to kick it out of its normal, substitutional lattice site. It is this interaction between dopant atoms and point defects that is the key to understanding dopant diffusion on an atomic scale. The difference in activation energies between self-diffusion and dopant diffusion plays a key role in how native point defects and mobile dopants interact as the temperature is changed, generating complex diffused profiles, as described in the following sections.

7.5.11 Dopant-Defect Interactions

We have seen that the macroscopic diffusivity of dopants is dominated by terms of the form $d_{AI}C_{AI}$ (or $d_{AV}C_{AV}$) on a microscopic scale. These terms describe the diffusivity of the mobile species and the concentration of the mobile species in the lattice. The mobile species come about via interactions of individual dopant atoms and individual point defects. We can consider a dopant atom A interacting with an interstitial I as follows:

$$A + I \leftrightarrow AI \qquad (7.77)$$

where AI is the actual interstitial-assisted mobile species on an atomic scale. The substitutional dopant is immobile by itself, unless it interacts with a point defect. (For a vacancy mechanism, I is replaced by V in this equation and the subsequent discussion.) This equation might represent either a migrating bound pair or a substitutional dopant "kicked out" of a lattice site by an interstitial to become a mobile dopant in an interstitial site. The justification for considering that the bound pair might migrate as a single entity comes from the difference in activation energy between self-diffusion and impurity diffusion. This activation energy difference indicates that there is a binding energy between an impurity and an intrinsic point defect. Our later mathematical treatment of Eq. (7.77) shows that there is no difference between a bound pair migrating and a pure dopant interstitial migrating. The overall picture in either case is that a silicon interstitial is temporarily consumed to generate the mobile species, which then migrates into the interior of the silicon sample, where the mobile dopant settles down on a lattice site to become a substitutional dopant once again, releasing a silicon interstitial in the process.

Eq. (7.77), although simple, contains a surprising amount of physics. It is clear that an interstitial supersaturation will drive more dopant atoms into a mobile state by shifting the equation to the right, thus enhancing the dopant diffusivity. This explains the phenomena of oxidation-enhanced diffusion, where the interstitial supersaturation during oxidation enhances the dopant diffusivity. A suppression of I, by nitridation of the silicon surface which introduces vacancies, can occur through the reaction

$$I + V \leftrightarrow Si_S \tag{7.78}$$

where Si_S represents a silicon atom on a lattice site. This will lead to retarded diffusion for dopant species that diffuse with interstitials.

Even under inert conditions where we consider the diffusion of a dopant from the surface (say, a shallow phosphorus implant), the atomic-level description says that the interior of the sample soon contains a lot of AI species, which drives Eq. (7.77) to the left, releasing interstitials in the interior when the mobile dopant regains a substitutional lattice position. In this way, silicon interstitials are "pumped" from the surface region into the interior region by the dopant diffusion. If there is a strong source of interstitials at surface kinks and ledges which maintains near-equilibrium levels of interstitials in the surface region, this leads to a supersaturation of interstitials in the interior of the sample. But a supersaturation of interstitials causes enhanced dopant diffusion. Thus, the tail of the phosphorus profile can diffuse much faster than expected, which is exactly what is observed experimentally. This demonstrates that diffusion pumping of point defects leads to point-defect supersaturations, which in turn feed back and enhance the diffusion process itself, as illustrated in Figure 7–40.

This description of "chemical pumping" of point defects also explains an anomalous interaction between emitter and base diffusions in some bipolar processes. The "emitter-push" or "emitter-dip" effect refers to the enhanced diffusion of the base profile below a phosphorus emitter as shown in Figure 7–41. This manifests itself in a much wider base width than expected. It is now understood to be caused by the pumping of interstitials by the diffusion of phosphorus which builds up a high supersaturation of

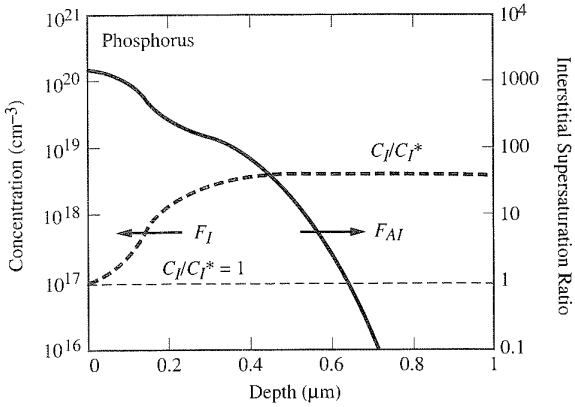


Figure 7–40 TSUPREM IV [7.14] simulation of the chemical or diffusion pumping of native point defects from the surface region into the bulk region by dopant-defect coupling. A quasi-steady state situation exists when the flux of free interstitials balances the flux of mobile dopant-interstitial pairs (or equivalently, mobile dopant interstitials). See text for a discussion of the C_I/C_I^* plots (right axis).

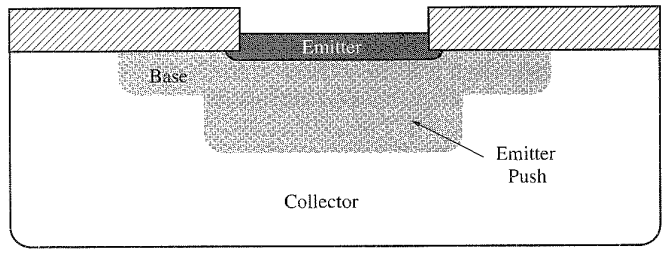


Figure 7–41 Schematic of the emitter-push effect in a bipolar transistor. The diffusion of the emitter junction affects the base region underneath the emitter and causes the base to "push out" further than expected [7.28]. This interaction is now known to be due to the chemical pumping of native point defects from the emitter region to the deeper bulk regions.

these point defects in the interior of the sample, well beyond the phosphorus diffusion front. The interstitial supersaturation then enhances the base (typically boron) diffusion in accordance with Eq. (7.76). We will return to this example later and show some SUPREM IV simulations that elucidate this coupling between dopant and self-interstitial fluxes.

In beginning this section we indicated qualitatively how a diffusing dopant atom transported interstitials into the bulk of the silicon, where it could enhance the tail diffusion of the diffusing dopant. We can now understand the origin of this effect if we properly account for the full coupling between the defects and the dopants when we consider the diffusion equation for the defects.

As seen in Figure 7–42, it is often important to include the full coupling between dopant and defects in simulations. Using only the concentration-dependent diffusivity for boron gives rise to a box-shaped simulated profile that does not match the experimental data. The β coefficient for boron [Eq. (7.57)] is the experimentally determined

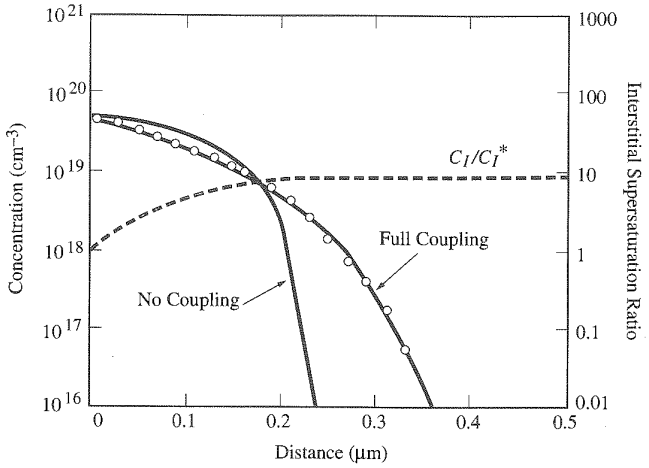


Figure 7–42 TSUPREM IV [7.14] simulations of boron diffused from a polysilicon source for eight hours at 850°C, with experimental data (circles) from [7.29]. The simulations use identical coefficients, but with and without full coupling between dopants and defects. The C_I/C_I^* curve (right axis) is for the fully coupled case. Without full coupling, $C_I/C_I^* = 1$.

value of 19 from highly doped extrinsic experiments (Table 7–5). One might be able to "fit" the experimental profile by adjusting the coefficients for neutral and positively charged boron diffusivity to make the profile less boxlike. However, by simply accounting for the coupling between boron and interstitials and therefore the enhanced diffusion in the tail region, the simulated profile agrees with the experimentally measured profile.

We want to be able to calculate how the supersaturation of point defects in the bulk is affected by the dopant transport in from the surface. The simplest interstitial diffusion equation is given by applying Fick's continuity equations to the native interstitial species:

$$\frac{\partial C_I}{\partial t} = -\frac{\partial}{\partial x}(F_I) \tag{7.79}$$

but this does not account for the fact that an interstitial is temporarily consumed and transported along when a mobile dopant atom diffuses through the lattice (this is most easily envisioned for the interstitialcy mechanism of dopant diffusion). The full coupling between the native interstitials and mobile dopant atoms is accounted for by including the flux of mobile dopant in the total interstitial continuity equation as follows:

$$\frac{\partial}{\partial t}(C_I + C_{AI}) = -\frac{\partial}{\partial x}(F_I + F_{AI}) \tag{7.80}$$

We can make some simplifying assumptions to estimate how this modification affects the interstitial supersaturation in the bulk. If we consider a quasi-steady state when the time derivative terms are small, then

$$\frac{\partial}{\partial x}(F_I + F_{AI}) = 0 \tag{7.81}$$

Integrating gives

$$F_I + F_{AI} = 0 \tag{7.82}$$

where we have taken the constant of integration to be zero. This is reasonable for many boundary conditions of interest. For example, the dopant flux (F_{AI}) deep in the crystal is certainly negligible and in quasi-steady state the interstitial flux deep in the crystal is also negligible, justifying our choice for the integration constant [7.29].
Thus,

$$F_I = -F_{AI} \tag{7.83}$$

which indicates that in steady state, the flux of dopant is balanced by a reverse flux of silicon interstitials, as shown in Figure 7–40. This balance between the dopant and defect fluxes is the key effect that results from a full coupling between the dopant and point-defect fluxes.

We can use this flux balance to investigate the magnitude of the interstitial supersaturation in the bulk, which is what causes the enhanced tail diffusion of dopants like phosphorus, the emitter-push effect, and boron profiles as in Figure 7–42. Since the flux of free interstitials is

$$F_I = -d_I \frac{\partial C_I}{\partial x} \tag{7.84}$$

we can use the flux balance equation [Eq. (7.82)] to write

$$\frac{\partial C_I}{\partial x} = \frac{F_{AI}}{d_I} \tag{7.85}$$

and we can obtain the depth dependence of the interstitial supersaturation by introducing the equilibrium interstitial concentration on both sides of the equation to give

$$\frac{\partial (C_I/C_I^*)}{\partial x} = \frac{F_{AI}}{d_I C_I^*} \tag{7.86}$$

We can relate the overall, observable macroscopic diffusion of the dopant to the microscopic diffusivity of the mobile species by realizing that the flux of the dopant is equivalent to the flux of the mobile species, giving

$$D_A^{\text{eff}} \frac{\partial C_A}{\partial x} = d_{AI} \frac{\partial C_{AI}}{\partial x} \tag{7.87}$$

Thus,

$$D_A^{\text{eff}} C_A = d_{AI} C_{AI} \tag{7.88}$$

which holds if the concentration of the mobile species bears some constant relationship to the substitutional dopant. This formulation for the effective dopant diffusivity brings out the importance of the (Diffusivity Concentration) or "DC" product of the mobile species in determining the overall motion of the dopant species. A more mathematical derivation of Eq. (7.88) is given in Section 7.5.12.

Since the flux of the mobile dopant species $F_{AI} = d_{AI}\nabla C_{AI}$ is equivalent to the overall flux of dopant $F_A = D_A^{\text{eff}}\nabla C_A$, we can write

$$\frac{\partial(C_I/C_I^*)}{\partial x} = \frac{D_A^{eff} C_A}{d_I C_I^*} \tag{7.89}$$

For an interstitial-type diffusion process, this equation indicates that the depth dependence of the interstitial supersaturation depends on the ratio of the dopant flux to the interstitial component of the self-diffusion flux. The severity of the chemical pump effect for different dopants relates to how the DC product for the dopant ($D_A^{\text{eff}} C_A$ or equivalently $d_{AI}C_{AI}$) compares with the DC product for the interstitial component of native silicon self-diffusion ($d_I C_I^*$).

Because of the difference in activation energies between self-diffusion (4.8 eV) and dopant diffusion (3–4 eV), Eq. (7.89) indicates that the magnitude of the interstitial supersaturation gets larger at lower temperatures, as shown in Figure 7–43. This is also the reason that the emitter-push effect is more pronounced for a fast diffusing dopant with a high solubility like phosphorus (high D_A^{eff}, high C_A) compared with a slow diffuser with a low solubility like arsenic, as shown in Figure 7–44. Thus, it explains why the emitter-push effect is more severe at lower temperatures and for fast diffusing emitter dopant species. Once again, this detailed understanding of the full coupling between mobile dopant atoms and native point defects on an atomic scale provides answers to questions that would otherwise remain a puzzle.

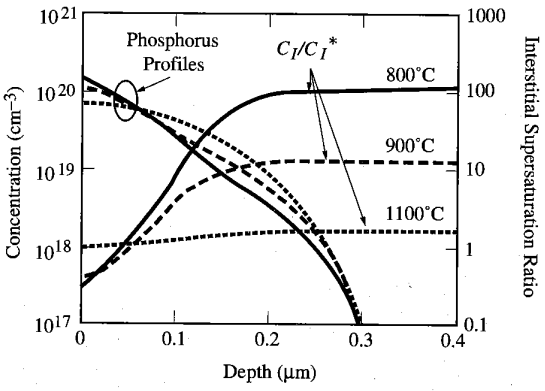


Figure 7–43 TSUPREM IV [7.14] simulations of the interstitial supersaturation generated by considering the fully coupled diffusion of phosphorus at different temperatures. The interstitial excess (right axis) generated by the diffusing dopant (left axis) can exceed a hundredfold at lower temperatures, making the fully coupled effect an important contribution to high-concentration diffusion modeling. The diffusion times at each temperature were chosen to give the same junction depth.

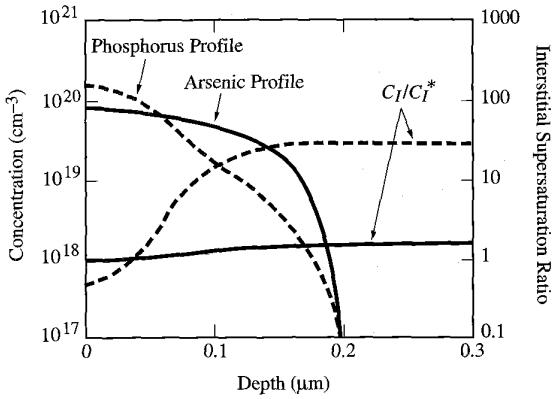


Figure 7-44 TSUPREM IV [7.14] simulations of the interstitial supersaturations generated by a phosphorus versus an arsenic diffusion to the same depth. The fast-diffusing phosphorus profile has a larger effect on C_I/C_I^* than the slow-diffusing arsenic profile.

7.5.12 Chemical Equilibrium Formulation for Dopant-Defect Interactions

In the previous section, we have seen how important a simple point-defect mediated reaction can be for determining how high-concentration dopants diffuse. The implications of such a reaction are now considered in more detail. Without loss of generality, if we consider an interstitial mediated reaction

$$A + I \leftrightarrow AI \tag{7.90}$$

to be in chemical equilibrium, then we can apply the law of mass action to the solid-state system and relate the concentrations by

$$C_{AI} = kC_AC_I \tag{7.91}$$

This particular description of the diffusion process is in "chemical equilibrium" in the sense that the mobile pairs are directly related to the reacting species by the equilibrium reaction constant. Experimentally this is justified because large enhancements in the diffusivity when excess interstitials are present can be seen for short diffusion times and remain the same for longer anneals. This indicates that the mobile species forms quickly and that the equilibrium assumption is valid.

Applying Fick's law to the mobile species C_{AI} gives

$$F_{AI} = -d_{AI}\frac{\partial C_{AI}}{\partial x} \tag{7.92}$$

At this point, our description of the diffusion flux is in terms of the unseen mobile species parameters ($d_{AI}C_{AI}$), the diffusivity and concentration of the mobile species in the lattice. These mobile species are not readily observable. Thus, the description as it stands is not very helpful. If we had numbers for the diffusivity of the mobile species itself and the concentration of the mobile species, we would have a useful description of the evolution of the profile shape. To make this description useful, we must relate these parameters to what can be observed, namely the overall dopant concentration (measured say,

by SIMS) and the apparent or effective diffusion coefficient of the total dopant concentration. A primary goal of this section will be to relate the microscopic diffusivity and concentration to parameters that can be measured at the macroscopic scale.

Differentiating Eq. (7.91) with respect to x and using the chain rule for differentiation gives

$$F_{AI} = -d_{AI}\left(kC_I\frac{\partial C_A}{\partial x} + kC_A\frac{\partial C_I}{\partial x}\right) \tag{7.93}$$

Now, we are interested in the overall observable diffusion of the dopant. But the substitutional dopant is immobile on an atomic scale, so F_A itself is zero. The only way for the substitutional concentration C_A to change in a small-volume element is if there is a difference in the flux of mobile species into and out of the volume element (similar to the argument used in Figure 7–7), giving

$$\frac{\partial C_A}{\partial t} = -\frac{\partial F_{AI}}{\partial x} = \frac{\partial}{\partial x}d_{AI}\left(kC_I\frac{\partial C_A}{\partial x} + kC_A\frac{\partial C_I}{\partial x}\right) \tag{7.94}$$

Note that the concentration of mobile species $C_{AI} << C_A$. We can infer that $C_{AI} << C_A$ experimentally. Silicon and phosphorus have similar mass and create similar interstitial and vacancy damage profiles in the lattice when they are implanted. Experimentally, the point defects from this damage enhance the diffusion of nearby buried dopant layers by the same amount (for low-implant doses, where there is no chemical pumping of the point defects). If a significant fraction of the damage-induced point defects were tied up with migrating phosphorus atoms (large C_{PI} or C_{PV} population), there would be a difference in the amount of enhanced diffusion seen at the buried layer for silicon and phosphorus implants at short diffusion times. There is not, so the dopant-defect complexes must be dilute.

The overall profile motion that is measured for a substitutional dopant C_A with an apparent measured diffusion coefficient D_A is in fact generated by a very small number of species on the atomic scale that become mobile intermittently (C_{AI}) and that migrate with a high diffusivity d_{AI} in the silicon lattice [illustrating the importance of the (DC) product again].

From Eq. (7.91), the equilibrium balance between the chemical species gives

$$C_{AI} = kC_AC_I \Rightarrow kC_I = \frac{C_{AI}}{C_A};\ kC_A = \frac{C_{AI}}{C_I} \tag{7.95}$$

so we can write the flux expression in the form

$$\frac{\partial C_A}{\partial t} = \frac{\partial}{\partial x}d_{AI}\left[\frac{C_{AI}}{C_A}\frac{\partial C_A}{\partial x} + \frac{C_{AI}}{C_I}\frac{\partial C_I}{\partial x}\right] \tag{7.96}$$

When we compare this to our original diffusion equation which claimed to describe the dopant diffusivity [Eq. (7.52)]

$$\frac{\partial C_A}{\partial t} = \frac{\partial}{\partial x}D_A^{\text{eff}}\left(\frac{\partial C_A}{\partial x}\right) \tag{7.97}$$

we see that there are both similarities and differences. Eqs. (7.96) and (7.97) are similar if the second term in Eq. (7.96) (the defect gradient term) is negligible, and then we can define the effective dopant diffusivity to be

$$D_A^{\text{eff}} = d_{AI}\frac{C_{AI}}{C_A} \tag{7.98}$$

which indicates that the macroscopic dopant diffusivity actually depends on the diffusivity of the mobile species and the fraction of the dopant that is actually mobile. This "effective diffusivity" is loosely considered to be the dopant diffusivity. It can easily be measured when the extra point-defect gradient term is negligible, which is most of the time for dopant diffusion in silicon. The effective diffusivity of a dopant should not be measured when there are large point-defect sources and sinks present, such as near a surface sink when large amounts of damage are present. Since we are considering an interstitial-assisted diffusion mechanism in this section, the effective diffusivity that can be measured is given by Eq. (7.65) for an interstitial type diffuser ($f_I = 1$)

$$D_A^{\text{eff}} = D_A^*\frac{C_I}{C_I^*} \tag{7.99}$$

Under equilibrium conditions where the interstitials are at their equilibrium level C_I^*, the measured diffusivity is the equilibrium, intrinsic, inert value D_A^*. The effective diffusivity that is enhanced by an interstitial supersaturation can also be measured by SIMS, providing there are no steep gradients in the interstitial profile in the region where the dopant is diffusing that could interfere with the measurement of the dopant diffusivity.

Combining the last two equations gives

$$d_{AI}C_{AI} = D_A^*C_A\frac{C_I}{C_I^*} \tag{7.100}$$

The microscopic parameters are now expressed in terms of the measured, macroscopic dopant diffusivity measured under equilibrium, inert conditions, the overall dopant concentration, and the supersaturation in the point-defect concentration which is easily estimated by observing how much the diffusion of an interstitial-type diffuser (P or B) is enhanced over its diffusion under normal, inert conditions (see problem 7.17). We have thus obtained a useful expression for the microscopic parameters of the mobile species that determine the diffusion flux.

7.5.13 Simplified Expression for Modeling

Because the derivative of the natural logarithm of x is $1/x$, we can further simplify the expression in Eq. (7.96) for the flux to

$$F_{AI} = -D_A^*C_A\frac{C_I}{C_I^*}\left[\frac{\partial}{\partial x}\ln(C_AC_I)\right] \tag{7.101}$$

Since C_I^* is a constant, we can include it in the gradient and in doing so we express the flux in terms of the gradient of the supersaturation rather than the absolute number of interstitials, giving

$$F_{AI} = -D_A^* C_A \frac{C_I}{C_I^*}\left[\frac{\partial}{\partial x}\ln\left(C_A\frac{C_I}{C_I^*}\right)\right] \tag{7.102}$$

This is really a nice accomplishment, because we have related the diffusion on an atomic scale to measurable parameters like the overall dopant diffusivity under inert equilibrium conditions, the overall dopant concentration, and the supersaturation of point defects which can be gauged by how much the dopant diffusivity is enhanced over its inert diffusivity. Thus, all the parameters in the equation are susceptible to measurement, even though the equation itself describes the atomic-scale motion of a short-lived intermittently migrating species which is responsible at the atomic scale for the overall motion of the dopant profile.

This expression is almost precisely the equation that is solved in most process simulators, with one omission. We left out the effect of an electric field on the diffusion of the mobile species. This modifies the flux equation for the mobile species to

$$F_{AI} = -d_{AI}\left[\frac{\partial C_{AI}}{\partial x} + \frac{q}{kT}C_{AI}\frac{\partial \psi}{\partial x}\right] \tag{7.103}$$

The potential ψ is given by

$$\psi = -\frac{kT}{q}\ln\frac{n}{n_i} \tag{7.104}$$

Thus, this simply adds another term in $\ln(n/n_i)$ to the overall flux equation, so with the inclusion of the electric field effects it becomes

$$F_{AI} = -D_A^* C_A \frac{C_I}{C_I^*}\left[\frac{\partial}{\partial x}\ln\left(C_A\frac{C_I}{C_I^*}\frac{n}{n_i}\right)\right] \tag{7.105}$$

The overall continuity equation for the dopant that is solved in a process simulator like SUPREM IV (with a similar term for the vacancy fraction of the diffusion flux) is

$$\frac{\partial C_A}{\partial t} = \frac{\partial}{\partial x}D_A^* C_A \frac{C_I}{C_I^*}\left[\frac{\partial}{\partial x}\ln\left(C_A\frac{C_I}{C_I^*}\frac{n}{n_i}\right)\right] \tag{7.106}$$

According to Eq. (7.106), there are three distinct effects that drive the diffusion flux—the gradient in dopant concentration, the gradient in the interstitial supersaturation, and the $\mathcal{E}$ field. The contribution from the point-defect gradient term is often overlooked [the interstitial gradient term in Eq. (7.106)]. The mathematical origin came from the application of the chain rule to the gradient in the mobile species, while the physical interpretation is that areas with high point-defect populations have faster diffusion, so a gradient in point defects can drive a diffusion flux. This cross term is of paramount importance in the neighborhood of point-defect sources and sinks, where

point-defect gradients are large. In these cases, it can dominate the overall diffusion of the dopant.

The full implications of this equation were realized only recently [7.30], when the impact of the gradient in point defects on the dopant profiles was shown to be responsible for altered threshold voltages in submicron MOS devices. The damage from implants in the source/drain regions (discussed in Chapter 8) combined with the local sink for point defects under the gate of the MOS device creates large gradients in the point-defect profiles in the channel region. An interesting practical example of this is the reverse short-channel effect in MOS devices, where point defects generated by implant damage in the source/drain regions flow toward the surface sink at the gate oxide interface, thereby creating large point-defect gradients which drive the overall dopant diffusion "uphill" toward the surface. These gradients in point defects drive dopant diffusion in a manner that is unexpected based on a simple application of Fick's law to the gradient in the doping profile itself. But Eq. (7.106) provides a complete description of the microscopic diffusion mechanisms and their otherwise unexpected impact on device operation. A simulation of how gradients in point-defect profiles affect the doping profile in a small MOS device is shown in Figure 7–45. The pileup of dopant under the gate oxide leads to an increase in the threshold voltage which we will discuss in Chapter 8. This effect is captured by the physically based diffusion model we have derived.

7.5.14 Charge State Effects

So far, we have ignored how the charge states of the defects or dopants affect the diffusion process, with the exception of mentioning that the overall concentration dependent dopant diffusivity can be written in the form

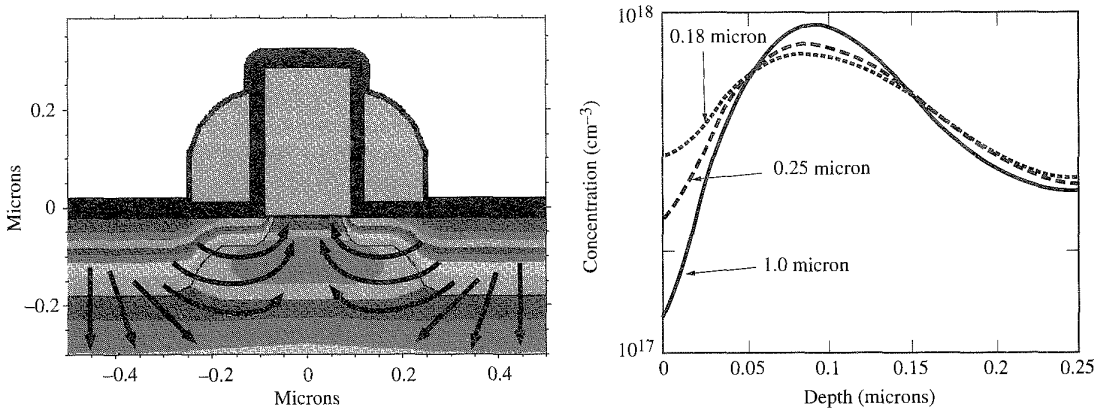


Figure 7–45 TSUPREM IV [7.14] simulation of short-channel MOS devices. The doping profiles are shown on the left with arrows that indicate the interstitial fluxes emanating from the diffusing heavily doped source drain regions. The doping profiles on the right are vertical plots through the center of the channel showing that as the device becomes smaller, more of the channel dopant piles up at the surface.

$$D_A^{\text{eff}} = D^O + D^-\left(\frac{n}{n_i}\right) + D^=\left(\frac{n}{n_i}\right)^2 \quad \text{N type} \tag{7.53}$$

$$D_A^{\text{eff}} = D^O + D^+\left(\frac{p}{n_i}\right) + D^{++}\left(\frac{p}{n_i}\right)^2 \quad \text{P type} \tag{7.54}$$

where each term is thought to originate from atomic-scale interactions with charged point defects.

However, because the dopant concentrations of interest in building devices are usually above the intrinsic concentration, high-concentration effects are often the dominant effect that should be taken into account in describing the diffusion profiles. This rather simplistic observation is the reason that effective diffusivities of the form shown in Eqs. (7.53) and (7.54) had such early success in process modeling programs—they capture the important first-order effects related to concentration-dependent diffusion that determine the shape of the diffused profiles. The concentration of charged point defects depends on the Fermi level. Thus, a deeper understanding of the interactions of dopants and charged defects to form mobile dopant species is required to understand the changes in the diffusion profiles that occur at high doping concentrations in real devices.

Fortunately, the process is very simple to sketch out in our description of the atomic-level transport process. We can take a concrete example of a boron dopant which interacts with an interstitial to form a mobile boron-interstitial species, with no loss of generality. Consider a substitutional boron acceptor (which donates a mobile hole to the electronic structure) that remains negatively charged on its lattice site, where it may interact with either neutral or positive interstitials. If no additional electrons or holes are involved in the reaction, this leads to the following two equations which describe the process:

$$B^- + I^0 \overset{K_A}{\leftrightarrow} BI^- \tag{7.107}$$

$$B^- + I^+ \overset{K_B}{\leftrightarrow} BI^0 \tag{7.108}$$

These two reactions give rise to fluxes of mobile species that we can sum to obtain the overall flux of mobile dopant:

$$F_{BI}^{\text{total}} = F_{BI^-} + F_{BI^0} \tag{7.109}$$

where each flux component can be specified using Fick's laws applied to the mobile species. In this case, we will take into account the effect of the electric field generated by the dopant profile only in the case where the mobile diffusing species is charged.

$$F_{BI^-} = -d_{BI^-}\left(\frac{\partial C_{BI^-}}{\partial x} + C_{BI^-}\frac{\partial}{\partial x}\ln\frac{p}{n_i}\right) \tag{7.110}$$

$$F_{BI^0} = -d_{BI^0}\left(\frac{\partial C_{BI^0}}{\partial x}\right) \tag{7.111}$$

Given the importance of understanding how charge-state information affects the diffusion process, it is desirable to try to connect the atomic-level description to the observed shape of the overall dopant diffusion profile. The assumption that the reaction of point defects and dopants are in chemical equilibrium is useful and is certainly valid for most of the processes of interest in silicon technology. This allows the concentration of the mobile species to be written in terms of the dopant and free point defects as follows:

$$C_{BI^-} = k_A C_{I^0} C_{B^-} \tag{7.112}$$

$$C_{BI^0} = k_B C_{I^+} C_{B^-} \tag{7.113}$$

Another major simplification occurs because electronic processes occur much faster than the atomic jumps that mediate diffusion steps, which occur on a time scale of the Debye frequency for lattice vibrations. Thus, the electronic charging process is fast compared to any of the diffusion processes that occur in silicon. This allows the concentration of defects in the charged state to be directly related to the concentration of neutral defects in the silicon, according to Shockley-Last statistics [7.31]. (See Chapter 3, Section 3.5.4.)

$$C_{I^+} = C_{I^0} \exp\left(\frac{E_{I^+} - E_F}{kT}\right) \tag{7.114}$$

By a similar analysis to the previous charge neutral case, we can show that

$$D_{BI}^{\text{eff}} = D_{BI}^{\gtrsim} \cdot \frac{\left(1 + \beta \frac{p}{n_i}\right)}{1 + \beta} \cdot \frac{C_{I^0}}{C_{I^0}^{\gtrsim}} \tag{7.115}$$

where β represents the enhancement in the boron diffusivity due to the Fermi level and is dependent on the diffusivity of the pairs, the concentrations of the pairs, and the energy level of the charged point defect in the bandgap.

Quite a bit of the microscopic physics is lumped into the β constant—it contains information on the diffusivities of the individual charged pairs, the charged concentrations of the native point defects that form the pairs, and the relative concentrations of the pairs in each charge state (through the ratio of the reaction constants). However, it has a very physical interpretation at the macroscopic scale—it simply represents how much the overall diffusivity is enhanced in highly doped regions and was already considered in an empirical model for high concentration diffusivity in Eq. (7.57). Its microscopic origins are now more clearly elucidated.

The overall diffusivity of the mobile boron pairs is thus given by a product of three easily interpreted terms—the intrinsic, inert equilibrium diffusion coefficient that can be measured for low concentration boron profiles, the extent to which this intrinsic diffusion coefficient is enhanced in highly doped material by the Fermi-level, and the supersaturation of interstitials.

The final equation representing the flux which includes all the Fermi-level effects can be written as

$$F_{BI}^{\text{tot}} = D_{BI}^{*} \cdot \left(\frac{1 + \beta \frac{p}{n_i}}{1 + \beta} \right) \cdot \frac{C_I^0}{C_I^{*0}} \cdot C_B \frac{\partial}{\partial x} \ln\left(C_B \frac{C_I^0}{C_I^{*0}} \frac{p}{n_i} \right) \tag{7.116}$$

This equation represents a remarkable step in our understanding of dopant diffusion in silicon. What is most interesting is that we started with a detailed atomic-level description of the diffusion by point defects and mobile dopant species in different charge states and ended up with an equation containing only macroscopic terms that can easily be measured experimentally. Each one of the terms represents a distinct physical effect that can easily be investigated experimentally by doing measurements on the total concentration of the dopant, using techniques such as SIMS. This gives us numbers to plug into the equation for parameters like D_{BI}^{*} and β.

But there is a downside to our successful development of Eq. (7.116)—it means that it is very difficult to say anything definite about the underlying atomic-level processes by making macroscopic measurements on the overall dopant motion. Instead, we can try to infer what is happening on an atomic scale by using clever test structures and by comparing and contrasting how different dopants diffuse under various conditions. But in the end, we must build an atomic-scale model for how the dopant moves and see if the equations based on that atomic-scale model are a useful description of the physical world where we make measurements.

7.6 Limits and Future Trends in Technologies and Models

A tremendous amount has been learned in the past 20 years about dopant diffusion and doping profiles in silicon. The simulation tools used today are very powerful and are widely used to develop new generations of technology. While we have focused our discussion on the issues associated with CMOS technology, these tools are also used in many other types of silicon device structures. In many cases they are crucial in reducing the number of experiments required to develop new processes. They are also very useful in understanding process sensitivities and manufacturing tolerances.

The SIA NTRS for mainstream CMOS technology makes it very clear that the main challenges for the future lie in making very shallow junctions to minimize short-channel effects, with low sheet resistances to maximize drive currents. An obvious limit to device scaling occurs when the junctions are so shallow that the doping concentrations are at the solubility limit—making the junctions even shallower only increases the sheet resistance and degrades drive current. Other doped regions (deeper junctions like wells, for example), will continue to be used, but these are easily fabricated with today's technology. There will thus be continued emphasis on understanding dopant diffusion and activation at the atomic level because the shallow doping profiles required in future devices may have metastable activation levels and will have to be controlled almost to the point of knowing where individual doping atoms are located. The physical understanding and the simulation tools described in previous sections will likely not be adequate for these kinds of applications. Thus the sophistication and the complexity of models

will likely continue to increase. We describe in this section a few extensions to the technologies and the models described above which indicate some of the directions current research is taking to address these issues.

7.6.1 Doping Methods

Most workers believe that the dominant doping method used today—ion implantation—will continue to dominate in the future. Ion implantation affords tremendous flexibility in terms of the doping species, the physical location of the dopants, and in manufacturing throughput. Deeper doped regions in device structures are straightforward to form by moderate- or high-energy implants. The very shallow junctions required in future devices can be formed by very low energy implants. The only real drawback to this technology is the damage produced by the implantation process and the transient diffusion and extended defects that result when implants are annealed. In order to minimize these effects, rapid thermal processing is likely to dominate annealing and diffusion processing in the future. We will discuss ion implantation and some of the issues associated with extending it to future device fabrication in Chapter 8.

A number of alternative doping methods are currently being investigated and may find some use in the future. These methods range from very old methods like outdiffusion from deposited thin films, to very new processes like laser-assisted annealing or doping. In the latter category, processes such as Gas Immersion Laser Doping (GILD) have been suggested, in which the wafers are placed in a controlled environment containing the dopant in a gas phase. A laser is used to heat the wafer either by scanning the laser beam or selectively using a mask. The dopant is incorporated into a thin surface layer where the silicon is melted by the laser [7.32]. This process can produce ultrashallow junctions, sometimes with very high, metastable active doping concentrations.

7.6.2 Advanced Dopant Profile Modeling—Fully Kinetic Description of Dopant-Defect Interactions

Even more sophisticated models than those described previously may be required for very shallow junctions produced by ion implantation and very short anneals. This is because in our previous modeling, we considered the reaction

$$A + I \leftrightarrow AI \tag{7.117}$$

to be in equilibrium. However, there are experimental hints that the chemical reaction between the dopants and the intrinsic point defects to form the intermediate mobile diffusing species may not be in chemical equilibrium for very short diffusion times at very low temperatures (e.g., 15 minutes at 450°C). The experimental observations show exponential tails on the diffusion profiles, rather than the expected Gaussian profiles, indicating a single hop or migration event for the mobile species [7.33]. A more general description of the diffusion process is to consider that the reaction terms are not yet in chemical equilibrium and are described atomistically by their forward and backward reaction rates, so that the fully kinetic reaction rate is given by

$$R = k_f C_A C_I - k_r C_{AI} \tag{7.118}$$

where k_f and k_r are the forward and reverse reaction rates. The diffusion of the dopant is then atomistically described by a series of reaction-diffusion equations. These are standard rate equations for chemical kinetics.

For example, the substitutional dopant A, which is immobile on an atomic scale, can only decrease its local concentration by reacting with an interstitial in a reaction which is proportional to both the local concentration of A and the local concentration of I,

$$\left.\frac{\partial C_A}{\partial t}\right|_{\text{forward}} = - k_f C_A C_I \tag{7.119}$$

Similarly, it can increase its local concentration by dissociation of the mobile species, that is, if the reaction goes to the left so that

$$\left.\frac{\partial C_A}{\partial t}\right|_{\text{reverse}} = + k_r C_{AI} \tag{7.120}$$

Similar terms govern the time evolution of the mobile species C_{AI} and C_I, with the exception that their time rate of change is also governed by their ability to diffuse. We know how to describe that process by Fick's laws, so that the overall system of reaction-diffusion equations that describes the diffusion process is given by

$$\frac{\partial C_A}{\partial t} = - k_f C_A C_I + k_r C_{AI} \tag{7.121}$$

$$\frac{\partial C_{AI}}{\partial t} = d_{AI} \frac{\partial^2 C_{AI}}{\partial x^2} + k_f C_A C_I - k_r C_{AI} \tag{7.122}$$

$$\frac{\partial C_I}{\partial t} = d_I \frac{\partial^2 C_I}{\partial x^2} - k_f C_A C_I + k_r C_{AI} \tag{7.123}$$

Here, we have assumed that d_{AI} and d_I are both constants. This is a fully kinetic formulation of the diffusion process.

The above example actually considered only a reduced set of the possible reactions that could occur when a particular dopant diffuses. Generalizing these ideas produces a complex set of equations with many parameters which at first glance seems almost intractable. However this approach has been taken successfully in a number of cases in recent years and may be required for more cases in the future. For example, the reactions which are thought to be involved in modeling high-concentration phosphorus diffusion have been solved based on the work of Dunham [7.27] and of Mulvaney and Richardson [7.34]. What is impressive is that a large number of these highly coupled, nonlinear differential equations can be solved numerically and can generate a lot of physical insight into the diffusion process.

One characteristic of the fully kinetic approach is that it contains a large number of parameters. Some method for estimating these parameters must be used, as there is no

easy means to directly measure these quantities. Fortunately, chemists have a fundamental interest in reaction rate kinetics and have well developed theories to describe the rate constants of simple reactions.

So, what is really known about the coefficients that are needed to solve these reactions and where does that information come from? About the only thing that is known with certainty is what can be directly measured, that is, the overall diffusion coefficient of a dopant species or of a silicon isotope in the silicon lattice. The goal of the fully kinetic approach to modeling is to estimate the microscopic diffusion parameters (such as the concentration of mobile species and their diffusivities), based on what is known.

Recently, due to breakthrough developments in theoretical methods and the availability of powerful computers, direct quantum mechanical calculations of some of the microscopic parameters have been made. These *ab initio* (from first principles) calculations currently give reliable information for the ground state properties of the system and calculations for the charged states are under development. More approximate quantum mechanical methods (tight-binding methods) allow the silicon lattice defects to be directly simulated at high temperatures, where their migration in the lattice can be observed and values for the diffusivity can be calculated. Such computational approaches will play a more important role in understanding the atomic-level details of the process kinetics in the future. However, the full set of parameters needed for solution of a system of reaction-diffusion equations has been attempted to date only in a few special cases.

7.7 Summary of Key Ideas

Doping of semiconductors is the fundamental process by which devices are fabricated. To make dopants electrically active and therefore useful in devices, they must generally sit on substitutional lattice sites. Both the concentration and the spatial variation (profile) of these dopants are critical to device performance. Diffusion is thus the basic process by which we produce these doping profiles.

Ion implantation is the dominant method used today to introduce dopants into silicon. This process allows precise control of the numbers of doping atoms and their physical position. Annealing and sometimes additional diffusion are required following implantation in order to allow the dopants to diffuse to their final, substitutional lattice sites. For short times, this process is dominated by transients associated with the implantation damage. During this period, the dopant diffusivity is much larger than its equilibrium value because of the high concentration of point defects present. Following this transient, the dopant diffusivity drops to a more reproducible value.

Dopants diffuse in silicon by interacting with point defects through a number of possible atomic-scale mechanisms. The dopant diffusivity is basically proportional to the number of such defects present in a crystal. Point-defect concentrations depend exponentially on temperature and as a result, dopant diffusivities do also. However point-defect concentrations are also changed by Fermi-level effects, by ion implantation damage, and by surface processes like oxidation. The result is that dopant diffusivities may vary spatially and with time during the diffusion of a particular doped region. This greatly complicates the modeling of dopant profiles and generally re-

quires that numerical simulation tools be used to calculate such profiles. Only in the simplest cases in which D is constant can simple analytical solutions be used to calculate impurity profiles.

Much progress has been made in recent years in developing physically based models for dopant diffusion and in implementing these models in simulators. These simulators today allow full coupling of the point defects and the dopants and can accurately predict very complex experimental dopant profiles in silicon. Many complex insights such as point-defect gradient-driven up-hill diffusion and chemical pumping of point defects appear as a trivial consequence of these physically based atomistic models of dopant diffusion. In the future, even deeper understanding of these atomistic mechanisms and more complete models may be required. The reaction-diffusion approach described above is suitable for treating nonlinear processes far from equilibrium (such as transient-enhanced diffusion after ion implantation) and is conceptually simple. It also offers the possibility that theoretical models of reaction pathways (i.e., binding energies and cross sections for reactions) can integrate seamlessly into the overall picture of diffusion in silicon. For these reasons, this description of transport processes is now favored for the silicon system.

7.8 References

[7.1]. R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of Ion-implanted MOSFET's with Very Small Physical Dimensions," *IEEE Journal of Solid State Circuits*, vol. 9, no. 5. p. 256, 1974.

[7.2]. "National Technology Roadmap for Semiconductors," SIA, 1997.

[7.3]. F. A. Trumbore, "Solid Solubilities of Impurity Elements in Germanium and Silicon," *Bell System Technical Journal*, vol. 39, p. 205, 1960.

[7.4]. L. S. Darken and R. W. Gurry, *Physical Chemistry of Metals*, McGraw-Hill, 1953.

[7.5]. J. C. Irving, "Resistivity of Bulk Silicon and Diffused Layers in Silicon," *Bell System Technical Journal*, vol. 41, p. 387, 1962.

[7.6]. R. Alvis, S. Luning, L. Thompson, R. Sinclair, and P. Griffin, "Physical Characterization of Two-dimensional Doping Profiles for Process Modeling," *J. Vac. Sci. Technology B*, vol. 14., no. 1, p. 231.

[7.7]. H. J. Cho, P. B. Griffin, and J. D. Plummer, "Cross sectional TEM Sample Preparation Using e-beam Lithography and Reactive Ion Etching," Paper Z2.18, 1997 MRS Spring Meeting, San Francisco, April 1997.

[7.8]. B. Schwartz and H. Robbins, "Chemical Etching of Silicon. IV. Etching Technology," *Journal of the Electrochemical Society*, vol. 123, no.12, p.1903, 1976.

[7.9]. D. P. Gold, J. H. Wills. G. R. Booker, M. C. Willson, and D. J. Godfrey, *Microscopy of Semiconductor Materials*, edited by A. G. Cullis and J. L. Hutchinson, Institute of Physics Conference Series, vol. 100, p. 537, 1989.

[7.10]. N. Khalil, J. Faricelli, D. Bell, and S. Selberherr, "The Extraction of Two-dimensional MOS Transistor Doping via Inverse Modeling," *IEEE Electron Device Letters*, vol. 16. no. 1. p.17, 1995.

[7.11]. Z. K. Lee, M. B. McIlrath, and D. A. Antoniadis, "Inverse Modeling of MOSFETs Using I-V Characteristics in the Subthreshold Region," *IEDM Technical Digest*, p. 683, 1997.

[7.12]. G. J. L. Ouwerling, "A Problem-specific Inverse Method for Two-dimensional Doping Profile Determination from Capacitance-voltage Measurements," *Solid State Electronics*, vol. 34, no. 2, p. 197, 1991.

[7.13]. SUPREM IV is a computer program written by Mark Law, Conor Rafferty, and Robert Dutton at Stanford University.

[7.14]. TSUPREM IV is the version of SUPREM IV by Avant! Inc.

[7.15]. ATHENA is the version of SUPREM IV by Silvaco Inc.

[7.16]. J. Crank, *The Mathematics of Diffusion*, Oxford University Press, 1979.

[7.17]. R. B. Fair and J. C. C. Tsai, "Theory and Direct Measurement of Boron Segregation in SiO_2 in Dry, Near Dry, and Wet O_2 Oxidation," *J. Electrochem. Soc.*, vol. 125, pp. 2050, 1978.

[7.18]. A. S. Grove, *Physics and Technology of Semiconductor Devices*, John Wiley & Sons, 1967.

[7.19]. R. Kasnavi, P. Pianetta, Y. Sun, R. Mo, P. Griffin, and J. D. Plummer, "Characterization of Arsenic Dose Loss at the Si/SiO2 Interface Using High Resolution X-ray Photoelectron Spectrometry," *IEDM Technical Digest*, p. 721, 1998.

[7.20]. S. M. Hu, private communication.

[7.21]. R. B. Fair, in *Impurity Doping Processes in Silicon*, edited by F. F. Y. Wang, North-Holland, 1981, p. 315.

[7.22]. S. M. Hu, "Formation of Stacking Faults and Enhanced Diffusion in the Oxidation of Silicon," *Journal of Applied Physics*, vol. 45, p. 1567, 1974.

[7.23]. E. A. Perozziello, P. B. Griffin, and J. D. Plummer, "Retarded Diffusion of Sb in a High Concentration as Background During Silicon Oxidation," *Applied Physics Letters*, vol. 61, no. 3. p.303, 1992.

[7.24]. P. M. Fahey, P. B. Griffin, and J. D. Plummer, "Point Defects and Dopant Diffusion in Silicon," *Reviews of Modern Physics*, vol. 61, p. 289, 1989.

[7.25]. S. M. Hu, "Nonequilibrium Point Defects and Diffusion in Silicon," *Materials Science and Engineering*, vol. R13, no. 3–4, pp. 105–192, 1994.

[7.26]. J. Zhu, T. Diaz de la Rubia, L.H. Yang, C. Mailhiot, and G. H. Gilmer, "Ab Initio Pesudopotential Calculations of B Diffusion and Pairing in Si," *Physical Review B*, vol. 54, no. 7, p. 4741, 1996.

[7.27]. S. T. Dunham, "A Quantitative Model for the Coupled Diffusion of Phosphorus and Point Defects in Silicon," *Journal of the Electrochemical Society*, vol. 139, no. 9, p. 2628, 1992.

[7.28]. R. B. Fair, "Explanation of Anomalous Base Regions in Transistors," *Applied Physics Letters*, vol. 22. no. 4. p. 186, 1973.

[7.29]. W. A. Orr Arienzo, R. Glang, R. F. Lever, R. K. Lewis, and F. F. Morehead, "Boron Diffusion in Silicon at High Concentrations," *J. Appl. Phys.*, vol. 63, no. 1, p. 116, 1988.

[7.30]. C. S. Rafferty, H. H. Vuong, S. A. Eshraghi, M. D. Giles, M. R. Pinto, and S. J. Hillenius, "Explanation of Reverse Short Channel Effect by Defect Gradients," *IEDM Technical Digest*, p. 311, 1993.

[7.31]. W. Shockley and J. Last, "Statistics of the Charge Distribution for a Localized Flaw in a Semiconductor," *Phys. Rev.*, vol. 107, p. 392, 1957.

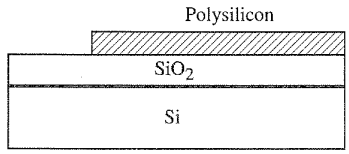
[7.32]. K. H. Weiner, P. G. Carey, A. M. McCarthy, and T. W. Sigmon, "Low-temperature Fabrication of p-n Diodes with 300Å Junction Depth," *IEEE Electron Device Letters*, vol. 13, no. 7, p. 369, 1992.

[7.33]. N.E.B. Cowern, G.F.A. van de Walle, P.C. Zalm, and J.J. Oostra, "Reactions of Point Defects and Dopant Atoms in Silicon," *Physical Review Letters*, vol. 69, no.1, p. 116, 1992.

[7.34]. B. J. Mulvaney and W. B. Richardson, "The Effect of Concentration-dependent Defect Recombination Reactions on Phosphorus Diffusion in Silicon," *Journal of Applied Physics*, vol. 67, no. 6, p. 3197, 1990.

7.9 Problems

7.1. A resistor for an analog integrated circuit is made using a layer of deposited polysilicon 0.5 μm thick, as shown below.

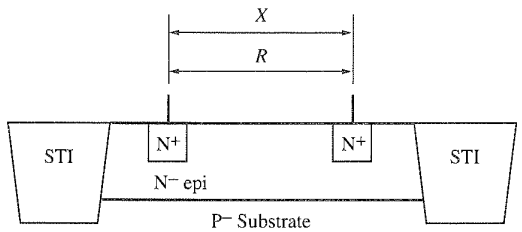


a. The doping of the polysilicon is 1×10^{16} cm^{-3}. The carrier mobility $\mu = 100$ cm^2 V^{-1} sec^{-1} is low because of scattering at grain boundaries. If the resistor has $L = 100$ μm, $W = 10$ μm, what is its resistance in ohms?

b. A thermal oxidation is performed on the polysilicon for two hours at 900°C in H_2O. Assuming B/A for polysilicon is $^2/_3$ that of <111> silicon, what is the polysilicon thickness that remains?

c. Assuming that all of the dopant remains in the polysilicon (i.e., does not segregate to oxide), what is the new value of the resistor in (a). Assume the mobility does not change.

7.2. A resistor is made as part of a high-frequency analog integrated circuit as shown below. The N^- epi layer forms the body of the resistor. If the width of the resistor in the direction into the paper is 2.5 μm, what should the length X be to give a resistor of approximately 50 kΩ?. The epilayer is doped with phosphorus at a concentration of 1×10^{15} cm^{-3} and is 3 μm thick.



7.3. A p-type (boron) diffusion is performed as follows:
Predep: 30 minutes, 900°C, solid solubility
Drive in: 60 minutes, 1000°C
a. What is the deposited Q?
b. If the substrate is doped 1×10^{15} cm^{-3} phosphorus, what is x_j?
c. What is the sheet resistance of the diffused layer?

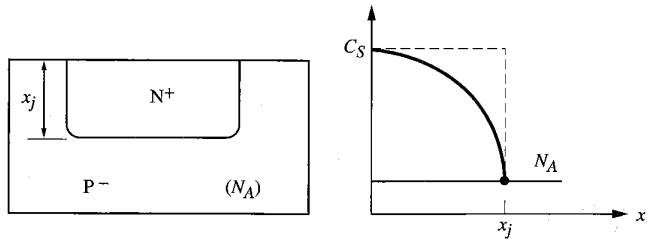
7.4. Suppose we perform a solid solubility limited predeposition from a doped glass source which introduces a total of Q impurities / cm^2.
a. If this predeposition were performed for a total of t minutes, how long would it take (total time) to predeposit a total of $3Q$ impurities / cm^2 into a wafer if the predeposition temperature remained constant?
b. Derive a simple expression for the $(Dt)_{\text{drive in}}$ which would be required to drive the initial predeposition of Q impurities / cm^2 sufficiently deep so that the final surface concentration is equal to 1% of the solid solubility concentration. This can be expressed in terms of $(Dt)_{\text{predep}}$ and the solid solubility concentration C_S.

7.5. A diffused region is formed by an ultrashallow implant followed by a drive in. The final profile is Gaussian. Derive a simple expression for the sensitivity of x_j to the implant dose Q. Is x_j more sensitive to Q at high or low doses?

7.6. From a process control point of view, predeposition times > 10 min are required. From an economic point of view, times < 10 hours are required. Equipment limitations restrict $700°C < T < 1200°C$. You need only consider simple Gaussian- and erfc-type profiles in this problem.
a. Is it possible to dope a MOS channel region by predeposition to shift threshold voltages? The required dose is 5×10^{11} cm^{-2}, boron.
b. What would be a reasonable schedule (T, t) for a MOS source drain predep? The required dose is 5×10^{15} cm^{-2}, arsenic. The junction depth cannot be deeper than 0.2 μm for device reasons. Assume the channel doping is 1×10^{18} cm^{-3}.
c. What sheet resistance would your profile in (b) have if the channel doping is 1×10^{18} cm^{-3} P type?

7.7. A boron diffusion is performed in silicon such that the maximum boron concentration is 1×10^{18} cm^{-3}. For what range of diffusion temperatures will electric field effects and concentration-dependent diffusion coefficients be important?

7.8. An N^+ region is formed in a P^- substrate with a junction depth is x_j as shown below. For the device being fabricated it is important to minimize the sheet resistance of the N^+ region.



a. If the ideal "box" profile shown could be obtained, it would provide an absolute minimum limit on ρ_s. Derive an approximate expression for this lower bound on ρ_s.

b. If As had been used as the dopant and a metastable doping at the solid solubility of 2×10^{21} cm^{-3} had been obtained in a box-shaped profile by laser melting, estimate the absolute minimum ρ_s for an x_j of 0.1 μm if the limiting carrier mobility in N^+ silicon is 85 cm^2 volt^{-1} sec^{-1}.

c. If a normal error function profile were used with the surface at the metastable solid solubility, what value of ρ_s would be realized with an x_j of 0.1 μm?

d. Repeat the calculation in (c) if the arsenic had deactivated to its normal "electrical solubility" in silicon.

7.9. A bipolar transistor is fabricated by implanting the boron base with a dose of $Q_B = 2 \times 10^{13}$ cm^{-2}. Then an in situ arsenic-doped polysilicon emitter region doped at 10^{20} cm^{-3} is deposited on the surface. A drivein is performed for 60 min at 1100°C. Assume that the implant can be treated as a delta function. The substrate is 10^{15} cm^{-3} N type. Calculate the base width of the final NPN bipolar transistor. You can neglect all second-order effects like concentration-dependent diffusion, $\mathcal{E}$-field effects, segregation, rapid diffusion in polysilicon, and so forth.

7.10. A special twin-well (twin-tub) CMOS technology requires that the wells have precisely the same depth at the substrate concentration of 1×10^{15} cm^{-3}, with arsenic used for the n tub and boron used for the p tub. A shallow implant dose of 1×10^{14} cm^2 is used for both and the slow diffusing arsenic is introduced first and partially driven in. Then the boron is introduced and the rest of the anneal is performed until both junctions reach 2.5 microns. Calculate all of the drivein times and temperatures used.

7.11. A process engineer on the day shift started a boron isolation diffusion for a structure in which the boron diffusion needs to penetrate completely through a 6-μm-thick N-type epitaxial layer that is lightly doped with phosphorus ($N_D = 1 \times 10^{15}$ cm^{-3}) on a P-type substrate ($N_A = 1 \times 10^{14}$ cm^{-3}). The purpose of the diffusion is to provide isolation between different N-type regions. The day shift engineer left no information on what he or she did. On a monitor wafer, you make

measurements and find that probing between adjacent N-type regions shows they are not isolated from each other and that the sheet resistance between the probes is 20 kΩ. The sheet resistance of the p-type diffusion measured using a four-point probe is 25 Ω/square. What time and temperature are needed to complete the isolation diffusion?

7.12. A bipolar transistor is fabricated by implanting the base and emitter regions and then driving them in together. Boron is used for the base and arsenic for the emitter. The implants are $Q_B = 2 \times 10^{13}$ cm^{-2} and $Q_E = 2 \times 10^{15}$ cm^{-2}. The drive in is 60 min at 1100°C. Assume that the implants can be treated as delta functions at the surface. The substrate is 10^{15} cm^{-3} N type.

a. Calculate and plot the resulting impurity profiles. Assume that the B and As diffusivities are the values from Table 7–5. You can neglect all second order effects like concentration dependent diffusion, $\mathcal{E}$-field effects, and so forth.

b. A parameter of interest in bipolar transistors is the base Gummel number which is the total dose (atoms cm^{-2}) in the base region under the emitter. Estimate this parameter for the device in part (a).

7.13. All semiconductor devices are really "unstable" at device operating temperatures because dopants will continue to diffuse and hence, given long enough, profiles will change and devices will stop functioning correctly. How long would it take at a 150°C operating temperature for the base width in the device in problem 7.12 to double? Assume the As does not move, just the boron moves.

7.14. An epitaxial layer is grown with uniform concentrations of B and P impurities ($B = 2 \times 10^{15}$ cm^{-3}, $P = 1 \times 10^{15}$ cm^{-3}) so that the net doping is p type. The wafer is oxidized at 1000°C for 60 minutes. Approximately (use $2\sqrt{Dt}$ as a measure of the profile motion) sketch the resulting impurity profiles for both boron and phosphorus under the oxide which result from the redistribution that occurs.

7.15. A silicon wafer is uniformly doped with boron (2×10^{15} cm^{-3}) and phosphorus (1×10^{15} cm^{-3}) so that it is net P type. This wafer is then thermally oxidized to grow about 1 μm of SiO_2. The oxide is then stripped and a measurement is made to determine the doping type of the wafer surface. Surprisingly it is found to be N type. Explain why the surface was converted from P to N type. *Hint*: Consider the segregation behavior of dopants when silicon is oxidized.

7.16. An engineer wants to use analytical solutions to diffusion equations in a programmable calculator to make rapid estimates for process changes on junction depths. Consider the following possible diffusion regimes: (a) high temperatures, (b) low temperatures, (c) long times, (d) short times. Which of them are most appropriate for analytical solutions (i.e., which would minimize $\mathcal{E}$-field or concentration-dependent effects). Explain.

7.17. A shallow phosphorus implant with a dose of 1×10^{14} cm^{-2} is covered in some regions with a deposited layer of inert nitride. An anneal is performed at 1000°C in dry O_2 and a junction depth below the original surface is measured in the inert region of 0.5 μm and of 1.2 μm under the oxidizing region. What is that diffusivity enhancement that the phosphorus in the oxidizing region experiences? If f_I for phosphorus is 0.9 and I-V recombination is efficient at 1000°C, what is the interstitial supersaturation that is generated by the oxidation?

7.18. Rapid Thermal Annealing (RTA) systems are becoming common for activating dopants. Silicon has quite a high thermal diffusivity of 0.88 cm^2 sec^{-1}, which describes how fast heat flows through silicon. Calculate the time required for a silicon wafer of thickness 500 μm to reach a constant temperature if a thin surface layer absorbs all the incident light and quickly reaches a reaches a steady-state temperature of T_S, that is, when does the center reach $T/T_S = 0.5$, since the wafer is heated from both sides?

7.19. By using Eq. (7.72) ($C_I C_V = C_I^* C_V^*$) and letting $C_V/C_V^* = x$, solve for x and use the fact that the solution must be real and positive to derive Eq. (7.73). Using the fact that $f_I = 1 - f_V$ and completing the square simplifies the solution. Alternatively, use Eq. (7.72) to substitute for C_V/C_V^* in Eq. (7.66) and take the derivative with respect to C_I/C_I^* and find an extreme (maximum or minimum).

7.20. Figure 7–38 shows that a wet oxidation produces a significantly higher C_I/C_I^* than does a dry O_2 oxidation. Explain quantitatively why this should be the case.

Ion Implantation

8

8.1 Introduction

Ion implantation has been the dominant doping technique for silicon ICs for the past 20 years. It is expected to retain this position of dominance for the foreseeable future. In this process, dopant ions are accelerated to hundreds or thousands of volts of energy and smashed into a perfect silicon lattice, creating a cascade of damage that may displace a thousand silicon atoms for each implanted ion. It is the purpose of this chapter to understand how such an energetic and violent technique has become the dominant and preferred method of doping silicon wafers in manufacturing. At first glance, it seems that the technique would not be of much use in the precise art of fabricating integrated circuits. Indeed, although the original patent for ion implantation was issued to William Shockley in 1954, it was not until the late seventies that ion implantation was used in manufacturing. To appreciate the power of the technique, we will investigate its ability to control precisely the distribution and dose of various dopants in the silicon and examine how the implant damage is annealed and how the dopant is introduced to active sites in the silicon lattice.

8.2 Historical Development and Basic Concepts

Ion implantation provides a very precise means to introduce a specific dose or number of dopant atoms into the silicon. The reason this is possible is because the electrical charge on the ion allows them to be counted by collection in a Faraday cup. Electrical measurements are very precise, so numbers of dopant atoms ranging from 1×10^{12} cm^{-2} to 1×10^{16} cm^{-2} are routinely introduced during the fabrication of a MOS device, and even lower and higher doses are possible. To increase the dose simply requires a longer implant time or a higher beam current. This allows subtle tailoring of implant doses for such critical steps as the threshold adjusting implant in a MOS device or the base dose in a bipolar device. This and other applications were illustrated in the CMOS process flow in Chapter 2. For example, Figures 2–22 and 2–23 illustrated threshold adjusting implants.

In spite of the precision with which the dose can be controlled, at its heart ion implantation is a random process, because each ion follows a random trajectory, scattering off the lattice silicon atoms before losing its energy and coming to rest at some location as illustrated in Figure 8–1. The reason the technique can be used successfully is because large numbers of ions are implanted so an average depth for the implanted dopants can be calculated.

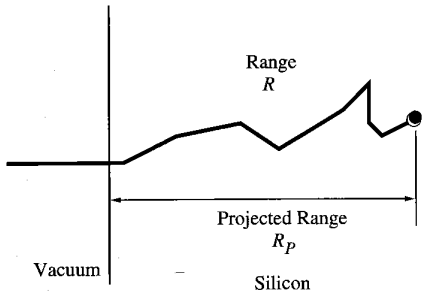


Figure 8–1 Schematic of the actual range of an implanted ion and the projected range normal to the surface. The projected range describes the peak of the implanted profile.

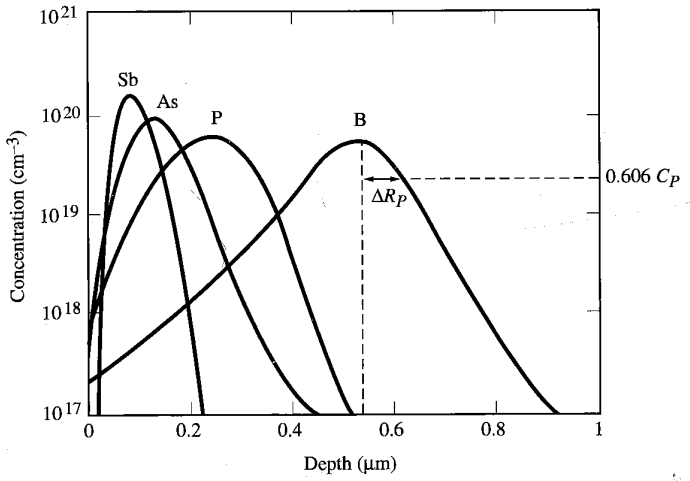


Figure 8–2 Distribution of ions implanted into crystalline silicon at an energy of 200 keV. The light ions travel further and have a broader distribution than the heavy ions.

The first step is to investigate the distribution of ions implanted at a given energy. Heavy ions like antimony do not travel as far in the crystal as light ions like boron. If different ions are implanted with the same energy, the heavy ions stop at a shallower depth, as shown in Figure 8–2. The projected range R_P depends on the energy that is used for the implant, with higher energies giving a deeper range. While each implanted ion itself follows a random trajectory with a range R, on average the distribution of a large group of ions will peak at a projected depth R_P below the surface of the wafer. Because of the random nature of the process, some ions will stop sooner because of more collisions, while some will travel further. The spread of the ions depends on the range traveled with deeper ranges allowing for more random stopping events. This gives rise to a distribution of ions where most of the ions are within a standard deviation $\pm\Delta R_P$ of the projected range. The heavy ions with the smaller range have a more narrow distribution than the light ions.

Since the number of ions implanted is usually greater than 10^{12} cm^{-2}, the distribution can be described statistically and is often modeled to first order by a symmetric Gaussian distribution given by

$$C(x) = C_p \exp\left(- \frac{(x - R_P)^2}{2\Delta R_P^2} \right) \tag{8.1}$$

where R_P is the average projected range normal to the surface, ΔR_P is the standard deviation or straggle about that range, and C_p is the peak concentration where the Gaussian is centered. Figure 8–3 plots the range and standard deviation for common dopants in silicon. Note that the profiles in Figure 8–2 are not perfectly symmetrical. We will return to these details later.

The total number of ions implanted is defined as the dose and is simply

$$Q = \int_{-\infty}^{\infty} C(x)dx \tag{8.2}$$

Making use of the fact that the sum (or integral) of Gaussian functions is an error function (Chapter 7), and using the formula which defines the error function gives

$$\int_{-\infty}^{\infty} \exp^{-u^2} du = \frac{\sqrt{\pi}}{2}[\text{erf}(\infty) - \text{erf}(-\infty)] \tag{8.3}$$

so that

$$Q = \sqrt{2\pi}\Delta R_P C_P \tag{8.4}$$

This provides a useful relationship between the dose and the peak concentration of the implant. The range and standard deviation for typical dopants in silicon have been measured and tabulated.

So far, we have considered a blanket implant over the whole silicon surface. It is interesting to think about what the implant profile would look like if we left the beam centered on a particular spot (x,y,z = 0,0,0). The distribution of 1000 ions implanted at a random position in a silicon unit cell centered at (0,0,0) is shown in Figure 8–4.

The side view shows the depth distribution of the ions which is approximated by Eq. (8.1), with a projected range of 50 nm and a straggle about the projected range of 20 nm. The beam direction view shows that the ions scatter laterally around the impact point centered at (y,z = 0,0). This lateral distribution around the peak can also be described statistically by a Gaussian distribution with a lateral straggle $\Delta R_\perp$ replacing the vertical straggle. Most of the time we are not concerned with implanting at one spot on a wafer, but we are very interested in implanting in a window with the surrounding areas masked from the implant. The two-dimensional projection near the window edge is of interest, because it indicates how many of the ions scatter under the window because of the lateral straggle. Because it is very difficult to experimentally measure lateral dopant profiles, the two-dimensional distribution is often assumed to be composed of just the product of the vertical and lateral distributions, so that

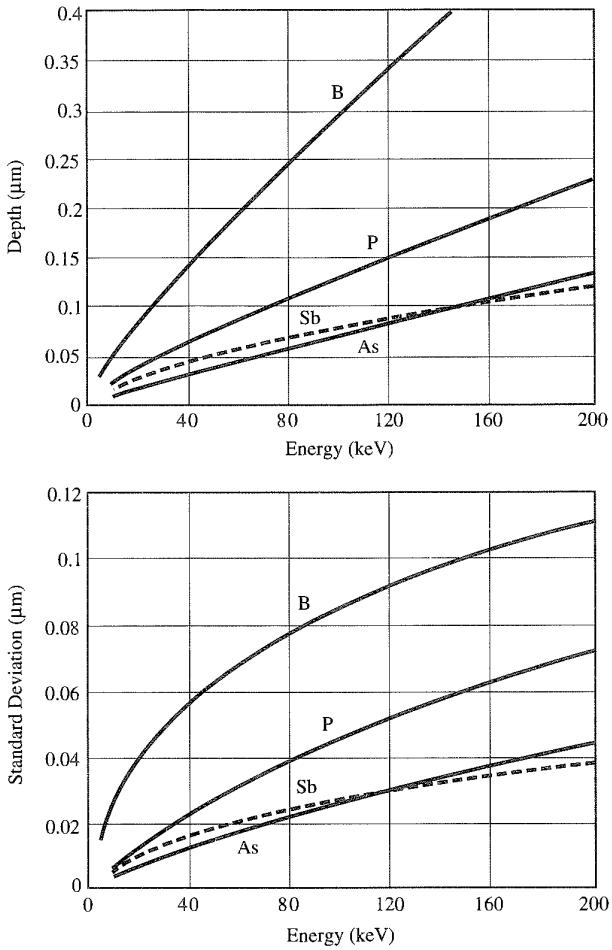


Figure 8–3 Plot of the range R_P (top) and standard deviation ΔR_P (bottom) of common dopants in crystalline silicon tilted and rotated to simulate a random direction. (After [8.1].)

$$C(x,y) = C_{\text{vert}}(x) \exp\left(-\frac{y^2}{2\Delta R_\perp^2}\right) \tag{8.5}$$

Any layer thick enough to capture the implanted ions can be used as a masking layer. For example, photoresist is often used as a convenient mask, because implants are performed at room temperature. Near a mask edge, the profile is dominated by the lateral straggle and is given by a sum of point response functions with the Gaussian form of Eq. (8.5). This means that the profile under a mask edge, such as under the gate

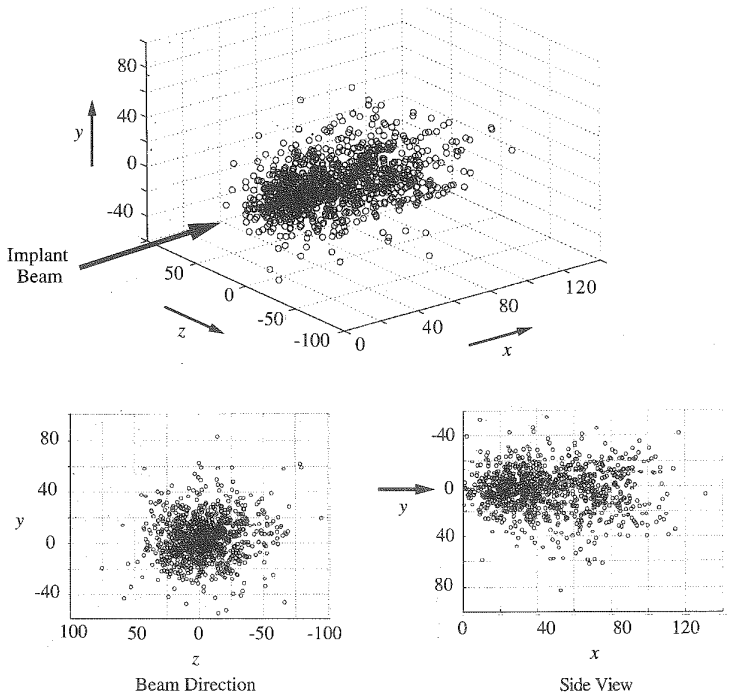


Figure 8–4 Monte Carlo simulations using UT-Marlowe [8.2] of the 3D distribution of 1000 phosphorus ions implanted at 35 keV. The 2D projection (side view) and top view are also shown. The implant beam is centered at $(y,z = 0,0)$ and the axes are in units of nm.

of a MOS transistor, can be dominated by the lateral straggle of the implant and by how far this moves during subsequent annealing. With decreasing device geometries, the implant straggle under the mask edge is becoming a more important feature in modern MOS devices.

How thick does a masking layer have to be to block the transmission of ions through the mask? The range can be measured in different materials and denser materials have better masking properties. The range and standard deviation are very similar in silicon and silicon dioxide, while a thin layer of metal or a thick photoresist layer could efficiently mask an implant. In order to act as an efficient mask, the thickness of the mask should be large enough that the tail of the implant profile in the silicon is at some specified background concentration as illustrated in Figure 8–5. We will use a superscript * to identify the ranges and standard deviations in the masking material, since they will in general be different from those in the silicon. Then, the criterion for efficient masking is that

$$C^*(x_m) = C_P^* exp\left(-\frac{(x_m - R_P^*)^2}{2\Delta R_P^{*2}}\right) \leq C_B \tag{8.6}$$

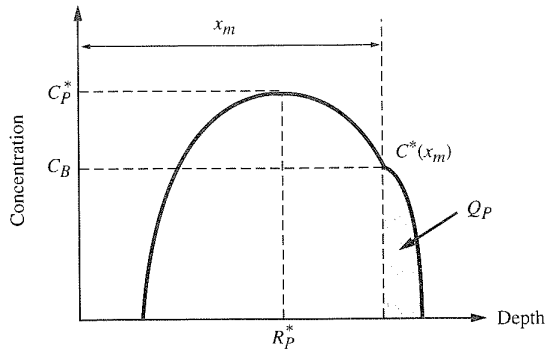


Figure 8–5 Schematic of a masking process, where a dose Q_P penetrates the mask of thickness x_m.

where $C^*(x_m)$ is the concentration at the far side of a mask of thickness x_m and C_B is the background concentration in the substrate. Setting $C^*(x_m) = C_B$ and solving for the mask thickness gives

$$x_m = R_P^* + \Delta R_P^* \sqrt{2\ln\left(\frac{C_P^*}{C_B}\right)} = R_P^* + m\Delta R_P^* \tag{8.7}$$

where the parameter m indicates that the mask thickness should be equal to the range plus some multiple m times the standard deviation in the masking material. Values of m for different levels of masking efficiency can easily be calculated from this equation.

If Q_P is the dose that penetrates the mask (the dashed area in Figure 8–5), then

$$Q_P = \frac{Q}{\sqrt{2\pi}\Delta R_P^*} \int_{x_m}^{\infty} \exp - \left[\frac{x - R_P^*}{\sqrt{2}\,\Delta R_P^*}\right]^2 dx \tag{8.8}$$

Since the integral or sum of Gaussian functions can be described as an error function, where

$$\int_d^{\infty} \exp(-u^2)du = \frac{\sqrt{\pi}}{2}\operatorname{erfc}(d) \tag{8.9}$$

the dose that penetrates is

$$Q_P = \frac{Q}{2}\operatorname{erfc}\left(\frac{x_m - R_P^*}{\sqrt{2}\,\Delta R_P^*}\right) \tag{8.10}$$

One case of practical interest which involves both of the previous calculations is the case where the mask edge is tapered or the implant is performed at a tilted angle. In these cases, the mask is not perfectly efficient and the distribution near the mask edge is more complex. An experimental result of etch delineated dopant contours under the

gate of a MOS device was shown in Figure 7–21. Note in that figure that the dopant profile extends under the edge of the polysilicon gate.

Example A process engineer building an NMOS device wants to dope the polysilicon at the same time as doing the arsenic source/drain diffusion. The source drain implant dose is 2×10^{15} cm^{-2} at an energy of 50 keV. (a) For the above implant conditions, what is the minimum polysilicon thickness that can be used if the implant is not to affect the channel doping which is 1×10^{16} cm^{-3} near the surface? (Assume the gate oxide is negligibly thick compared to the polysilicon.) (b) Assuming that this polysilicon thickness is actually used, how much of the implant dose will penetrate the polysilicon mask if the process engineer decides to change the implant energy to 80 keV?

Answer (a) From Figure 8–3, 50-keV arsenic has a range of 35 nm and a standard deviation of 15 nm, so the implant profile in the polysilicon is equal to the channel profile when

$$C(x_m) = 1 \times 10^{16} = \frac{Q}{\sqrt{2\pi}\Delta R_P} \exp\left[- \frac{(x_m - R_P)^2}{2\Delta R_P^2} \right]$$

$$C(x_m) = 1 \times 10^{16} = \frac{2 \times 10^{15}}{\sqrt{2\pi}(150 \times 10^{-8})} \exp\left[- \frac{(x_m - (350 \times 10^{-8}))^2}{2(150 \times 10^{-8})^2} \right]$$

$$\therefore x_m = 10.5 \times 10^{-6} \text{ cm} = 0.105 \mu\text{m}$$

(b) If the implant energy is raised to 80 keV, the range from Figure 8–3 is 55 nm and the standard deviation on the implant is 22.5 nm. The dose that penetrates the mask of thickness $x_m = 0.105$ μm is

$$Q_P = \frac{Q}{2} \text{erfc}\left[\frac{x_m - R_P^*}{\sqrt{2}\Delta R_P^*} \right]$$

$$Q_P = \frac{2 \times 10^{15}}{2} \text{erfc}\left[\frac{0.105 \times 10^{-4} - 550 \times 10^{-8}}{\sqrt{2} \times (225 \times 10^{-8})} \right]$$

$$= 1 \times 10^{15}(1 - \text{erf}(1.571)) = 2.7 \times 10^{13} \text{ cm}^{-2}$$

In the general case in which the mask edge is not vertical or an angled implant is performed, numerical methods must be used to calculate the 2D doping profile. There may be reasons to use a high-angle implant to try to introduce dopants underneath the gate in a MOS device. For example, in order to minimize short-channel effects in small

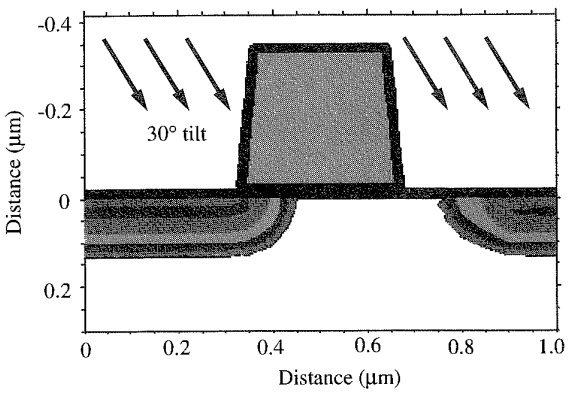


Figure 8–6 TSUPREM IV [8.1] simulation of a 50-keV phosphorus implant at a high-tilt angle of 30°, showing the asymmetrical implant distribution and the shadowing cause by the gate polysilicon.

devices, some doping may be introduced below the tip extension region under the gate. This "halo" of doping is formed by a high-tilt implant under the edge of the gate. Tilted implants can cause shadowing effects due to the topography already on the wafer. A numerical simulation of a tilted implant near a mask edge is shown in Figure 8–6, which clearly shows the asymmetric distribution and the shadowing caused by the gate polysilicon thickness. To get a symmetrical distribution on all devices, the wafer is often rotated during the implant or is implanted at four separate rotations. To completely avoid shadowing effects requires that the implant be performed at a zero-tilt angle.

To understand how these implanted profiles evolve in time during subsequent annealing, we can compare the Gaussian formulation from the implant distribution with that from a delta-function distribution that has diffused. The solutions for both Gaussian distributions in a semi-infinite medium are

$$C(x) = C_p \exp\left(-\frac{(x - R_p)^2}{2\Delta R_p^2}\right) \quad \Leftrightarrow \quad C(x) = C(0)\exp\left(-\frac{x^2}{4Dt}\right) \tag{8.11}$$

Implanted Diffused

By comparing these solutions, we can see that an implanted Gaussian profile with a standard deviation ΔR_p has the same form as an initial delta-function distribution that has diffused for an effective time-temperature cycle of $\Delta R_p = \sqrt{2Dt}$. Because we simply add successive Dt cycles to predict how a diffused profile continues to evolve with time (Section 7.2.9 in Chapter 7), we can easily consider the effect of additional time-temperature cycles on the implanted Gaussian distribution by using the formula

$$C(x,t) = \frac{Q}{\sqrt{2\pi(\Delta R_p^2 + 2Dt)}} \exp\left(-\frac{(x - R_p)^2}{2(\Delta R_p^2 + 2Dt)}\right) \tag{8.12}$$

From this, it is clear that a Gaussian remains a Gaussian and that it preserves its shape upon further annealing in an infinite medium although its standard deviation or straggle about the peak concentration increases with the diffusion distance as shown in Figure 8–7.

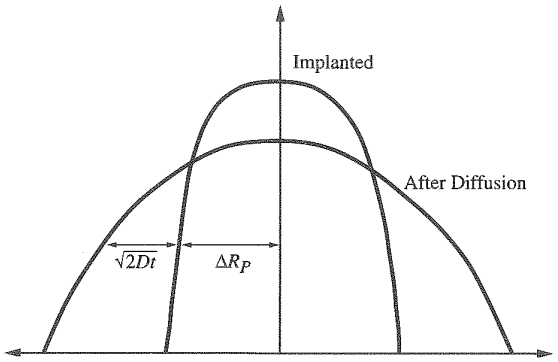


Figure 8–7 Evolution of a Gaussian profile after annealing. The Gaussian preserves its shape as it diffuses in an infinite medium.

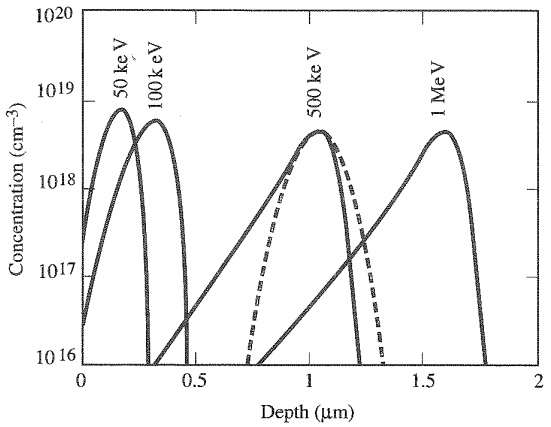


Figure 8–8 TSUPREM IV [8.1] simulated profiles of boron implanted into amorphous or polycrystalline silicon at a range of energies. The profiles show a noticeable skew toward the surface because of backscattering of the light boron ions. These skewed profiles are described by a four-moment Pearson distribution. A Gaussian fit to the profile at 500 keV is shown as a dashed line and only matches the peak of the profile. (After [8.3].)

While these Gaussian distributions are very simple mathematically, we find experimentally that they often only match the central (peak) region of the implanted profiles, while there are significant deviations in the tails of the distributions. For example, in Figure 8–8, the boron distribution is skewed toward the surface and a Gaussian profile provides a good fit only near the peak of the distribution.

We can understand this physically by considering how light ions like boron scatter off silicon atoms compared with how heavy atoms like arsenic scatter off silicon atoms. There is a greater tendency for light ions to backscatter and fill in the front side of the distribution and vice versa for heavy ions. Figure 8–9 shows how dramatic this effect can be for light and heavy ions implanted into amorphous silicon. Clearly, to accurately describe implanted profiles, higher-order moments other than the range and standard deviation are needed.

An arbitrary distribution can be described by a series of moments. The normalized first moment is the projected range and is given by

$$R_p = \frac{1}{Q}\int_{-\infty}^{\infty} xC(x)dx \tag{8.13}$$

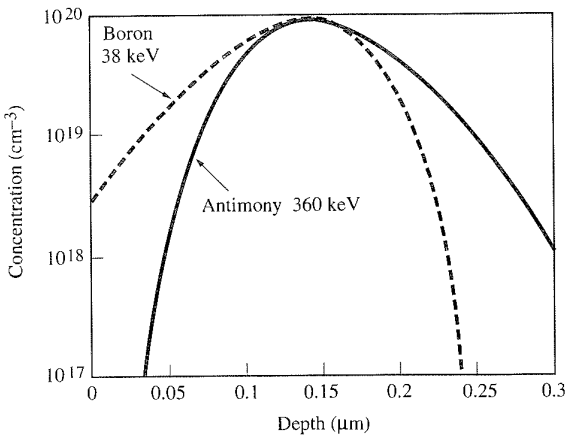


Figure 8–9 TSUPREM IV simulation [8.1] of ion-implanted boron and antimony profiles with the implant energy adjusted to give the same range. The light boron ion skews toward the surface while the heavy antimony ion skews toward the bulk.

The second moment is the straggle or standard deviation

$$\Delta R_p = \sqrt{\frac{1}{Q}\int_{-\infty}^{\infty} (x - R_p)^2 \, C(x)dx} \tag{8.14}$$

The third moment describes the skewness and is given by

$$\gamma = \frac{\int_{-\infty}^{\infty} (x - R_p)^3 C(x)dx}{Q\Delta R_p^3} \tag{8.15}$$

and the fourth moment is the kurtosis

$$\beta = \frac{\int_{-\infty}^{\infty} (x - R_p)^4 C(x)dx}{Q\Delta R_p^4} \tag{8.16}$$

These four moments can be used to get a more accurate description of a wide range of implanted species in silicon. The integral representation of the moments is not particularly convenient, so in practice the equivalent moments are described by coefficients in a functional form known as Pearson's equation. Tabulated values of the moments can then be used to generate a profile. A plot of the first two moments for the common dopants was shown in Figure 8–3.

At this stage, we have an accurate description of implanted profiles in an amorphous or random silicon matrix, perhaps capped with an amorphous masking material like oxide or nitride. This provides a useful description for implants in amorphous silicon and fine-grain polycrystalline silicon. The description can also be applied to crystalline silicon if care is taken to make the substrate appear amorphous by tilting and rotating the

wafer. However, the level of sophistication needed in describing implant profiles in most devices requires that we consider the atomic-scale structure of the silicon lattice to accurately describe profiles.

8.2.1 Implants in Real Silicon—The Role of the Crystal Structure

One crucial property of real silicon wafers that we have ignored is the crystalline symmetry of the lattice which leads to planar and axial channels that can have quite dramatic effects on the implant profile. Images from a silicon crystal viewed in different directions are shown in Figure 8–10, showing some of the different channels that exist and including a "random" orientation where the crystal appears amorphous. Even a wafer tilted and rotated to appear amorphous, as in Figure 8–10, can allow a scattered ion to enter the channels which, of course, continue to exist no matter how we rotate the crystal in space.

Most of the scattering events that do occur are through relatively small angles, so that a tilt and rotation can minimize the number of ions that immediately find the relatively open channels in the silicon lattice. However, once an ion does enter a channel, the small-angle scattering events from the atoms that line the walls of the channel mean that the ion can be steered quite a long distance along the channel before coming to rest from the electronic drag forces or from a sharp collision that causes the ion to exit the channel. The effect of channeling on the implant profile is to cause a tail that continues

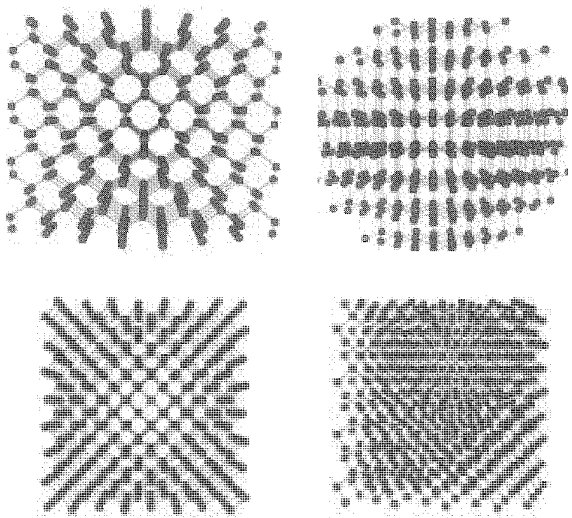


Figure 8–10 Image of a silicon crystal looking down the 110 axial channels (top left), the 111 planar channels (top right), the 100 axial channels (lower left), and with a tilt and rotation to simulate a "random" direction (lower right).

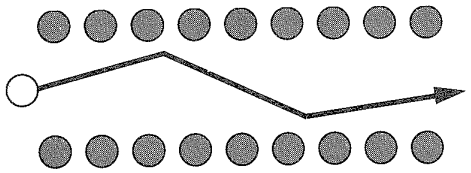


Figure 8–11 Schematic of a channeling direction in silicon where an ion undergoes many small-angle deflections.

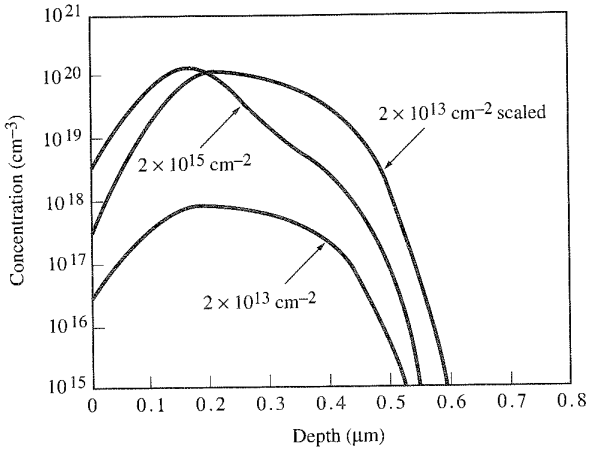


Figure 8–12 TSUPREM IV simulations [8.1] of boron profiles implanted into <100> crystalline silicon wafers with zero tilt and rotation. The implant energy is 35 keV. Note the deep channeling tails in crystalline silicon, which depend on the implant dose, in marked contrast with Figure 8–8 (After [8.4] and C.S. Rafferty [private communication].)

much further than expected. Small-angle scattering from silicon channels is depicted in Figure 8–11.

If we examine the channeled profiles in Figure 8–12, they appear to be composed of two portions, one corresponding to the main peak and the other corresponding to the channeled part. Note that the origin of this skewed tail is different from the skewness caused by heavy ions being implanted in amorphous silicon. In the case of crystalline silicon, it is due to some fraction of the ions entering channels in the silicon lattice and traveling further because of fewer collisions with lattice atoms. However, it has a similar effect in that it modifies the profile so that it requires higher-order moments to fit it. The effects of channeling in Figure 8–12 are so severe that two distributions are required to fit the profiles—one distribution to describe the main part of the profile and another to describe the channeled part of the profile. This leads to a dual-Pearson approach to modeling implant profiles in crystalline silicon, with an additional parameter to specify the ratio of the dose in the main and channeled profiles. A further parameter is needed to describe the dose dependence that is present in the profiles. As shown in

Figure 8–12, simply scaling up a low-dose implant does not give the same profile as a high-dose implant. As the crystal gets damaged by the implanted ions, the channels are less evident to subsequent incoming ions. Relatively speaking, higher implanted doses are less prone to channeling effects.

To attempt to minimize channeling, a thin screen oxide which is amorphous is often used, causing some randomization of the incident beam before it enters the lattice. For example, this was done in the implants used in the CMOS process in Chapter 2. A complete description of the implanted profile requires that the effect of the thickness of the screen oxide on the Pearson parameters be included. Finally, this functional form should be tabulated for each tilt and rotation angle that is likely to be used, because the distributions in the channels clearly vary depending on the orientation of the wafer.

Though tabulating all possible parameters that affect implant profiles is a tedious task, it does provide simulation programs with the ability to use a look-up table of coefficients and to estimate the shape of a profile for various implant conditions. Because of extensive characterization using SIMS, the implanted profiles can be well modeled in crystalline silicon for blanket implants.

Often however, the two-dimensional distribution of the ions near a mask edge is critical for device performance. To illustrate the complexity of this task, several different implant models are compared in Figure 8–13. The profile is shown close to a polysilicon gate edge coated with a thin liner oxide (similar to the structure in Figure 7–21). Three of the models use look-up tables of varying complexity, while the fourth model takes a more atomistic view of the implant process. The disagreement between the phenomenological models is perhaps expected, as they are usually calibrated to one-dimensional SIMS data and the lateral straggle is arbitrarily taken to be some fraction of the vertical straggle. Implanting near a mask edge may cause an ion to travel through multiple different layers (thin spacer, polysilicon, gate oxide) before ever entering the silicon and all of these affect the ion's trajectory, particularly in the lateral direction. The biggest problem in verifying which model is correct is that two-dimensional measurements of doping profiles are very difficult, although one-dimensional measurements in large areas using SIMS are easy. Recent developments in 2D dopant profiling such as scanning capacitance microscopy (Figure 7–23) will help to resolve these issues however.

It is at this stage that the phenomenological description of the implanted profiles becomes insufficient to describe many of the effects that are crucial for VLSI processes. Of particular importance is how well the lateral profile under a mask edge is modeled by these ad hoc descriptions. In order to achieve an adequate description, an atomic-scale view of the implant process will be discussed in the modeling section.

8.3 Manufacturing Methods and Equipment

Conventional ion implanters have much in common with linear accelerators. Indeed, the early history of ion implantation made much progress when research in nuclear physics shifted to such high energies that laboratory machines were no longer useful and were instead used to investigate the relatively low energies of ion-solid interactions that are useful for ion implantation. A schematic of an ion implanter is shown in Figure 8–14.

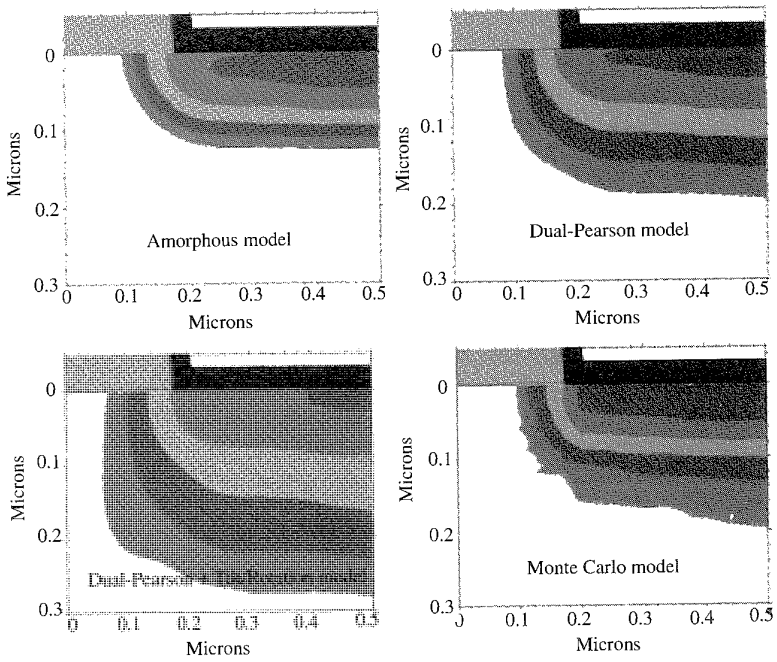


Figure 8–13 TSUPREM IV [8.1] implant simulations using four different implant models of varying complexity. The implant contours are shown near a polysilicon gate edge (thick enough to block the implant) coated with a thin oxide layer. Only the Monte Carlo model considers the atomic structure of the silicon lattice in detail, a factor which is lacking in the other models which use tabulated coefficients for Gaussian, dual Pearson, or dual Pearson with a tilt/rotation dependence.

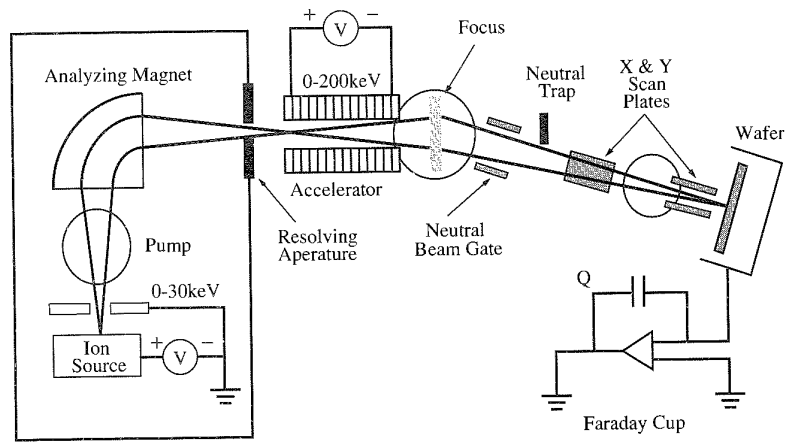


Figure 8–14 Schematic of an ion implanter.

The basic requirement is a source of ions of sufficiently high density to be useful. Either a solid source that is vaporized or a gas source is conventionally used to deliver the material to the ion implanter. Arsine, phosphine, and BF_2 gas sources are extremely toxic and have historically been used in dilute mixtures (15% in H_2) in high-pressure (> 400 psi) cylinders. Because of safety concerns, solid sources of elemental arsenic or phosphorus are sometimes preferred. The advantage of a solid feed source is that any element can be vaporized and implanted. New gas source systems use a zeolite matrix which acts as a molecular sieve to adsorb and store the gas in cylinders below atmospheric pressure, neatly avoiding the possibility of high-pressure rupture and release. Even a catastrophic failure releases gas at a rate driven only by molecular diffusion. These gaseous systems allow rapid beam tuning, which is important if multiple implants are to be performed through a single mask.

The gas from the feed source is ionized by energetic electrons boiled off a hot filament or by a plasma discharge. The ions are extracted by a voltage bias on a grid and mass analyzed to select only one ion species and perhaps even a single isotope of an ion species. This mass analysis is necessary because a gas like BF_2, for example, will dissociate into many ions such as B^{++}, B^+, F^+, BF^+, and BF_2^+ for both the boron mass-10 and mass-11 isotopes. The mass analysis relies on balancing the force exerted on the charged ions in a magnetic field with their centrifugal force, so that the path of the ion is given by the solution of the equation

$$\frac{mv^2}{R} = q \cdot \vec{v} \times \vec{B} \tag{8.17}$$

where m is the mass of the ion, q is the charge on the ion, v is the ion velocity, R is the radius of curvature, and B is the magnetic field intensity. The magnetic field intensity depends directly on the current in the electromagnet coils, so that $B = \alpha I$ where α depends on the magnet design. Because the magnetic field is designed to be perpendicular to the ion velocity, $|\vec{v} \times \vec{B}| = vB$. The ion velocity is related to the extraction voltage by

$$v = \sqrt{\frac{2E}{m}} = \sqrt{\frac{2qV_{ext}}{m}} \tag{8.18}$$

The field can be tuned to allow an ion of particular mass m to exactly follow an arc of radius r and exit the analyzer through a narrow slit. Using the above equations, we can show that

$$\sqrt{m} = \frac{q}{\sqrt{2E}}(\alpha r)I \tag{8.19}$$

For a fixed magnet design and fixed position of the resolving aperture in a machine, this indicates that tuning the analyzer magnet current will cause different mass ions to be selected depending on the square root of their mass. This is the reason that very heavy isotopes are difficult to separate using electromagnetic methods, though it is not a severe problem for most of the species used in semiconductor fabrication. In ultralow energy

machines no further acceleration of the selected ions takes place and they continue toward the sample at the extraction energy.

For most situations, the selected ions are further accelerated in a small linear accelerator to the final energy of implantation. Some beam neutralization may occur during this acceleration phase and this is a problem because neutral atoms cannot be electrostatically scanned or counted as ion current. They implant an uncontrolled dose in the center of the wafer. For this reason, the ion path typically undergoes an electrostatic deviation from the linear path just before final implantation, which acts to trap the neutrals which continue undeflected. This bend in the column is often noticeable near the end station. The ion beam may then undergo an *x-y* electrostatic deflection which scans the beam onto the wafer, perhaps combined with a mechanical translation and scan of the wafer holder.

The implant dose is measured by locating the sample at the end of a deep Faraday cup, which collects the current and integrates it over time. The normalized dose is given by

$$Q = \frac{1}{A}\int \frac{I}{q}dt \tag{8.20}$$

where I is the collected beam current, A is the implant area, t is the integration time, and q is the charge on the ion.

8.3.1 High-Energy Implants

While most implanters cover a range of 30 keV to 200 keV for singly ionized species, this imposes practical limits on the depth that dopant atoms can be introduced into the silicon. The upper energy limit is determined by breakdown in systems with air insulation. The effective ion energy can be extended by using doubly charged species, but the beam current is lower since fewer doubly charged species are generated in the source. However, there are particular applications where implant energies in the MeV range are useful, such as for the deep wells in a CMOS technology. Implanting the well allows cost savings because multiple sequential implants can be performed through the same mask, allowing an optimized profile for the deep well, a shallower punchthrough implant, and a threshold voltage adjust implant to be performed in the same load step and pump-down. Because of the nature of ion implantation, it is also possible to produce a retrograde well profile with higher doping deep in the substrate and low doping at the surface, something that is impossible to do using diffused wells. For example in the P and N wells produced in the CMOS process in Chapter 2, a high-temperature diffusion was used after the implants (Figure 2–12). This would result in wells in which the highest dopant concentrations are near the surface.

We can examine the impact of introducing high-energy implantation technology in a standard CMOS process. One of the major problems with CMOS devices is the susceptibility to latchup, whereby a parasitic n-p-n-p thyristor latches in the on state, destroying the circuit operation. Figure 8–15 illustrates this problem. The CMOS structure inherently contains parasitic NPN and PNP bipolar transistors. Under normal CMOS

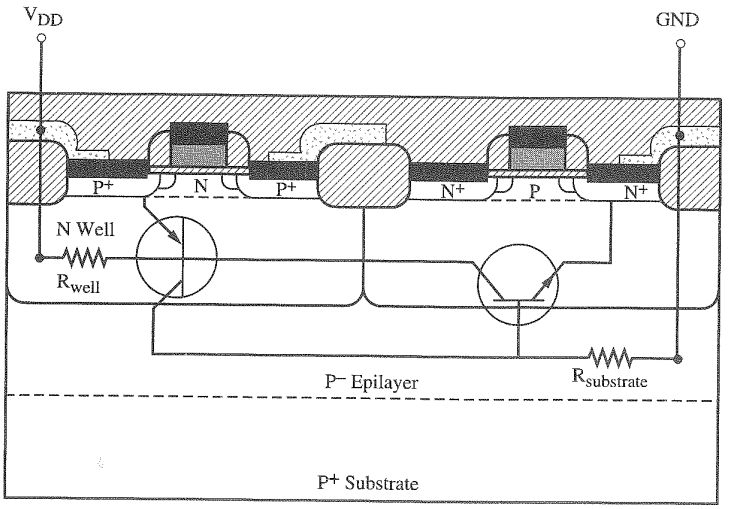


Figure 8–15 Schematic of latchup structure which occurs in a CMOS technology. The parasitic bipolar transistors are overlaid on the technology discussed in Chapter 2. The resistors are normally connected as shown in the schematic.

circuit operation, these devices are not active since $V_{BE} = 0$ volts. However under transient conditions, IR drops in the parasitic resistors shown in Figure 8–15 can become large enough (≈ 0.7 volts) to turn on one or both bipolar devices since voltage drops across these resistors also appear across the bipolar base emitter diodes. If this happens, the bipolar devices can behave like a thyristor because they are inherently wired in a cross-coupled configuration. Once triggered, the thyristor structure can carry large currents, shorting the supply voltage to ground and causing catastrophic failure of the chip.

This problem can be solved for all practical purposes by building CMOS devices in an epitaxial layer above a highly doped substrate, using the high conductivity of the substrate to reduce the parasitic resistances in the critical parasitic path. But epitaxial wafers are expensive and there is a possibility of using high-energy MeV implants to introduce a similar buried layer shunt path in a bulk wafer. This requires a relatively high dose (above 10^{14} cm^{-2}) but could be used to replace epitaxial wafers.

The more immediate use of high-energy implants is to simplify the well-formation process in CMOS while simultaneously decoupling the isolation design from the transistor design. A typical sequence of implants includes all of the implants for the n well such as a deep well, a punchthrough implant, and a threshold adjust implant, using thick (2-micron) photoresist as the implant mask. Note that the high-energy implants can be performed after the isolation regions have been formed. This can provide some advantages in terms of reduced thermal cycles when the implants are performed after the field oxide formation or the shallow trench isolation step. The p-well regions can subsequently be defined and a similar series of implants performed to optimize the various

dopant profiles. The thick photoresist provides an adequate mask but is often not hard-baked (Chapter 5) to avoid photoresist flow and degradation of the sidewall angle. The masking qualities of the resist can be measured by doing implants in the resists and subsequently measuring the implanted profiles using SIMS analysis.

It is likely that MeV implant technology will be increasingly used to reduce the number of process steps associated with forming the wells and for the ability to optimize each sequential implant, thereby decoupling the isolation design and the transistor design.

8.3.2 Ultralow Energy Implants

As seen from the technology roadmap (Table 7–1 in Chapter 7), the ability to form junctions shallower than 100 nm is needed to continue CMOS scaling. Forming such shallow junctions is not an inherently difficult problem—for example, energies below 100 eV have been used for doping shallow layers grown by Molecular Beam Epitaxy (MBE). At this low energy, the ions are not implanted into the material but land softly onto the growing surface where they are buried by the incoming molecules from the MBE evaporation source. The real question for this chapter is whether it is possible to form shallow junctions by ion implantation in an economically viable manner, allowing an extension of current technology to future technology generations.

The reason most current ion implanters cover an energy range above 30 keV is because the low-energy limit is set by the extraction voltage for the ions from the source plasma. In order to get sufficient beam current, a reasonably large extraction voltage has traditionally been used. The beam current determines the throughput and hence the economic viability of using ion implantation in large-scale production and needs to be in the range of several milliamps.

The source current available from a typical plasma source is not a limiting factor on the ion flux that is available. Instead, it is the space charge in the beam that limits the beam current at low energies. At some current level, the amount of charge in transit shields the source from the applied field, limiting the current in the beam. The space charge limited ion current density is given by Child's Law:

$$J \propto V_{\text{ext}}^{1.5} d^{-2} \tag{8.21}$$

where V_{ext} is the extraction voltage and d is the distance over which the voltage is applied. Thus, the available beam current scales as the extraction voltage $V^{1.5}$, so the beam current at 2 keV is only one-quarter of that available at 5 keV. Careful attention to the ion optics and the design of the extraction electrodes has enabled useful currents to be obtained at low energies, within this scaling limitation.

Since the extracted ions are often accelerated to higher energies, why is it not possible to simply decelerate the beam? Deceleration can produce higher currents at low energies. This particular mode of operation is sometimes used (decel mode) but may cause subtle problems in manufacturing. Because charge exchange between neutrals and ions in the beam may occur continuously, some of the neutrals will not be decelerated and there can be a variable energy spectrum leading to a deeper than expected

profile which directly affects the junction depth. This means that extraordinary care has to be taken with the machine design. However, the latest generations of optimized low-energy machines have beam currents that are comparable with those achieved in decel mode, without the problems of beam instabilities. These machines operate in a drift mode, where ions are extracted from the source at the final energy required, mass analyzed, after which the ions drift toward the sample. Implants at energies less than 250 eV can be performed, suggesting that there is no lower limit on the economical throughput for ion implantation in shallow junction formation. The absolute lower energy is determined by the energy needed to penetrate the lattice and is approximately 50 eV for boron. High doses cannot be introduced at extremely low energies, because the incoming ions sputter off some surface atoms, leading to a self-limiting dose. For very low energies, the profiles appear to be still affected by Transient Enhanced Diffusion (TED), discussed in Section 8.5.9. The final junction may need careful annealing to minimize TED, but junctions a few tens of nanometers deep are still within reach. Similar {311} defects (described later) are seen, as in higher-energy implants, though with a preferred orientation parallel to the surface, suggesting that the TED kinetics will continue to be controlled by evaporation of interstitials from clustered defects. It appears that there is no unusual behavior for ultralow energy implants, other than enhanced interactions with the surface. We will return to many of these issues later in the chapter.

8.3.3 Ion Beam Heating

The temperature rise in the silicon caused by ion implantation is determined by the dose, acceleration voltage, and thickness of the silicon wafer. The energy deposited is

$$E_{\text{dep}} = V \int I dt = VQ \tag{8.22}$$

One important limiting temperature regime is at about 120°C, where photoresists which are used for masking begin to flow, crack, and deteriorate. For many normal dose-energy combinations even an uncooled wafer will not reach this temperature.

8.4 Measurement Methods

Since ion implantation is used to provide doped regions in silicon, the measurement methods associated with this process are generally designed to measure dopant profiles. Thus the same methods as those described in Chapter 7 are used. Total chemical concentration profiles are generally measured with SIMS. Electrically active profiles are measured with spreading resistance, sheet resistance, or CV methods. The more advanced methods described in Chapter 7 (cross sectional TEM, scanning capacitance microscopy, and inverse electrical measurements) can also be used to characterize implanted profiles in 2D.

Because the ion implantation process creates damage, most dopants are not electrically active at the end of the implantation. Thus profile measurement at this point must carefully distinguish between total chemical concentrations and the portion of the dopant that is electrically active. Following an appropriate activation anneal, most of the dopant is electrically active, unless the concentrations are so high that precipitation, clustering, or other processes occur.

8.5 Models and Simulations

While the phenomenological description of ion distributions in silicon in terms of distribution functions is useful, a complete description of the detailed atomic-scale events that take place along an ion trajectory has proven to be the only viable method for accurately modeling the final ion distribution in crystalline silicon. Much of the theoretical description of the stopping of charged ions by matter came from original work on the structure of the nucleus. Following Ziegler [8.5], we provide a brief review of some of the interesting historical highlights.

Rutherford, in 1911, observed that about 0.01% of alpha particles were backscattered from thin aluminum foils and proceeded to show that these came from a single collision with a positively charged nucleus, thereby altering the prevailing view of the nature of matter. In the terminology of ion implantation, the alpha particle is just a double ionized helium atom, though the source of Rutherford's particles came from radioactive decay. Soon after, Bohr attempted one of the first unified theories of ion stopping by matter and made the hypothesis that the energy loss of ions could be divided into two components: the nuclear energy loss to the positive atomic cores of the target and the electronic energy loss to the free electrons in the target. The major unknown in calculating the energy loss was the effective charge state of a moving ion in the target. Bohr assumed that the ion's electrons with an orbital velocity less than the ion velocity would be stripped off, leaving a highly charged particle interacting with the orbiting target electrons which could be treated as harmonic oscillators. Bethe and Bloch later introduced a quantum mechanical treatment of the electronic energy loss of a charged particle to a free electron gas, while Fermi considered how the charged particle would polarize the electron gas and alter the interaction. Fermi and Teller extended this work and found that the electronic energy loss would be directly proportional to the particle's velocity. The nuclear energy loss was treated as an elastic collision of the charged particle with the Coulomb field of a positively charged nucleus, partially screened by the outer valence electrons.

The culmination of this work by Lindhard, Scharff, and Schiott in 1963 in what has become known as LSS range theory brought all the pieces together and allowed the stopping powers to be calculated for any arbitrary atomic species and elemental targets within a factor of two. The improvements in range calculations since the LLS theory have come about from considering the shell structure of the atoms using quantum mechanical atomic structure calculations, by numerically calculating the path of the ion trajectory during interactions with a target atom rather than making asymptotic collision

approximations, by improved effective charge approximations [8.6], and by considering the lattice structure of crystalline materials.

Computers are now powerful enough to literally follow the atomic-scale trajectory of a representative sample of a few thousand ions through the lattice, accurately accounting for an arbitrary screen oxide thickness together with any tilt and rotation used during implantation. By combining the information on the final resting positions of the ions, a picture of the final distribution can be obtained. Because of the random nature of the starting condition, this approach is often called a Monte Carlo simulation. In amorphous silicon, the position of the silicon atoms is also randomly chosen so the overall density is correct. Given the atomic density for the target, the average distance between atoms in an amorphous target is $(1/N)^{1/3}$, which gives the average path length between nuclear interactions in amorphous material.

In crystalline silicon, the trajectory is completely determined by the structure of the lattice after the ion is set in motion. The ion scatters deterministically from target atoms through angles determined from classical two-body collision theory, slowing down by an additional drag force from electronic interactions. The rate at which an ion loses energy depends on both the nuclear and electronic stopping power of the target, giving

$$\frac{dE}{dx} = -N[S_n(E) + S_e(E)] \tag{8.23}$$

where N is the target atom density (5×10^{22} cm^{-3} for silicon) and $S_n(E)$ and $S_e(E)$ are the nuclear and electronic stopping powers in eV cm^2. Both are in general functions of energy.

If $S_n(E)$ and $S_e(E)$ are known, then the range of the ion can be calculated using

$$R = \int_0^R dx = \frac{1}{N}\int_0^{E_0} \frac{dE}{S_n(E) + S_e(E)} \tag{8.24}$$

8.5.1 Nuclear Stopping

When an ion with mass m_1, atomic number Z_1 and kinetic energy E_0 collides with a stationary target atom with mass m_2 and atomic number Z_2, it loses energy by interacting with the electric field of the nucleus of the target. Not all collisions are head-on collisions, so the nuclear energy loss depends on the distance of closest approach, often called the impact parameter. As the ion interacts with the electric field of the nucleus, it exchanges its kinetic energy for potential energy, reaching a maximum transfer at the distance of closest approach. This potential energy is partitioned between the ion and target atom in accordance with their masses, and the ion continues on a deflected path while the lattice atom recoils. The velocities and trajectories can be found from the conservation of momentum and energy in a classical treatment of two colliding particles. Thus, the nuclear energy loss is elastic in that the energy lost by the ion is transferred to the lattice atom, providing the atomic-scale basis for the damage created by the incoming ions. These nuclear interactions give rise to scattering and deflected trajectories, with the deflection angle depending on the impact parameter. The energy transferred

by the incoming ion and its subsequent scattering angle depends on the distance of closest approach to the target atom. If the ion and target atoms were bare nuclei, the scattering potential would be given by the simple Coulomb potential

$$V(r) = \frac{q^2 Z_1 Z_2}{4\pi\varepsilon\, r} \tag{8.25}$$

Because of the electrons which surround the target nucleus, the full effect of the positive core potential is screened from the incoming ion. Modifying the potential by an exponential screening function gives the Thomas-Fermi model of the atom

$$V(r) = \frac{q^2 Z_1 Z_2}{4\pi\varepsilon r} \exp\left(-\frac{r}{a}\right) \tag{8.26}$$

where a is some screening distance. To predict the scattering angle, the interaction of the ion with the screened Coulomb potential is integrated over the path length where the nuclear force is important, as shown in Figure 8–16. Since the scattering is deterministic in the binary collision approximation, tables of precalculated scattering angles for each impact parameter and energy speed up the calculations in practice, trading memory for speed in the simulations.

For a head-on collision, the maximum energy transferred in the nuclear encounter is given from classical two-body collision theory

$$E_{\text{Trans}} = \frac{4m_1 m_2}{(m_1 + m_2)^2} E \tag{8.27}$$

where E is the impact energy. This allows an estimate of the maximum recoil energy given to an atom in the substrate.

The nuclear stopping power $S_n(E)$ depends on the ion energy. The nuclear energy loss is small at very high energies, because fast particles have less interaction time with

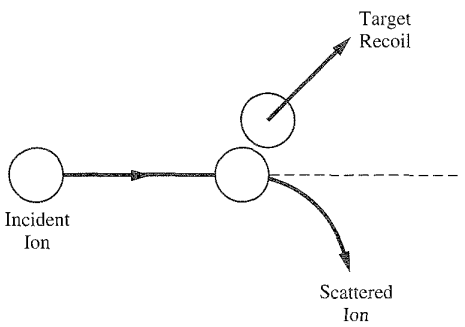


Figure 8–16 Schematic of a nuclear collision or scattering encounter between an ion and a stationary target atom.

the scattering nucleus (reduced cross section). Thus, the nuclear energy loss tends to dominate toward the end of the range when the ion has lost much of its energy and where the nuclear collisions produce most of the damage. While more sophisticated models are used in modern computer simulation tools, $S_n(E)$ can sometimes be usefully approximated as a constant value at energies below where electronic stopping is important by [8.7]

$$S_n(E) = 2.8 \times 10^{-15} \frac{Z_1 Z_2}{(Z_1^{2/3} + Z_2^{2/3})^{1/2}} \frac{m_1}{m_1 + m_2} \text{ eV cm}^2 \tag{8.28}$$

where m_1 and Z_1 refer to the ion and m_2 and Z_2 are the substrate atom mass and atomic number. Note that this simple expression does not include any energy dependence in $S_n(E)$. We will look at more accurate ways of modeling $S_n(E)$ later (Figure 8–20).

8.5.2 Nonlocal Electronic Stopping

The other important components of energy loss are due to electronic effects. The major component of electronic stopping is due to the drag that a moving ion experiences in a dielectric medium. Consider first a stationary ion in a dielectric medium. A polarization of the dielectric medium occurs around the ion to minimize the overall electric field. As the ion is set in motion, the polarization field must adjust by realignment of the electrons in the valance bonds, which at sufficient ion velocities causes the polarization field to lag behind the charged ion, leading to a drag force on the moving ion as illustrated in Figure 8–17.

An analogy is sometimes made with particle transport in a viscous medium, where the viscosity is caused by the sea of electrons from the target lattice. This drag force is directly proportional to the ion velocity at all energies of interest for ion implantation. It depends strongly on the effective ionization state of the ion in the medium (the Brandt-Kitagawa effective charge [8.6]). Because this electronic stopping mechanism is an overall characteristic of the target medium, it is referred to as nonlocal electronic stopping. It acts as a dissipative energy loss which does not alter the direction of the ion's trajectory.

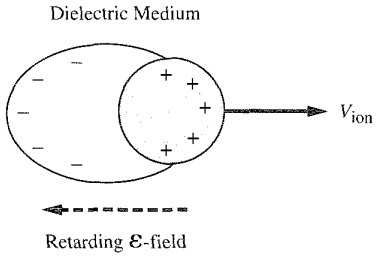


Figure 8–17 Schematic of drag force exerted on a moving ion in a dielectric medium. (After [8.2].)

Much of the progress in modeling implanted profiles has come from better models for the electronic stopping. One effect we have not considered is that the shell structure of the ion varies with the atomic mass of the ion and there are thus oscillations in the electronic stopping powers as the ion mass increases. These subtle but important effects are not shown in Figure 8–19 (discussed later), where an average electronic stopping power versus energy is plotted.

Recent work [8.8] has used a single adjustable parameter which provides an effective ionization radius to describe electronic stopping for different species in silicon. It remains to be seen if this is a sufficiently good description of the electronic stopping, or if an additional local component for electronic stopping is also needed.

8.5.3 Local Electronic Stopping

An argument can be made that the local electronic environment around the lattice atoms also contributes to electronic stopping, through collisions with electrons and consequent momentum transfer. If an ion passes close enough to a lattice atom so that their electron wavefunctions overlap, there can be charge exchange and momentum exchange, as shown in Figure 8–18. This reduces the ion energy and gives rise to a force on the ion which slows it down. The ion-target electronic collisions give rise to a momentum transfer $m_e v_{\text{ion}}$, where m_e is the electron mass and v_{ion} is the velocity of the ion. Because of the momentum transfer, it can subtly alter the ion trajectory, though this effect is minor compared to the scattering from the nuclear core. To compute the total momentum transfer, the momentum transfer from a single electron is multiplied by the electron current transferred between the ion and target during the course of an electronic interaction. The momentum transfer is proportional to the local electron concentration in the vicinity of the target atom multiplied by the orbital velocity of the electrons. Both the local electron concentration and the orbital velocity of the electrons depend strongly on the distance from the atoms. This gives rise to a long-range (compared with nuclear stopping) local electronic stopping which depends directly on the ion velocity and decreases as the fifth power of the impact parameter.

Since both forms of electronic stopping depend directly on the ion velocity, we can write

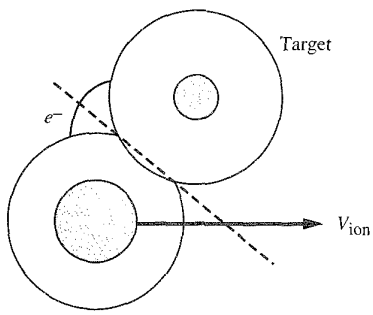


Figure 8–18 Schematic of charge and momentum exchange due to local electronic interactions. (After [8.2].)

$$S_e(E) = cv_{\text{ion}} = kE^{1/2} \tag{8.29}$$

where c and k are parameters that depend on the ion, the substrate, and the particular electronic stopping process being modeled. In the simplest case, in which the substrate is assumed to be amorphous, k is roughly independent of the ion being implanted and is given by [8.7]

$$k \cong 0.2 \times 10^{-15}\ \text{eV}^{1/2}\ \text{cm}^2 \tag{8.30}$$

In summary, electronic stopping is inelastic and is responsible for dissipating the energy of the incoming ions. Both the overall dielectric response of the target medium and the local electronic structure of the target atoms contribute to the electronic stopping.

8.5.4 Total Stopping Powers

Both the nuclear and electronic stopping powers depend on energy and contribute to the total stopping, as seen in Figure 8–19. For some range of ion energies, the total stopping power is approximately constant, which explains why the ion ranges in Figure 8–3 are approximately linear with energy.

Clearly, the local environment (such as whether the ion is in a channel or amorphous region) will have a dramatic effect on the nuclear encounters that will be faced by the ion. To adequately model implanted profiles, especially in channeling directions, accurate models for the electronic stopping must also be available. Many modern processes have implants at zero tilt to avoid shadowing effects (Figure 8–6). In addition, some channeling component always exists because of the inherent structure of the lattice. For these reasons, accurate modeling of both the nuclear and electronic stopping in a Monte Carlo simulator is the only feasible way to obtain accurate implanted profiles.

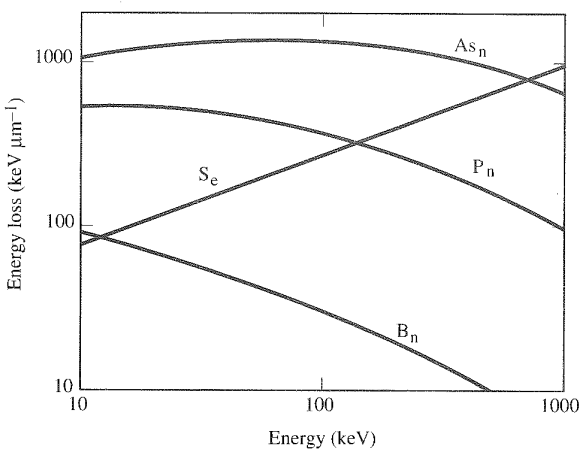


Figure 8–19 The total stopping power is composed of the sum of the nuclear and electronic stopping. An average electronic stopping power is plotted, though the details depend on the shell structure of the incident ion. The nuclear stopping dominates at low energies while the electronic stopping dominates at higher energies. (After [8.2].)

8.5.5 Damage Production

The major disadvantage of ion implantation is that the nuclear stopping creates a lot of damage in the lattice that must be repaired by annealing at high temperatures. The energy required to displace a silicon atom just far enough from its lattice site to create a stable separated interstitial and vacancy (a Frenkel pair) is the displacement energy E_d and is approximately 15 eV in silicon. Because the implant energy is often in the range of kilovolts, a large number of displacements will be caused.

Example How many displaced lattice atoms are created by a single incident 30-keV arsenic atom?

Answer For a heavy ion like arsenic, this energy is mainly dissipated due to nuclear collisions [8.9]. From Figure 8–3, a 30-keV arsenic implant has a range of 25 nm, which means it transverses 100 planes of atoms (0.25 nm average interplanar spacing), so that on average it loses 300 eV in a nuclear collision at each plane, as illustrated in Figure 8–20. This is much more than the displacement energy, so the lattice atom that is displaced by the incident ion (the primary knock-on) can continue and itself lose energy and displace other secondary lattice atoms. This is essentially equivalent to a silicon implant with 300 eV of energy at each point where a nuclear collision occurred. When the energy of the incident ion or the secondary knocked-on atoms reach E_d, they can be considered stopped, because if they do manage to transfer all their energy to a lattice atom, they can cause a single displacement but remain at rest in the lattice position themselves. Thus, an energetic particle can only increase the number of moving particles if it has energy greater than $2E_d$. This allows an estimate of the number of displaced atoms created by an energetic particle to be made (the Kinchin-Pease formula):

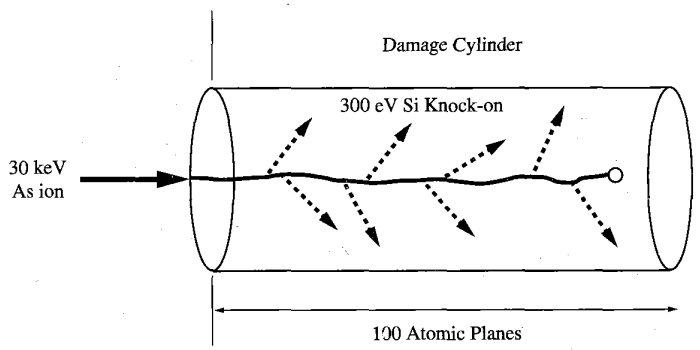


Figure 8–20 Schematic of the damage caused by a 30-keV arsenic implant into silicon, with a range of 25 nm. The ion crosses approximately 100 atomic planes. (After [8.9].)

$$n = \frac{E_n}{2E_d} \tag{8.31}$$

where E_n is the energy lost in nuclear collisions. For a 30-keV arsenic ion,

$$n = \frac{30{,}000}{2 \times 15} = 1000 \text{ displaced atoms}$$

Equivalently, n = 100 planes $\times \frac{300}{2 \times 15}$ atoms/plane = 1000 atoms

Thus, each incident arsenic ion creates a trail or cascade of 1000 displaced lattice atoms.

What is the time scale on which this damage production occurs? We can estimate that the time it takes the ion to come to rest is given by the range divided by its velocity:

$$t = \frac{R_p}{\sqrt{2E/m}} \approx 10^{-13} \text{ sec} \tag{8.32}$$

while the thermal vibrations decay on the order of the Debye frequency (10^{-12} sec). Some recombination occurs within the cascade due to elementary diffusion hops on the time scale of 10^{-9} sec after which the primary damage generated by the incident ion can be considered stable. This damage is primarily small defect clusters and dopant-defect complexes and some isolated interstitials and vacancies. Such a detailed picture of the damage evolution has been obtained from molecular dynamics calculations of ions impacting a crystalline silicon lattice. Since the damage evolution is primarily determined by the silicon atoms that are knocked on, the well-characterized silicon interatomic potentials should provide a physically accurate description of the damage evolution. An example of these calculations is shown in Figure 8–21.

The question now is how this primary damage accumulates when other ions are implanted. Some of the defects generated can recombine with defects from other cascades, so the damage accumulation depends on the existing local defect density. Thus the amount of damage that accumulates [8.11] depends on the number n generated in an isolated cascade, on the fraction f_{rec} that recombine within a cascade and from overlapping cascades, along with a factor for the existing local defect density. The increment in the primary damage in a volume element for an additional implanted ion is given by

$$\Delta n(x) = n f_{\text{rec}} \left(1 - \frac{N}{N_\alpha}\right) \tag{8.33}$$

where N is the local defect density and N_α is a threshold defect density where the crystal is considered amorphous. Clearly, if the crystal is amorphous $N = N_\alpha$ and $\Delta n = 0$,

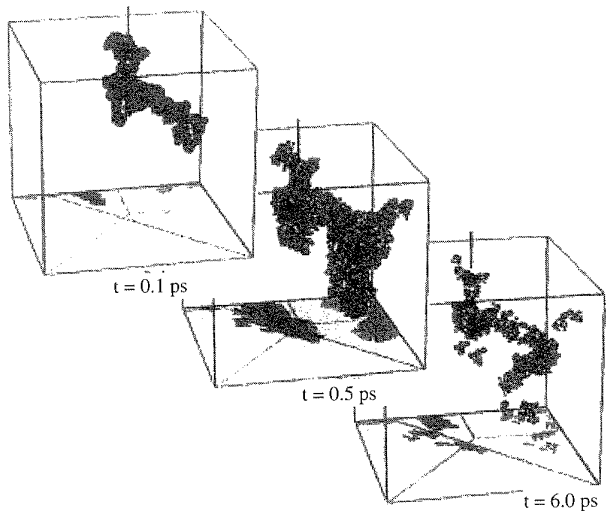


Figure 8–21 Molecular dynamics simulations of a 5-keV boron implant into silicon, showing the damage evolution and stabilization. The simulation temperature was 40K and the atoms plotted have a potential energy greater than 0.2 eV. Reprinted with permission of *Nucl. Instrum. Methods Phys.* [8.10].

so that there is no incremental disorder introduced by additional cascades. In UT-Marlowe [8.2], the values of f_{rec} are chosen to be 0.1, 0.4, 0.6, 0.6 for boron, phosphorus, arsenic, and BF_2 implants. Physically, the dense damage cascades from heavy species like arsenic and BF_2 allow for more efficient recombination than the more dispersed damage distribution from a light ion like boron. The threshold damage density for amorphization, N_α, is often taken to be 10% of the silicon lattice density. It is clear that the primary damage builds up until eventually an amorphous state is reached, after which no channeling occurs and the damage accumulation saturates.

Since the number of displacements n produced by an ion depends on the amount of energy deposited into nuclear collisions, the damage production is a maximum where the nuclear energy losses are highest. For heavy ions like arsenic, whose stopping may be dominated by nuclear collisions, this means that the damage profile is relatively flat over the whole range up to R_P. For lighter ions like boron, which have an appreciable component of electronic stopping at higher energies, the damage accumulation is concentrated near R_P where the ion energy is lower. For light ions then, the amorphous region will form first at the peak of the damage density profile near R_P and expand on both sides of this depth as the implant dose is increased. Experimental measurements of the amorphization layer thickness as a function of implant dose support this picture and indicate a U-shaped distribution for the amorphous layer thickness depending on implant dose, as shown in Figure 8–22. A buried amorphous layer first forms, and it may

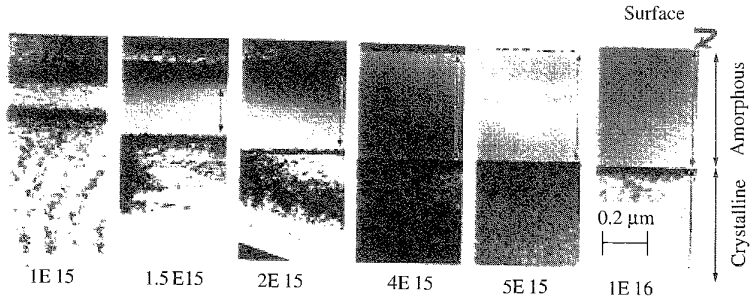


Figure 8–22 Cross-sectional TEM images of the formation of an amorphous layer for increasing Si implant doses. The layer begins to form where the nuclear damage is highest, spreading wider with higher doses. Reprinted with permission of J. Appl. Phys. [8.12].

take a considerably higher implant dose before a continuous amorphous layer extending down from the surface forms [8.12]. The experimental results also indicate that it is easier to form an amorphous layer at low temperatures (liquid nitrogen) rather than at room temperature or higher. This observation is easily explained by the larger fraction of ions that recombine within a cascade at higher implant temperatures.

These ideas suggest the possibility of preamorphizing the silicon substrate to eliminate ion channeling, perhaps by preimplanting the substrate with silicon or germanium ions. If an amorphous layer is created, then a subsequent implant of a dopant will not channel and a shallow implant profile can be formed.

8.5.6 Damage Annealing

The goal of damage annealing is to remove the primary damage created by the implant and restore the silicon lattice to its perfect crystalline state, leaving the dopants on active substitutional sites. This damage removal can be divided into two distinct regimes: one below and the other above the amorphous threshold.

The primary damage begins to anneal at relatively low temperatures of about 400°C. First, the vacancy complexes (di-vacancies and vacancy-type clusters) begin to break up and the released vacancies annihilate with interstitials. The final result of the initial stages of annealing is to remove most of the Frenkel pairs, leaving only interstitial-type defects whose origin stems from the extra atom introduced into the lattice. This gives rise to the "+1" model for residual damage due to implants, that is, to a good approximation, all of the original damage recombines, leaving only one interstitial created when the implanted ion finally occupies a lattice site. Figure 8–23 illustrates some of these effects. Note the very short time scale on which the Frenkel pairs recombine in this simulation through bulk I-V recombination [8.13]. Some vacancies recombine independently at the surface. After 10^{-2} seconds, only interstitials remain in small clusters, and these slowly dissolve by recombining interstitials at the surface. Though this

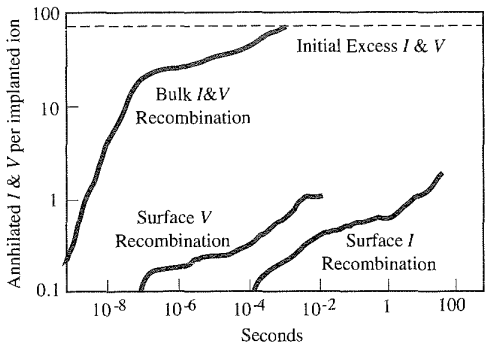


Figure 8–23 Monte Carlo simulation of the recombination of I and V damage generated by an implant. After a short time, only excess interstitials remain and these form clusters in a fraction of a second at 800°C. (After [8.13].)

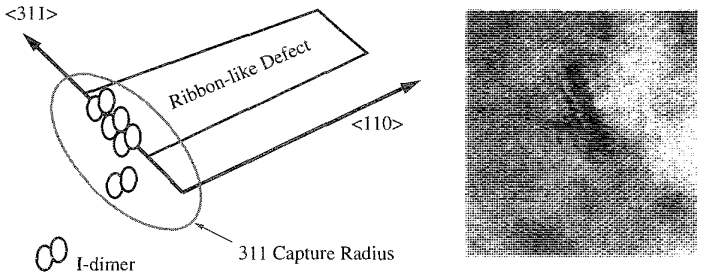


Figure 8–24 Schematic (left) and cross section TEM (right) of {311} clusters that form by capturing a row of interstitial dimers that lie on the {311} plane and grow by extending in the 110 directions. Photo courtesy of H. Chao, Stanford University.

picture is simple, it captures much of the essential physics that occurs during implant damage annealing.

The silicon interstitials generated by this "+1" amount of damage very quickly condense into characteristic rod-shaped defect clusters upon further annealing at temperatures above 400°C. These rod-shaped defects are composed of ribbons of silicon atoms that lie on {311} planes as illustrated in Figure 8–24. An actual TEM image of such a defect is also shown in Figure 8–24.

The {311} defects are typically composed of rows of a few silicon dimers that precipitate on {311} planes and extend in the <110> directions to form a planar defect. After a 5-second anneal at 900°C, there may be a very high concentration ($\approx 10^{11}$ cm^{-2}) of these {311} defects, which are only about 10 nm long. The {311} defects start to dissolve upon further annealing by the evaporation of silicon interstitials from the ends of the defects.

If the amount of primary damage is below a critical value, these {311} defects dissolve completely and the crystal is restored to its perfect state. Above this critical damage level, some of the {311} defects grow at the expense of others which shrink. The larger

{311} defects can turn into stable dislocation loops which are much more difficult to remove. These stable dislocation loops are called secondary defects and remain when the primary damage is completely annealed. The dislocation loops can be approximated by an extra circular atomic layer of silicon atoms precipitated on a {111} plane of silicon. They can be shown from TEM images to be extrinsic loops, that is, they are composed of an extra plane of silicon interstitials.

Stable dislocation loops are a much more dominant feature of higher-dose implants, which drive the silicon amorphous. The largest concentration of these loops is seen at the interface between amorphous and crystalline silicon after an anneal which regrows the amorphous region by solid phase epitaxy. These defects at the original amorphous/crystalline (a/c) interface are referred to as End-Of-Range (EOR) defects as illustrated in Figure 8–25. They occur because there is a large amount of damage which is just below the threshold of amorphization beyond the a/c interface. By definition, just beyond the a/c interface is the maximum possible amount of damage that can exist in the crystal at the implant temperature without its being amorphous. This damage is sufficiently high to be able to nucleate a mixture of {311} defects and a band of dislocation loops in a very narrow region just beneath the a/c interface on the crystalline side of the interface.

The loops have a mean radius of approximately 10 nm and ripen to a mean radius of approximately 20 nm during subsequent annealing at a typical RTA anneal temperature of 1000°C, as shown in Figure 8–26 [8.14]. This ripening of the loops happens because a few big loops are energetically more favorable than many small loops since this minimizes the total circumference and the total energy of the system. The ripening process appears to conserve the total number of interstitials trapped in the loops, which

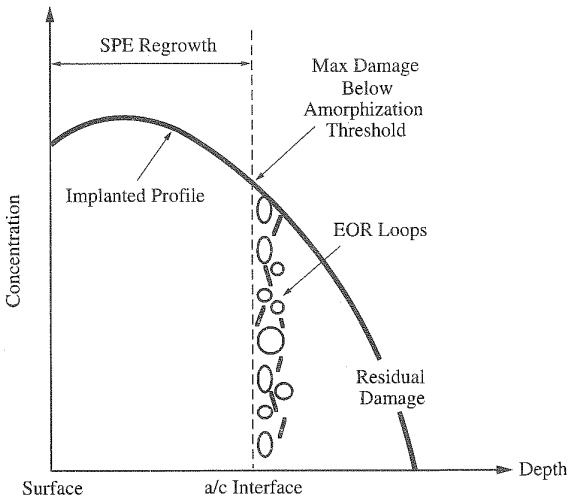


Figure 8–25 Schematic of the stable End-Of-Range (EOR) dislocation loops that form at the amorphous/crystalline (a/c) interface after solid-phase epitaxial regrowth.

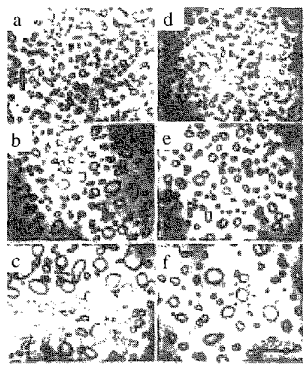


Figure 8–26 Plan view TEM images of the EOR dislocation loops for annealing times of 5, 60, and 960 min at 850°C (a–c) and for 1, 60, and 400 sec at 1000°C (d–f). The loops show some ripening (large circumference loops grow at the expense of smaller loops to minimize the total energy) but remain very stable for typical anneal temperatures. Reprinted with permission of *J. Appl. Phys.* [8.14].

indicates just how stable this secondary damage is at 1000°C. The dislocation loops will disappear at high enough temperatures, and times of 60 sec at 1100°C have been reported to remove loops in some instances.

If the dislocation loops are in the depletion region of a PN junction during device operation, they can give rise to large leakage currents, in particular if they become decorated with metal impurities. Thus, they are often detrimental to device performance. Annealing cycles are chosen to cause just enough dopant diffusion so that the loops are contained in the highly doped regions and are shielded from any depletion regions during device operation. This is a critical optimization problem, since there are often conflicting requirements to minimize the amount of dopant diffusion and maintain shallow junction depths.

The implant process can also be optimized to minimize the nucleation of the loops. In particular, if the implant is performed at liquid nitrogen temperatures rather than room temperature, a lower implant dose can be used to form the same amorphous layer depth. Because of this, the amount or dose of damage in the tail of the damage distribution beyond the a/c interface is much less for the liquid nitrogen temperature implant and may be small enough to preclude stable dislocation loop formation. For example, 75-keV Germanium with a dose of 5×10^{14} cm^{-2} gives rise to an amorphous layer 100 nm deep for a room temperature implant, while only 4×10^{14} cm^{-2} is needed to form the same amorphous layer thickness at liquid nitrogen temperatures and the a/c interface is also sharper. A band of EOR loops at a depth of 100 nm is seen for the room temperature implant, where the loop diameter is 25 nm and their concentration is about 10^{10} cm^{-2} after an anneal at 900°C for 15 min. No loops are seen at all for the LN_2 sample after the same anneal [8.15].

8.5.7 Solid-Phase Epitaxy

If the substrate is amorphous, then it can regrow by layer-by-layer epitaxial realignment beginning from the amorphous/crystalline interface. This process is similar to the crystallization process that occurs from the melt onto a single-crystal seed or from a gas

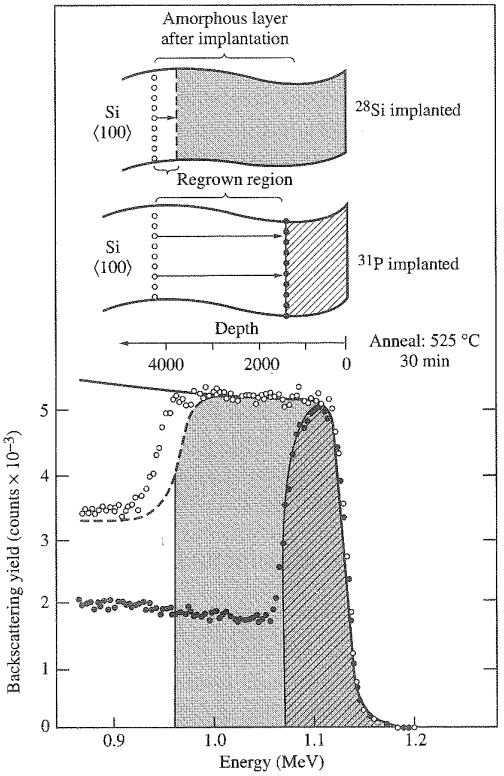


Figure 8–27 Channeling RBS images of the regrowth of an amorphous layer during annealing. The layer regrows from the substrate and undoped silicon regrows more slowly than doped silicon. Reprinted with permission of *J. Appl. Phys.* [8.16].

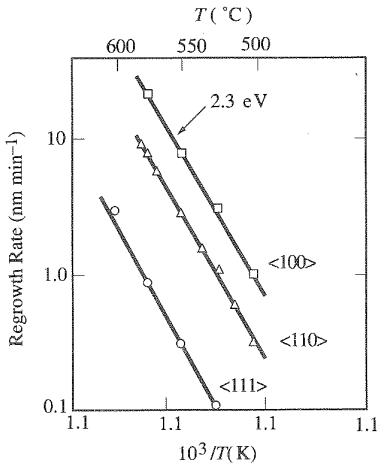


Figure 8–28 Plot of the regrowth velocity during solid phase epitaxy as a function of temperature. (After [8.17].)

phase onto a crystalline substrate, except that it is now occurring from a solid phase rather than a liquid or gas phase. Figure 8–27 shows a time lapse series of channeling Rutherford Backscattering Spectrometry (RBS) images of this regrowth process. It occurs quite quickly, with regrowth rates in undoped silicon at 600°C of 50 nm min^{-1} for <100> orientations, 20 nm min^{-1} for <110> orientations, and 2 nm min^{-1} for <111> orientations, as illustrated in Figure 8–28. The activation energy for the regrowth process on any orientation is 2.3 eV, which is indicative of a silicon-silicon bond breaking and formation process. The regrowth rate is given by

$$v = A \exp\left(-\frac{2.3}{kT}\right) \tag{8.34}$$

where v is the regrowth rate and A is an experimentally determined parameter. The regrowth rate can be enhanced by a factor of 10 for doping levels characteristic of the

source and drain in MOS devices. The regrowth at these low temperatures quickly eliminates all the primary damage in the amorphous region, so that no anomalous dopant diffusion occurs there. Most of the dopant atoms in the amorphous regions are incorporated onto substitutional lattice sites during regrowth, so that high levels of activation are possible even at these low temperatures. Indeed, were it not for the residual band of EOR defects just below the a/c interface, solid-phase epitaxy would provide an ideal method for simultaneously removing all the implant damage and obtaining high levels of dopant activation without any noticeable dopant diffusion.

If the implant energy is high enough, a buried amorphous layer can be created. Solid-phase regrowth occurs from both the top and bottom interfaces and leaves a band of dislocations in the center when the two moving interfaces collide.

8.5.8 Dopant Activation

In order to contribute to the electrical activity, the implanted dopant atoms must occupy substitutional sites in the lattice. In addition, the broken bonds in the lattice must be repaired so that the carrier mobility is returned to a reasonable value. At a sufficiently low level of primary damage, all the damage anneals out and high activation levels are obtained. As indicated above, for very high levels of damage where amorphization takes place, solid-phase epitaxial regrowth provides a nearly ideal method to remove damage and obtain high levels of dopant activation. It is much more difficult to anneal a sample that has only been partially damaged at levels below the amorphization threshold. This difficult regime is where secondary damage forms—then the annealing behavior is more complex. For example, Figure 8–29 shows both regimes clearly. Above the amorphization threshold ($\approx 1 \times 10^{15}$ cm^{-2} dose), an anneal temperature of 600°C is sufficient to get high activation. For very low doses ($\approx 10^{12}$ cm^{-2}), the damage also anneals out at 600°C, giving high levels of activation. However, for increasing doses where secondary defects form, up to the amorphous threshold, the damage becomes increasingly difficult to anneal. Typically, temperatures of 950°C to 1050°C are required to obtain useful levels of activation when secondary damage has formed.

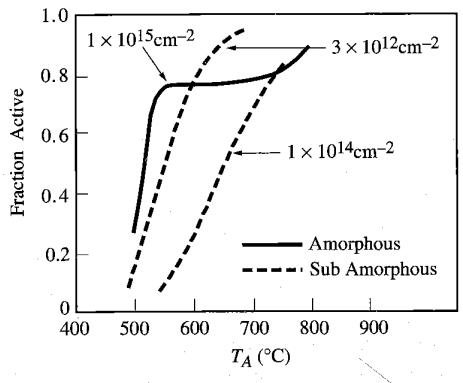


Figure 8–29 Plot of the fraction of an implant activated as a function of anneal temperature for various doses. Note the distinct difference between doses below and above the amorphization threshold (After [8.18].)

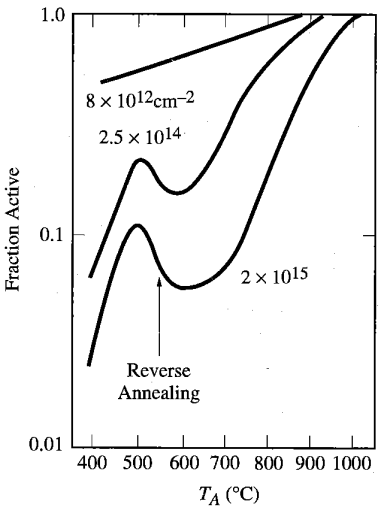


Figure 8–30 Plot of the fractional activation of a boron implant as a function of anneal temperature, showing a region where there is reverse activation. This reverse annealing behavior is thought to be due to competition between boron atoms and native silicon point defects for substitutional lattice sites. Isochronal anneals (30 min at each temperature) were used in these experiments. (After [8.19].)

The behavior of the activation process as a function of temperature provides some indication of the complex interactions that occur between the dopants and the defects. Figure 8–30 shows anneals for different implanted boron doses below the amorphization threshold as a function of temperature. For very low doses, much of the boron is active even after very low temperature anneals and it quickly gains full activity as the anneal temperature is increased. For higher doses, the initial activation level is very low. The damage created by the implant creates deep-level traps which increase the resistivity in either n- or p-type silicon. Some of these traps begin to anneal out at temperatures below 400°C, causing a reduction in the resistivity and an increase in the carrier concentration. However, between approximately 450°C and 550°C, the interstitial damage that remains is responsible for a "reverse annealing" effect, where the carrier concentration abruptly drops. This is caused by silicon interstitials competing with boron atoms for substitutional sites or by interstitials pairing with boron to form inactive complexes. Above 550°C, the activation level slowly rises, approaching full activation at the highest anneal temperatures. Similar annealing behavior is seen for phosphorus below the amorphization threshold.

The problem with using these higher anneal temperatures to activate the dopant is that diffusion also occurs, making it difficult to form shallow junctions. The problem of minimizing the amount of dopant diffusion after implantation is exacerbated by the presence of the excess interstitials which give rise to a transient enhancement in the dopant diffusion (TED). This transient-enhanced diffusion is anomalous because there are conditions where the dopant can diffuse more at lower temperatures than at higher temperatures. It might seem that only the $\sqrt{Dt}$ product should determine the extent of profile motion, that is, a longer time at low temperature could be chosen to be equivalent to a short time at high temperature. A longer time at lower temperature would allow a controllable anneal to be performed in a furnace environment. However, when

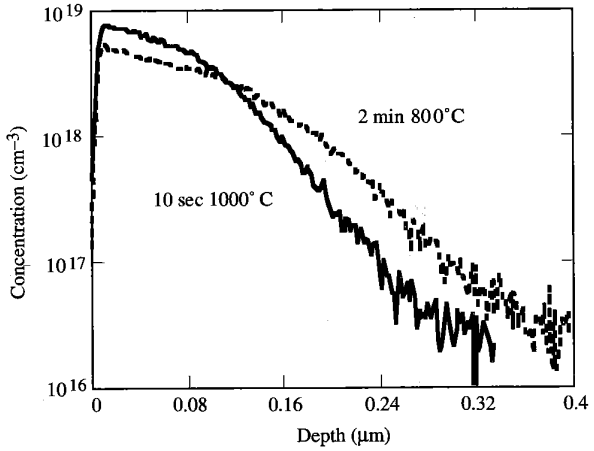


Figure 8–31 Temperature dependence of transient-enhanced diffusion, showing more diffusion for lower temperatures. (After [8.20].)

implant damage is present, the result is very surprising and is shown in Figure 8–31 [8.20]. Here, the profile annealed at the lower temperature diffused further than the one annealed at high temperature.

The reason is that at lower temperatures the damage can stay around longer and enhance the dopant diffusion, while at higher temperatures the damage annihilates faster. Historically, long anneals at high temperatures masked this effect, because eventually normal thermal diffusion will occur and is exponentially faster at higher temperatures. However, for the small thermal budget anneals used in present day devices, this anomalous TED has become the dominant effect that determines how far an implanted dopant moves during an anneal.

The magnitude of this anomalous TED can be quantified by examining how large the enhancement in dopant diffusivity is at each temperature. The results are astonishing, with enhancements ranging from 20,000 at 700°C to 400 at 1000°C. In the following section, we will examine the phenomena of TED and show how its impact on devices can be mitigated.

8.5.9 Transient-Enhanced Diffusion

There now exists a much deeper understanding of how Transient-Enhanced Diffusion (TED) due to implant damage affects dopant motion and consequently device performance. This is a direct result of understanding at an atomic level the physical processes that control TED. Transient-enhanced diffusion occurs when an attempt is made to anneal implant damage and restore the lattice to its crystalline perfection. It consists of a burst of diffusion many thousands of time faster than what is normally observed for similar anneals when no implant damage is present.

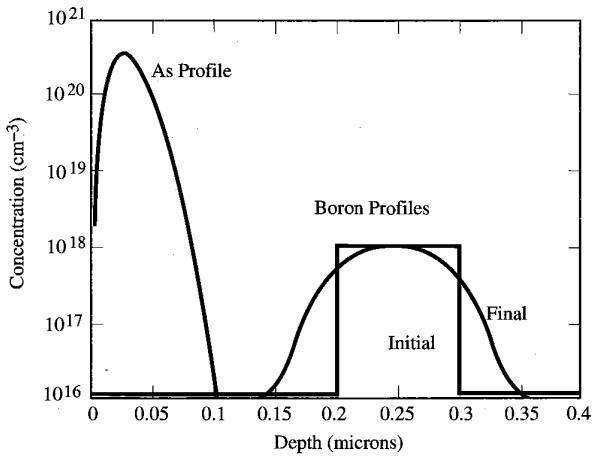


Figure 8–32 Simulation of an arsenic layer implanted above a buried boron layer that gives rise to damage beyond the amorphization/crystalline interface that is able to cause substantial diffusion in the nearby boron layer during a 1-sec, 900°C anneal. (After [8.21].)

Where is TED important and how are its effects measured? Consider for example the arsenic implant in Figure 8–32, which is performed above a buried, lightly doped boron layer. The boron layer might be grown using epitaxial silicon growth, followed by a lightly doped intrinsic silicon layer. The arsenic implant amorphizes the surface region of the substrate, and solid-phase epitaxial regrowth removes much of the implant damage. But there is still a sufficient amount of damage below the a/c interface to cause an enormous enhancement in the diffusion of the nearby boron buried layer, which would not normally show any appreciable motion for the short anneal at 900°C. The arsenic profile itself also shows some slight but significant diffusion, but it is less for arsenic than for boron because of the lower component of arsenic diffusion that takes place by an interstitial mechanism and also because of the lower intrinsic diffusivity of arsenic.

Figure 8–32 illustrates the effect that TED can have in determining how the damage introduced by one dopant affects its own diffusion and also that of a nearby dopants. The structure in Figure 8–32 represents a unique experimental structure that also provides a means to quantify the amount of enhancement seen by a "marker layer" when damage is introduced into the crystal. It is a particularly useful structure because the amount of profile motion can easily be measured using SIMS. Its most important feature is that it decouples the damage itself from the effects of the damage on a nearby region. This test structure is a one-dimensional analog of the source/drain to channel coupling in a MOS device. Thus, experiments on this test structure allow models to be developed which are directly relevant to how TED affects MOS device characteristics. This kind of test structure has allowed many crucial experiments about the nature of TED to be performed.

8.5.10 Atomic-Level Understanding of TED

Atomic-level understanding of TED has come from key experimental [8.22] and theoretical studies. An important experimental finding was that the point defects responsible for TED are interstitial-type extended defects in the form of {311} clusters, whose evaporation rate matches the kinetics of TED [8.23]. Figure 8–33 shows the experimentally measured dissolution rate of the {311} defects. The same time dependence is seen experimentally in TED. These {311} defects appear in transmission electron microscope images as ribbonlike structures that lie on {311} planes and grow and extend in the <110> direction. They form during the very early stages of the anneal, being visible after 1 second at 600°C. The key theoretical results come from molecular dynamics simulations of the collision cascade process in silicon [8.10] and from Monte Carlo simulations of the point-defect evolution and clustering [8.13]. The molecular dynamics simulations show that although a large cascade of damage is generated immediately after an ion strikes the silicon lattice, there is much local recombination and only a small amount of damage remains after a nanosecond. (Recall Figure 8–21.) The Monte Carlo simulations of the subsequent anneal indicate that the small vacancy clusters first decay and the free vacancies start to annihilate interstitial-type defects. After a hundredth of a second at 815°C, only interstitial-type defects remain that are caused by the imbalance introduced in the lattice by the extra implanted ion, as shown in Figure 8–23. Recent experimental work using Deep-Level Transient Spectroscopy (DLTS) confirms this picture [8.24], where interstitial-type defects are found after annealing ion implanted samples but are not found after annealing comparable damage produced by electron beam impact where extra atoms are not introduced.

The microscopic explanation of the energy dependence of TED is even more interesting for the insights it provides into the nature of the residual implant damage. In a clever experiment, Giles and colleagues [8.25] implanted a dopant at the same depth below the surface using both a normal angle of incidence for the ion beam and using a higher energy at a tilted angle of incidence. One expects that the higher energy creates more damage, but the amount of TED in both profiles is identical. This indicates that though more primary damage is created, the bulk recombination process is efficient

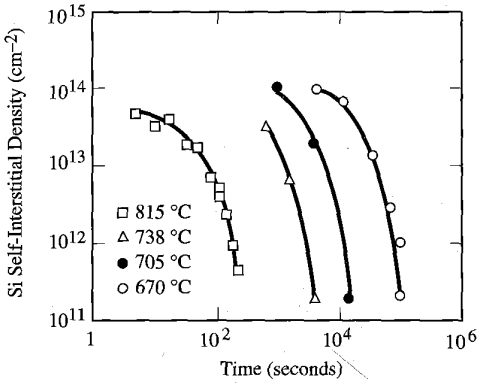


Figure 8–33 Plot of the number of interstitials contained in {311} defect clusters as a function of anneal time at various temperatures. The {311} defects dissolve and provide the interstitial supersaturation that controls TED. (After [8.23].)

enough that essentially all the interstitials and vacancies produced by the implant annihilate in the early stages of the anneal, leaving only a dose of interstitials due to the extra ions introduced into the lattice. This provides a remarkable confirmation of the "+1" approach to implant damage modeling [8.26], where the assumption is made that the net effect of an implant is to introduce one extra interstitial when a dopant atom is introduced into the lattice. This "+1" dose of interstitials then proceeds to cluster and control the kinetics of the TED.

A simulation of how this "+1" interstitial profile from a boron implant evolves in time is shown in Figure 8–34. Rapid clustering of interstitials into {311} defects reduces the enormous supersaturation of interstitials immediately after implant in a time scale of 10^{-2} seconds, and subsequently the supersaturation level is maintained approximately constant while the clusters are dissolving, until the clusters eventually disappear. Most of the TED occurs during this period when the supersaturation of interstitials is at a high, constant level allowing the maximum doping motion to occur.

While the +1 model is more physically correct than considering the thousandfold damage excess introduced by an incoming ion, it should perhaps be better termed a "+n" model where n is of order unity. Clearly, the details of the cascade evolution and in particular the effects of atoms lost to sputtering near the surface along with any of the net vacancies captured in the screen oxide all modify the effective dose of interstitials introduced by the implant. In most simulation programs, this number is taken as a fitting parameter, though detailed Monte Carlo calculations of the implant cascades and

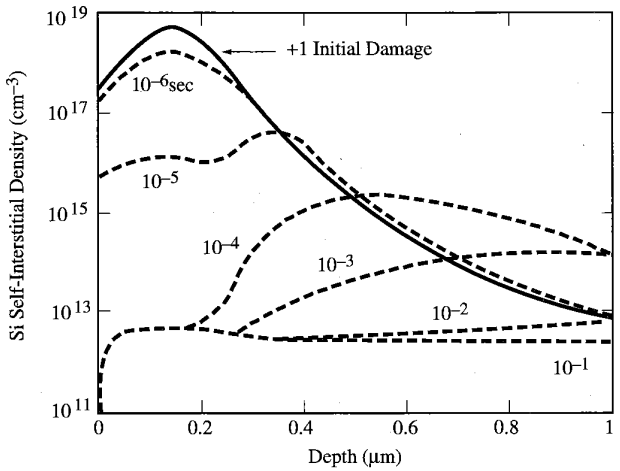


Figure 8–34 TSUPREM IV simulation [8.1] of the time evolution of the "+1" damage from a 40-keV, 10^{14} cm^{-2} boron implant, for anneals from 10^{-6} sec to 10^{-1} sec at 750°C. The equilibrium interstitial concentration is approximately 10^{8} cm^{-3}, so the flat concentration profile at 10^{-1} sec represents an interstitial supersaturation of more than 10,000 fold, and TED occurs until surface recombination reduces this to equilibrium levels.

their subsequent evolution allow trends in the "$+n$" value to be obtained as a function of implant species, dose, and energy.

Given these atomic-level insights, we can now quantitatively model the impact of TED on dopant diffusion [8.27] and hence device characteristics. The fact that the {311} defects form remarkably quickly, as shown in Figure 8–34, implies that the TED transient is controlled by the evaporation of interstitials from these aggregated defects. The process of cluster growth and evaporation can be atomistically described by

$$I + Cl_n \Leftrightarrow Cl_{n+1} \tag{8.35}$$

where the subscript n indicates a cluster with n interstitials. If we assume that the properties of the cluster remain the same with size, we can write an equation [8.27] for the time rate of change of the cluster growth as

$$-\frac{\partial C_I}{\partial t} = \frac{\partial Cl}{\partial t} = k_f C_I Cl - k_r Cl$$
$$= \text{growth} - \text{shrinkage} \tag{8.36}$$

where the forward reaction rate for {311} cluster growth is given by k_f and the reverse reaction rate for cluster shrinkage is given by k_r. We can assume that the growth of the clusters is diffusion limited so that the forward reaction rate is given by

$$k_f = 4\pi a d_I \tag{8.37}$$

where a is an encounter distance equal to the nearest neighbor distance in silicon and d_I is the interstitial diffusivity. The reverse reaction constant which describes the rate of evaporation will be given by the attempted hop frequency of the bound interstitial d_I/a^2, times a Boltzmann factor to account for the binding energy to the cluster, so that

$$k_r = \frac{d_I}{a^2}\exp\left(-\frac{E_b}{kT}\right) \tag{8.38}$$

For longer anneal times, the growth of the clusters is not important, leaving only the second term

$$\frac{\partial Cl}{\partial t} = -k_r Cl \tag{8.39}$$

which is the origin of the exponential decay kinetics of the clusters and the consequent change in the dopant diffusivity for longer time anneals. This accounts for some attempts to model the effective diffusivity during TED with terms of the form

$$D_{eff} = D_0 + D_{II}\exp\left(-\frac{t}{\tau}\right) \tag{8.40}$$

However, this formulation ignores the initial stages of the anneal, where a steady state exists between the cluster growth and evaporation. The steady-state period turns out to be the dominant source of the amount of TED observed. This balance between growth

and evaporation when $\partial Cl/\partial t = 0$ gives rise to a constant interstitial concentration during the steady state period of

$$C_I^{\max} = \frac{k_r}{k_f} = \frac{1}{4\pi a^3} \exp\left(-\frac{E_b}{kT}\right) \tag{8.41}$$

It is more useful to write this as a constant interstitial supersaturation during the steady-state period, since this will determine the enhancement in the dopant diffusivity, giving

$$\frac{C_I^{\max}}{C_I^*} = \frac{1}{4\pi a^3 C_I^0} \exp\left(-\frac{E_b - E_F}{kT}\right) \tag{8.42}$$

where we used the formula for the equilibrium interstitial concentration from Chapter 3,

$$C_I^* = C_{\text{Si}} \exp\left(-\frac{(E_F - TS_F)}{kT}\right) = C^0{}_I \exp\left(-\frac{E_F}{kT}\right) \tag{8.43}$$

Example Calculate the interstitial supersaturation versus temperature using the above equations. Assume the binding energy of an interstitial to a {311} cluster is 1.77 eV.

Answer From Table 7–5, we have the self-diffusion coefficient for silicon

$$D_{\text{Self}} = 560 \exp\left(-\frac{4.76}{\text{kT}}\right) \text{cm}^2 \text{ sec}^{-1}$$

This is the overall measured diffusion coefficient for silicon atoms, so the self-diffusion flux is given by

$$D_{\text{Self}} C_{\text{Si}} = 560 \exp\left(-\frac{4.76}{\text{kT}}\right)[\text{cm}^2 \text{ sec}^{-1}](5 \times 10^{22})[\text{cm}^{-3}]$$

$$= 2800 \times 10^{22} \exp\left(-\frac{4.76}{\text{kT}}\right)[\text{cm}^{-1} \text{ sec}^{-1}]$$

The interstitial fraction of the self-diffusion (using Table 7–6) is

$$d_I C_I^* = (0.6)2800 \times 10^{22} \exp\left(-\frac{4.76}{kT}\right) = 1680 \times 10^{22} \exp\left(-\frac{4.76}{kT}\right)[\text{cm}^{-1}\text{s}^{-1}]$$

This equation describes the magnitude and temperature dependence of the interstitial flux. Note that the activation energy is large compared to the activation energy for dopant diffusion. This will become important in the later discussion.

A reasonable value for the diffusion coefficient of interstitials is given by

$$d_I = 51 \exp\left(-\frac{1.8}{kT}\right)[\text{cm}^2\text{s}^{-1}]$$

allowing C_I^* to be calculated

$$C_I^* = 33 \times 10^{22} \exp\left(-\frac{2.96}{kT}\right) [\text{cm}^{-3}]$$

The jump distance can be taken as half the lattice constant, so that

$$a = \frac{5.41 \times 10^{-8}}{2} = 2.7 \times 10^{-8} \text{ cm}$$

The binding energy of interstitials to {311} clusters is 1.77 eV, which was obtained by fitting the decay kinetics of the 311 defects in Figure 8–33 [8.27]. It is a pure coincidence that this is so similar to the activation energy for interstitial diffusion. Using these values gives

$$\begin{aligned}\frac{C_I^{\max}}{C_I^*} &= \frac{1}{4\pi a^3 C_I^0} \exp\left(-\frac{E_b - E_F}{kT}\right) \\ &= \frac{1}{4\pi(2.7 \times 10^{-8})^3 (33 \times 10^{22})} \exp\left(-\frac{1.77 - 2.96}{kT}\right) \\ &= 1.22 \times 10^{-2} \exp\left(\frac{1.19}{kT}\right)\end{aligned}$$

which is plotted in Figure 8–35.

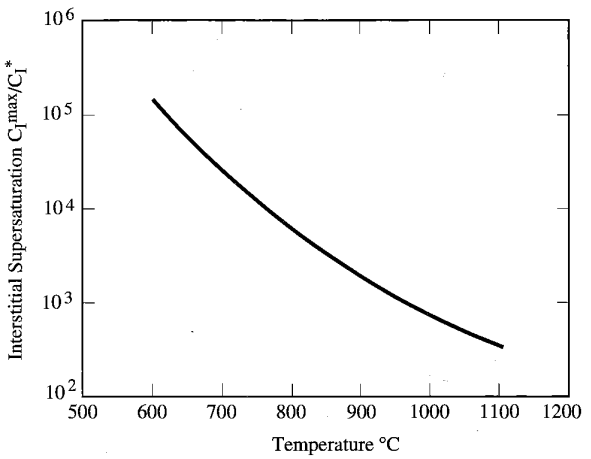


Figure 8–35 The maximum interstitial supersaturation as a function of anneal temperature.

The enhancement in diffusivity observed during TED is determined by the balance between cluster growth and evaporation. As long as there are a sufficient number of clusters around, some will shrink and others will grow, maintaining a steady-state interstitial supersaturation. It is during this steady-state condition that most of the profile motion occurs because the interstitial supersaturation is large and constant. This supersaturation has been measured experimentally and is as large as 10,000 at 750°C [8.21]. It reduces to a factor of 800 at a typical RTA temperature of 1000°C.

An example is shown in Figure 8–36, where the effects of TED on boron profile motion are dramatic for times as short as 1 minute at 750°C. The reason is that the normal, inert diffusivity of the boron is enhanced by a factor of 10,000 during the steady-state period when the clusters exist. The clusters continually supply interstitials to maintain this supersaturation by dissolving slowly, and the excess interstitials diffuse into the bulk producing the flat supersaturation profile and also recombine at the surface. After approximately 10 minutes at 750°C, the interstitial supersaturation decays rapidly and the boron profile will have experienced the maximum profile motion. Further annealing at 750°C will move the profile only with the inert diffusivity.

A second critical parameter is how long this steady-state condition lasts. If we assume the "+1" model for damage, then an implanted dose Q introduces an equivalent dose Q of excess interstitials which rapidly form clusters, and we wish to determine how long these excess interstitials survive. The interstitials in the clusters will eventually diffuse into the bulk or recombine at the surface. Since the damage is in general near the surface, the surface is the dominant sink for the excess interstitials and the flux of interstitials toward the surface can be estimated from Figure 8–37 to be $d_I C_I^{max}/R_p$, where R_p is the range of the implant.

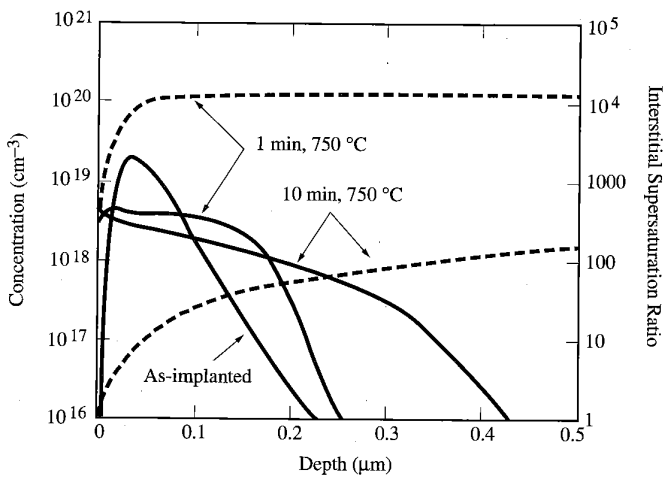


Figure 8–36 TSUPREM IV simulations [8.1] of boron TED (solid lines) at 750°C. A 10-keV, 10^{14} cm^{-2} boron implant experiences a steady-state enhancement of 10^4 while TED lasts at 750°C. The TED effect is beginning to decay rapidly after 10 min at 750°C. The dashed lines are the interstitial supersaturations.

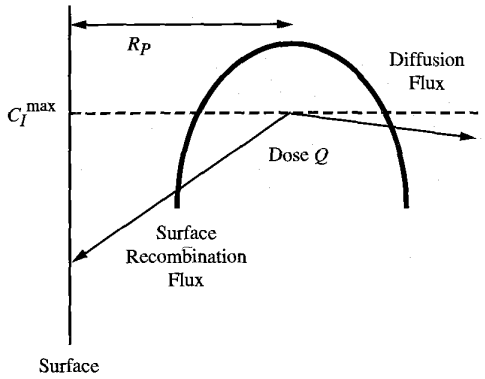


Figure 8–37 A schematic of the recombination flux at the surface which removes the +1 damage introduced by an implant at a range R_p [8.27].

The time to dissolve the clusters is given by the dose divided by the flux,

$$\tau_{enh} = \frac{4\pi a^3 R_P Q}{d_I} \exp\left(\frac{E_b}{kT}\right) \tag{8.44}$$

To clearly see the activation energy for the steady-state period, we can substitute for the self-interstitial diffusivity

$$d_I = d_I^0 \exp\left(-\frac{E_m}{kT}\right) \tag{8.45}$$

giving

$$\tau_{enh} = \frac{4\pi a^3 R_P Q}{d_I^0} \exp\left(\frac{E_b + E_m}{kT}\right) \tag{8.46}$$

Example Calculate and plot how the duration of TED depends on temperature for a 40-keV, 1×10^{14} cm^{-2} phosphorus implant into silicon.

Answer From Figure 8–3, a 40-keV phosphorus implant has a range of 60 nm:

$$\therefore \tau_{enh} = \frac{4\pi a^3 R_P Q}{d_I^0} \exp\left(\frac{E_b + E_m}{kT}\right)$$

$$= \frac{4\pi(2.7 \times 10^{-8}\text{ cm})^3(60 \times 10^{-7}\text{ cm})(1 \times 10^{14}\text{ cm}^{-2})}{51[\text{cm}^2\text{ sec}^{-1}]} \exp\left(\frac{1.77 + 1.8}{kT}\right)$$

$$= 2.9 \times 10^{-15} \exp\left(\frac{3.57}{kT}\right) \text{ sec}$$

which is plotted in Figure 8–38.

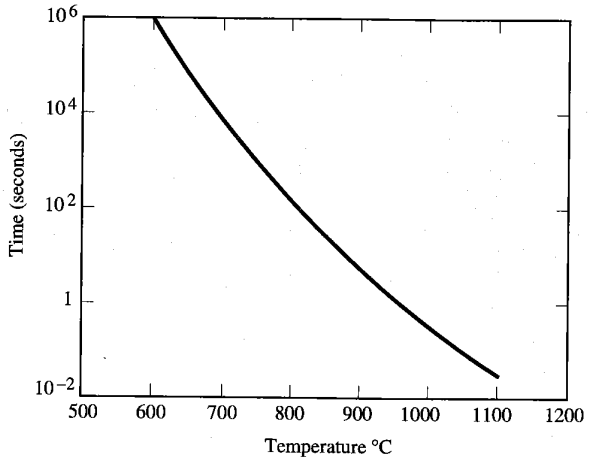


Figure 8–38 Duration of TED plotted versus temperature.

We can estimate the overall amount of profile motion after TED by combining the above information and taking the product of the dopant diffusivity, the steady-state interstitial supersaturation, and the steady-state time to give

$$D_A^{\text{eff}} t_{\text{eff}} = \left(D_A \frac{C_I^{\max}}{C_I^*} \right) \tau_{\text{enh}} = D_A \times \frac{QR_P}{d_I C_I^*} = \frac{D_A}{f_I^{\text{self}} D^{\text{self}} C_{\text{Si}}} \times QR_P \qquad (8.47)$$

which immediately makes clear the anomalous temperature dependence of TED. Though the dopant diffusivity has an activation energy of approximately 3.5 eV, it is overwhelmed by the activation energy of self-diffusion D^{self} (i.e., $E_F + E_m$) $\approx$ 4.8 eV, which causes more profile motion at lower temperatures. Thus, this equation neatly explains the single most anomalous observation about TED—the fact that profiles diffuse more at lower temperatures than at higher temperatures, even for similar Dt as shown in the experimental SIMS plots in Figure 8–31. Though puzzling, this observation can be rationalized by considering that a fixed amount of damage is introduced by the implant, while the background point-defect concentration and interstitial self-diffusion coefficient fall sharply with temperature. Thus, the supersaturation in the point-defect levels rises sharply with falling temperature and the excess defects remain around longer due to the lower interstitial diffusion coefficient, leading to greater overall motion in the dopant profile. A simulation of how this comes about is shown in Figure 8–39, where the higher interstitial supersaturation which also lasts longer at low temperatures causes more profile motion at the lower temperatures.

The general form of the TED dependence in Eq. (8.47) also indicates that the amount of profile motion measured by $\sqrt{Dt}$ has a square root dependence on the dose of damage and also on the range at which the damage is introduced. For implants close to the surface, this also contains information on how the energy of the implant will

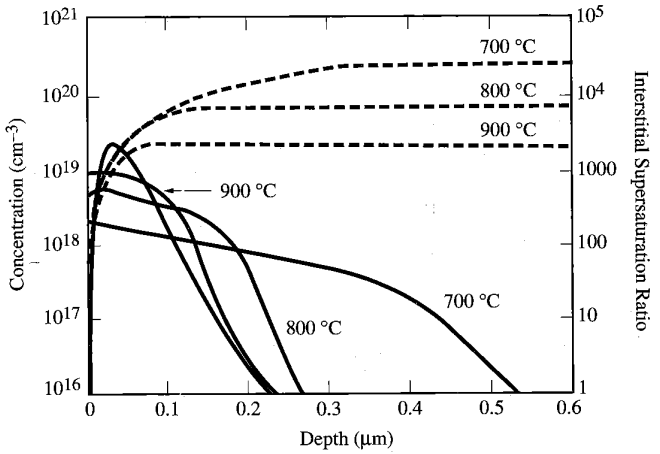


Figure 8–39 TSUPREM IV simulations [8.1] of boron profile motion after TED at various temperatures. The steady-state enhancement in the interstitial supersaturation is shown on the right axis and is higher at low temperatures and also lasts longer at low temperatures, leading to more profile motion at low temperatures.

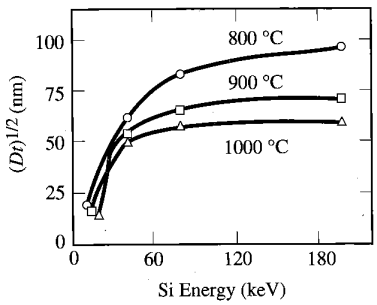


Figure 8–40 Energy dependence of TED. In these experiments, an experimental structure similar to that shown in Figure 8–32 was used, with different Si implant energies used to create the surface damaged region and the impact of TED on a buried marker layer was observed. (After [8.22].)

affect the overall profile motion. Since the range is roughly linearly proportional to the energy, it indicates that higher energies will give rise to higher overall profile motion. For large enough energies, the surface is no longer the dominant sink and diffusion to the bulk limits the flux away from the clusters. At this point, the amount of profile motion will be independent of the implant energy. Higher implant energies do give rise to more TED, as shown in Figure 8–40. The explanation appears to be obvious, in that higher implant energies create more damage. Though eminently reasonable, this is not the whole story for the energy dependence of TED and the physics contained in Eq. (8.47) provides a more complete explanation of the shape of the curves versus energy.

Another puzzling experimental observation relates to the dose dependence of TED—how do different damage doses give rise to the same amount of TED for short anneal times, but more damage produces more TED for longer anneal times? An example of

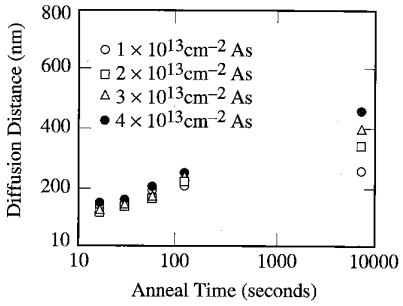


Figure 8–41 Dose dependence of TED. In these experiments, an experimental structure similar to that shown in Figure 8–32 was used, with different As implant energies used to create the surface damaged region and the impact of TED on a buried marker layer was observed. (After [8.22].)

the amount of profile motion after different As implant damage doses is shown in Figure 8–41. These data suggest that there is a maximum cap in the amount of enhancement that a dopant can experience, but a detailed understanding of the dose dependence requires an atomic-level model for the damage evolution. The microscopic cluster mechanism solves the dilemma of the unusual dose dependence of TED. Supersaturations higher than C_I^{max}/C_I^* do not exist because of the balance between cluster growth and dissolution. This provides a maximum cap on the diffusion enhancement regardless of the implant dose. Larger implant doses simply mean that the steady-state interstitial supersaturation is maintained for a longer time period before the clusters eventually decay. This provides the atomic-level explanation for why the enhancement is the same for different doses at short times but is greater for higher doses at longer times.

Thus, two parameters control the duration of TED for practical purposes—the critical supersaturation level and the time that the steady-state condition lasts, as shown schematically in Figure 8–42. This has been validated experimentally for As, P, and Si implant damage [8.21]. Boron TED appears to be more complex, perhaps because of the enormous strain that boron atoms introduce in the lattice, leading to the preferential formation of Boron Interstitial Complexes or BICs [8.28]. There are fewer {311} defects seen in high-concentration boron layers, but there is still substantial TED. Small dots that appear on TEM images may be related to Boron Interstitial Clusters or BICs. There appears to be competition between BICs and {311} defects for the excess interstitials. The BICs appear to have a higher binding energy for interstitials, so retain them longer during an anneal, giving rise to a longer time transient for boron TED than for the other dopants. However, the general picture holds, in that the magnitude and duration of TED are controlled by the evaporation of interstitials from microscopic defect aggregates.

8.5.11 Effects on Devices

MOSTEK in the early seventies was the first company to make major use of ion implantation to adjust the threshold voltage in MOS transistors. This application is ideally suited to the strengths of ion implantation and is still one of its key uses today. Because ion implantation can control even very low implant doses accurately, this allows the threshold voltage of MOS devices to be made uniform, without regard to the initial

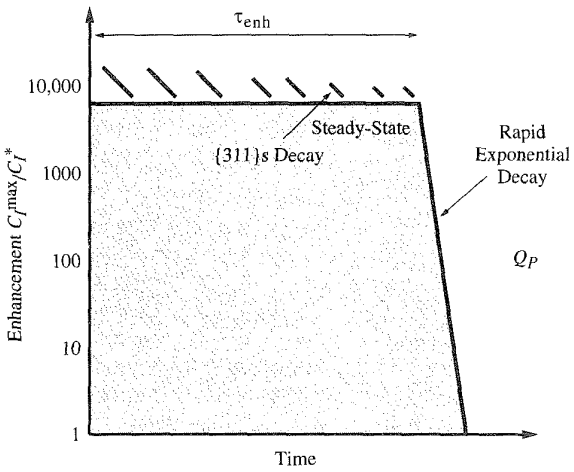


Figure 8–42 Schematic of the magnitude and duration of TED, determined from Eqs. (8.42) and (8.46). The shaded area is primarily responsible for the profile motion observed during TED.

variations in the substrate. When an ion implant dose Q is introduced into the channel of a MOS device such that all of the dose is contained in the depletion region during normal operation, this introduces a threshold voltage shift of

$$\Delta V_T = \frac{qQ}{C_{OX}} \tag{8.48}$$

where C_{OX} is the oxide capacitance per unit area, given by

$$C_{OX} = \frac{\varepsilon_0}{x_O} \tag{8.49}$$

In this way, MOS transistors can be fabricated on a lightly doped substrate or in a lightly doped well without much regard to the precise surface doping level. Later in the process, a threshold adjust implant can tailor the surface concentration in the channel region to the precise value needed for correct device operation. The ability of ion implantation to control the dose of ions introduced means that uniform thresholds can be obtained. This approach is universally used today and was illustrated in the CMOS process in Chapter 2.

Ion implantation has also made a major contribution to MOS scaling by allowing a deeper implant under the channel to be introduced, thereby avoiding punchthrough between the source and the drain through the bulk of the silicon. Once predicted to be a limiting factor on the scaling of MOS devices, this bulk punchthrough can now easily be avoided by using a punchthrough implant below the channel of the MOSFET. Typically, this is a higher dose than used for the threshold adjust implant, but the ability of ion implantation to introduce a retrograde profile below the surface was a key factor to its success.

Today, the focus of much effort is on even deeper implants to replace the deep well diffusions in a CMOS device. The ability to introduce a retrograde well profile, a punchthrough implant, and a threshold adjust implant through the same mask allows for considerable process simplification and mask savings. The reader may note that the implants in Figures 2–10 and 2–22 in Chapter 2 use exactly the same mask, for example.

Quite apart from the possibility that dislocations may be introduced by implants, there are more subtle effects on devices that are caused by ion-implant-induced TED. One example is the so-called Reverse Short-Channel Effect (RSCE), where a gradient in the interstitial profile toward the gate oxide causes the channel dopant to pile up under the gate giving rise to a higher threshold voltage for shorter gate lengths [8.29]. (Recall Figure 7–45 and the associated text.) Understanding TED allows suggestions to be made for avoiding the RSCE. For example, the dopant can be piled up in both long- and short-channel devices by using a silicon damage implant at the same time as the threshold adjust implant. The existing gradient of dopant atoms will inhibit further pileup, thus reducing the RSCE effects from later thermal anneals and source/drain implants. Alternatively, indium can be used as the channel dopant instead of boron. Even though indium exhibits diffusion characteristics very similar to boron, including identical susceptibility to TED, it segregates very strongly to an oxide interface. Thus, while some of the channel dose is lost to the oxide interface, further indium does not pile up at the interface during TED. It thus maintains its retrograde profile, which gives rise to less severe RSCE.

In summary, recent advances in understanding the physically based origins of TED on an atomic scale have led to breakthrough abilities to model its impact on devices. Key device effects such as DIBL, short-channel rolloff behavior, and RSCE can now be quantitatively modeled.

8.6 Limits and Future Trends in Technologies and Models

Ion implantation seems destined to retain a key role in doping semiconductors into the foreseeable future. The limits of the technology range from energies low enough to "soft-land" dopant atoms on the silicon surface at one extreme, to energies in the MeV range which may provide a replacement for buried layer epitaxial wafers. Energies at either extreme can be realized in a conventional implanter, consisting of ion separation, ion acceleration, and ion counting in a Faraday cage.

Alternative technologies have been suggested in the ultralow-energy regime, where conventional ion sources may have difficulty producing the required beam current. When the implant energies are in the hundreds of electron volts, they are comparable to the ion energy that is available across the sheath in an ion plasma (see Chapters 9 and 10). A plasma source can provide a large ion current and thus allow high doses to be implanted at low energies. The one drawback is that there is no ion separation possible, so multiple ion complexes are implanted into the wafer. In spite of this drawback, plasma source ion implantation is being investigated as a high-dose, low-energy doping method.

Diffusion from doped layers still remains attractive for obtaining the shallowest possible junctions, without the complication of transient-enhanced diffusion caused by implant damage. The possibility that the ultrashallow tip or extension region in a MOS

device can be formed by out-diffusion from a doped spacer layer remains attractive. Dopant is introduced at the solid solubility limit without accompanying damage, providing the minimum possible junction depth.

The future trends in ion implantation are dominated by the need for the equipment manufacturer to supply a process module that will provide a shallow junction meeting the NTRS requirements. Many of the auxiliary steps not directly related to the ion implant equipment play a key role, such as how to anneal the implant damage to provide a shallow junction with high activation. The equipment maker must also be familiar with the associated technology to provide the complete solution required by the semiconductor manufacturer. The trend is for the equipment makers to combine several steps, such as ion implant and RTA, and provide detailed process recipes that produce an optimized process step. In the future, it is likely that a detailed physical understanding of the underlying atomic-scale mechanisms that affect implantation and diffusion will be required by the equipment process developer. Models for the process steps, not just raw machinery, will likely be provided to the semiconductor customer.

8.7 Summary of Key Ideas

Ion implantation is a technique which combines ion separation, acceleration, and dose measurement to introduce a precise number of dopant atoms into the silicon substrate. The ions that impact the silicon wafer slow down by a combination of nuclear and electronic interactions with the lattice atoms. Ion implant energies range from hundreds of eV to millions of eV, with higher energy producing a deeper doping profile. This energy is much larger than the 15 eV needed to displace a silicon atom from the lattice and create a stable interstitial and vacancy pair. Large numbers of displaced atoms are created by a single implanted ion. The damage can accumulate and eventually produce a completely disordered or amorphous layer where the implant has occurred. Damage produced by an implant causes a large enhancement in the diffusion coefficient of dopants, until such time as the damage is completely repaired. This effect is known as transient-enhanced diffusion of ion-implanted layers and often limits how shallow a junction can be made.

Simple models for the ion implanted profile rely on a Gaussian distribution, though higher-order moments are needed to characterize profiles even in amorphous material. The relatively open structure of the silicon lattice leads to channeled profiles at low doses, where ions that enter a planar or axial channel travel further. Accurate models for ion implanted profiles in silicon require that the nuclear and electronic collisions be considered for a number of ions so that the implanted profile can be constructed. Provided that means can be found to minimize TED and provide highly active layers, ion implantation is likely to remain the technology of choice for doping silicon.

8.8 References

[8.1]. TSUPREM4 is a version of SUPREM-IV from Avant! Inc. SUPREM IV was originally written at Stanford University by M. E. Law, C. S. Rafferty, and R. W. Dutton.

[8.2]. Simulation performed with computer code UT-MARLOWE, distributed by the University of Texas, Austin.

[8.3]. W. K. Hofker, D. P. Oosthoek, N. J. Koeman, and A.M. De Grefte, "Concentration Profiles of Boron Implantations in Amorphous and Polycrystalline Silicon," *Radiation Effects*, vol. 24, no. 4, p. 223, 1975.

[8.4]. S. J. Morris, B. Obradovic, S. H. Yang, and A. F. Tasch, "Modeling of Boron Phosphorus and Arsenic Implants into Single-crystal Silicon over a Wide Energy Range (Few keV to Several MeV)," *IEDM Technical Digest*, p. 721, 1996.

[8.5]. J. F. Ziegler, "The Stopping and Range of Ions in Solids," in *Ion Implantation Science and Technology*, edited by J. F. Ziegler, Academic Press, 1984.

[8.6]. W. Brandt and M. Kitagawa, "Effective Stopping-power Charges of Swift Ions in Condensed Matter," *Physical Review B (Condensed Matter)*, vol. 25, no. 9, p. 5631, 1982.

[8.7]. J. F. Gibbons, "Ion Implantation in Semiconductors, Part 1, Range Distribution Theory and Experiments," *Proc. IEEE*, March 1968, vol. 56, no. 3, p. 295.

[8.8]. D. Cai, C. M. Snell, K. M. Beardmore, and N. Gronbech-Jensen, "Simulation of Phosphorus Implantation into Silicon with a Single parameter Electronic Stopping Power Model," *International Journal of Modern Physics C*, vol. 9, no. 3, p. 459, 1998.

[8.9]. I. Brodie and J. J. Muray, *The Physics of Microfabrication*, Plenum Press, 1982.

[8.10]. M. J. Caturla, T. Dias de la Rubia, and G. H. Gilmer, "Disordering and Defect Production in Silicon by keV Ion Irradiation Studied by Molecular Dynamics," *Nucl. Instrum. Methods Phys. Res. B (Beam Interactions with Materials and Atoms)*, vol. B106, no. 1–4, p. 1, 1995.

[8.11]. S. Tian, M. F. Morris, S. J. Morris, B. Obradovic, G. Wang, A. F. Tasch, and C. M. Snell, "A Detailed Physical Model for Ion Implant Induced Damage in Silicon," *IEEE Transactions on Electron Devices*, vol. 45, no. 6, p. 1226, 1998.

[8.12]. W. P. Maszara and G. A. Rozgonyi, "Kinetics of Damage Production in Silicon During Self-implantation," *Journal of Applied Physics*, vol. 60, no. 7, p. 2310, 1986.

[8.13]. M. Jariz, G. H. Gilmer, J. M. Poate, and T. D. de la Rubia, "Atomistic Calculations of Ion Implantation in Si, Point Defect and Transient Enhanced Diffusion Phenomena," *Applied Physics Letters*, vol. 68, no. 3, p. 409, 1996.

[8.14]. G. Z. Pan, K. N. Tu, and A. Prussin, "Size-distribution and Annealing Behavior of End-of-range Dislocation Loops in Silicon-implanted Silicon," *Journal of Applied Physics*, vol. 81, no. 1, p. 78, 1997.

[8.15]. J. R. Liefting, J. S. Custer, R. J. Schreutelkamp, and F. W. Saris, "Dislocation Formation in Silicon Implanted at Different Temperatures," *Materials Science and Engineering B (Solid-State Materials for Advanced Technology)*, vol. B15, no. 2, p. 173, 1992.

[8.16]. L. Csepregi, E. F. Kennedy, T. J. Gallagher, J. W. Mayer, and T. W. Sigmon, "Reordering of Amorphous Layers of Si Implanted with P, As and B Ions," *J. Appl. Physics*, vol. 48, p. 4234, 1977.

[8.17]. J. Fletcher, J. Narayan, and O. W. Holland, "Studies of Defects and Solubility Limits in SPE grown In and Sb Implanted Silicon," *Microscopy of Semiconducting Materials*, 1981, Proceedings of the Second Oxford Conference, p. 295, editors, A. G. Cullis and D. C. Joy, 1981.

[8.18]. B. L. Crowder and F. F. Morehead, "Annealing Characteristics of N-type Dopants in Ion Implanted Silicon," *Appl. Phys. Lett.*, vol. 14, p. 313, 1969.

[8.19]. T. E. Seidel and A. U. MacRae, "The Isothermal Annealing of Boron Implanted Silicon," in *First International Conference on Ion Implantation*, editors, F. Eisen and L. Chadderton, Gordon and Breach, 1971.

[8.20]. S. W. Crowder, "Processing Physics in Silicon-on-Insulator Material," Ph.D. thesis, Stanford University, 1995.

[8.21]. H. S. Chao, S. W. Crowder, P. B. Griffin, and J. D. Plummer, "Species and Dose Dependence of Ion Implantation Damage Induced Transient Enhanced Diffusion," *Journal of Applied Physics*, vol. 79, no. 5, p. 2352, 1996.

[8.22]. P. A. Packan and J. D. Plummer, "Transient Diffusion of Low-concentration B in Si Due to Si Implantation Damage," *Applied Physics Letters*, vol. 56, no.18, p. 1787, 1990.

[8.23]. D. J. Eaglesham, P.A. Stolk, H.-J. Gossmann, and J. M. Poate, "Implantation and Transient B Diffusion in Si: The Source of the Interstitials," *Applied Physics Letters*, vol. 56, no.18, p. 2305, 1994.

[8.24]. J. L. Benton, S. Libertino, P. Kringhoj, D. J. Eaglesham, J. M. Poate, and S. Coffa, "Evolution from Point to Extended Defects in Ion Implanted Silicon," *Journal of Applied Physics*, vol. 82, no. 1, p. 120, 1997.

[8.25]. M. D. Giles, S. Yu, H.W. Kennel, and P.A. Packan, "Modeling Silicon Implantation Damage and Transient Enhanced Diffusion Effects for Silicon Technology Development," in *Defects and Diffusion in Silicon Processing Symposium*, San Francisco, 1997, edited by T. Diaz de la Rubia, S. Coffa, P.A. Stolk, and C. S. Rafferty, Materials Research Society, 1997.

[8.26]. M. D. Giles, "Transient Phosphorus Diffusion from Silicon and Argon Implantation Damage," *Applied Physics Letters*, vol. 62, no. 16, p. 1940, 1993.

[8.27]. C. S. Rafferty, G. H. Gilmer, M. Jaraiz, D. Eaglesham, and H.-J. Gossmann, "Simulation of Cluster Evaporation and Transient Enhanced Diffusion in Silicon," *Applied Physics Letters*, vol. 68, no. 17, p. 2395, 1996.

[8.28]. K. S. Jones, J. Liu, and L. Zhang, "Evidence of Two Sources of Interstitials for TED in Boron Implanted Silicon," in *Proceedings of the Fourth International Symposium on Process Physics and Modeling in Semiconductor Technology*, p. 116, edited by G. R. Srinivasan, C. S. Murthy, and S. T. Dunham, Pennington, NJ, Electrochemical Society, 1996.

[8.29]. C. S. Rafferty, H.-H. Vuong, S. A. Eshraghi , M. D. Giles, M. R. Pinto, and S. J. Hillenius, "Explanation of Reverse Short Channel Effect by Defect Gradients," *IEDM Technical Digest*, p. 311, 1993.

8.9 Problems

8.1. Arsenic is implanted into a lightly doped p-type Si substrate at an energy of 75 keV. The dose is 1×10^{14} cm^{-2}. The Si substrate is tilted 7° with respect to the ion beam to make it appear amorphous. The implanted region is assumed to be rapidly annealed so that complete electrical activation is achieved. What is the peak electron concentration produced?

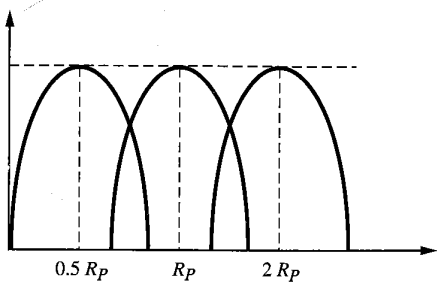
8.2. An engineer worried about avoiding punchthrough in an NMOS device decides to perform a deep punchthrough implant while keeping the surface concentration at a maximum value of 1×10^{17} cm^{-3}. What is the maximum implant dose that would be suitable if the peak of the profile is aligned at twice the depth of the deep source/drain contact junction in a 180 nm technology? (See NTRS Table 7–1 for parameter values.)

8.3. We want to design an implant step which will implant phosphorus ions through 50 nm of SiO_2 into an underlying silicon substrate such that the peak concentration in the substrate is 1×10^{17} cm^{-3} and the concentration at the SiO_2/Si interface is

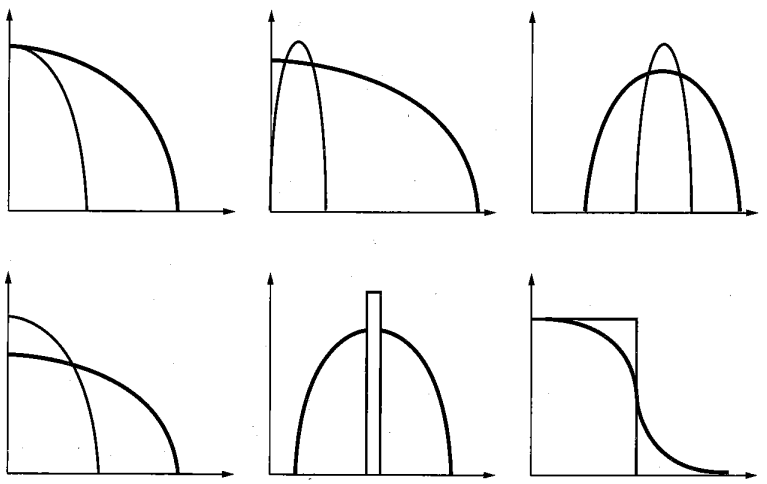
1×10^{15} cm^{-3}. What energy and dose would you use to achieve these conditions? Assume that the stopping power of SiO_2 is the same as that of silicon. Neglect channeling effects.

8.4. How thick does a mask have to be to reduce the peak doping of an implant by a factor of 10,000 at the mask/substrate boundary? Provide an equation in terms of the range and the standard deviation of the implant profile.

8.5. In a particular application, it is important to produce a fairly flat profile over an extended distance by ion implantation, as indicated below. One way to do this is to superimpose several implants at different energies. If phosphorus implants with energies of 0.5 R_P, R_P, $2R_P$ are used with a dose of 1×10^{14} cm^{-2} in the middle peak, approximately what doses should be used in the adjacent peaks if the initial peak concentrations are to be the same?



8.6. The equations below provide a reasonable analytical description for some of the diffusion processes indicated schematically in the diagrams on the following page. Put the equation number [(a)–(f)] on each figure that is the best match. Equations may be reused, or multiple equations may describe the same figure. A brief explanation is required for each figure.



a. $C(x,t) = \frac{C}{2}\,\mathrm{erfc}\left(\frac{x}{2\sqrt{Dt}}\right)$

b. $C(x,t) = \frac{Q}{\sqrt{\pi(D_1t_1 + D_2t_2)}}\exp\left(-\frac{x^2}{4(D_1t_1 + D_1t_1)}\right)$

c. $C(x,t) = \frac{Q}{2\sqrt{\pi Dt}}\exp\left[-\left(\frac{x}{2\sqrt{Dt}}\right)^2\right]$

d. $C(x,t) = C\,\mathrm{erfc}\left(\frac{x}{2\sqrt{Dt}}\right)$

e. $C(x,t) = \frac{Q}{\sqrt{\pi Dt}}\exp\left(-\frac{x^2}{4Dt}\right)$

f. $C(x,t) = \frac{Q}{\sqrt{2\pi(\Delta R_P^2 + 2Dt)}}\exp\left(-\frac{(x - R_P)^2}{2(\Delta R_P^2 + 2Dt)}\right)$

8.7. An 80-keV, 5×10^{13} cm^{-2} boron implant is performed into bare silicon. A subsequent anneal at 950°C is performed for 60 min.

a. Can the annealed profile evolution be described by the following formula?

$$C(x,t) = \frac{Q}{\sqrt{\pi D_B t}}\exp\left(-\frac{x^2}{4D_B t}\right)$$

b. Assume that all the dopant remains in the silicon and that none evaporates to the ambient. By considering a virtual or imaginary image profile on the ambient side of the interface (a reflecting boundary), calculate the surface concentration after a 60-min 950°C anneal.

8.8. A 1×10^{14} cm^{-2} phosphorus implant through a 200-nm SiO_2 mask layer is performed so the peak concentration is at the silicon/SiO_2 interface. An anneal is then performed for 30 min at 1000°C. Calculate the location of the junction with the substrate doped at 1×10^{15} cm^{-3}. Assume no diffusion in the masking layer and ignore any segregation effects. Assume the same range statistics for SiO_2 and Si.

8.9. Phosphorus is implanted at 50 keV with a dose of 1×10^{14} cm^{-2}. Calculate the junction depth where the phosphorus meets the substrate (p type, 1×10^{15} cm^{-3}) after (a) a 900°C, three-hour anneal and (b) a 1200°C, three-hour anneal.

8.10. An implant machine for 300-mm wafers is required to have a throughput of 60 wafers per hour. What beam current is required in order to implant a source/drain region in a CMOS device with a dose of 1×10^{15} cm^{-2}?

8.11. In the ion implantation process, positively charged ions impact on the semiconductor surface. Normally these ions are neutralized by capturing an electron from the conducting substrate. However, when the mask is an insulator like SiO_2, the charge on the ions may not be neutralized as easily. Consider the case where a dose Q is implanted into the surface of an SiO_2 layer (assume all the charge resides at the oxide surface). Further assume that the oxide can withstand an electric field of 10^7 Vcm^{-1} before it breaks down. What implant dose Q is required to cause electrical failure of the mask? That is, what dose will cause a field of 10^7 Vcm^{-1} across the oxide?

8.12. An engineer investigating solid-phase epitaxial regrowth after amorphizing ion implants of various species (P, B, Si, Ge, As, Sb) makes the following observations:

1. N-type dopants of very different size or atomic radius (e.g., antimony versus phosphorus) show identical regrowth rates, approximately an order of magnitude faster than the regrowth of silicon implanted and amorphized with silicon ions.
2. P-type and N-type ion implanted regions have much faster regrowth rates than Si or Ge implanted regions, although they are not identical.
3. B-doped regions compensated with an equal dose of an arsenic implant show identical regrowth rates to Ge implanted and amorphized regions.

Construct a unified physical explanation of all three phenomena, considering possibilities such as size or stress effects, dopant charge or electric field effects, or point-defect-based effects (no calculation required).

8.13. In two separate experiments, As and then B are implanted through a thin SiO_2 layer into the underlying substrate. As a result of the implantation, some of the oxygen atoms in the SiO_2 layer are knocked into the silicon substrate. Would you expect the As or the B to produce more oxygen knock-ons? Why?

8.14. A boron implant is performed into silicon at 100 keV. The boron beam is aligned with the silicon crystal so that channeling is present. Estimate the range of the channeled boron profile by considering that electronic stopping is the only mechanism for slowing the boron ions.

8.15. An amorphizing implant is performed using a high dose of arsenic (5×10^5 cm^{-2}, 200 keV) and TEM cross-section images indicate complete amorphization to a depth of 50 nm. If Solid-Phase Epitaxial (SPE) regrowth instantly removes all of the damage in the amorphous region, calculate the fraction of the "+1" implanted dose that is available for TED.

8.16. Calculate the change in junction depth for a 40-keV boron threshold adjust implant of 5×10^{13} cm^{-2} annealed at 750°C in a furnace or at 1000°C in an RTA for a time just long enough to remove all the damage that causes TED. Assume a uniform well doping of 5×10^{16} cm^{-3}.

8.17. The diffusion of a buried antimony layer is only 20% of its normal diffusion in an inert ambient after a phosphorus implant is performed at the surface of a silicon wafer. The phosphorus diffusion coefficient is itself enhanced by a factor of 4 (400%).

a. Suggest a qualitative reason for this observation.

b. Assuming that the phosphorus diffuses completely by an interstitial mechanism, calculate how much of the antimony diffusion is mediated by interstitials and how much by vacancies.

8.18. A phosphorus implant is performed into a bare p-type silicon wafer with a background doping of 1×10^{15} cm^{-3}, at an energy of 50 keV, and a dose of 1×10^{14} cm^{-2}. After an anneal at 900°C for 10 min, the junction depth is measured to be 0.25 μm. Calculate the enhancement in the phosphorus diffusion coefficient that was caused by the implantation damage.

8.19. An engineer wants to form a shallow boron-doped source/drain junction for an advanced technology. The manager wants to know whether the company should buy an inexpensive batch furnace and achieve the required junction depth using a low-thermal budget anneal (one hour at 800°C) or an expensive single-wafer RTA (Rapid Thermal Annealer) using a high-temperature anneal.

a. Calculate the time required to achieve the target junction depth if the annealing temperature is 1050°C in the RTA.

b. Make a crude estimate ($\sqrt{Dt}$ approximation) of how far the dopants move during these anneals

c. Now, consider that the boron is introduced using an implant and that Transient Enhanced Diffusion (TED) due to the implant will be important. Using the charts for the expected enhancement in diffusivity and the time TED lasts, calculate how far the dopants move at each temperature.

d. Which anneal would you recommend?

8.20. An NMOS transistor is being built and an ion implantation is done after the gate oxide is grown and before the gate polysilicon deposition, in order to adjust the threshold voltage by +1volt.

a. Which dopant should be used?

b. Calculate the dose of the dopant if the oxide thickness is 10 nm.

8.21. A phosphorus implant is performed into a bare p-type silicon wafer with a background doping in $1 \times 10^{15}\text{cm}^{-3}$, at an energy of 50 keV and a dose of $1 \times 10^{14}\text{cm}^{-2}$. After an anneal at 900°C for 10 min, the junction depth is measured to be 0.25 μm. Calculate the time averaged enhancement in the phosphorus diffusion coefficient that was caused by implantation damage.

Thin Film Deposition

9

9.1 Introduction

The CMOS process flow in Chapter 2 contained many layers above the silicon substrate. These layers include dielectrics, semiconductors, and metals. In most cases these layers must be deposited which implies that there are typically many deposition steps in an IC process. In this chapter we describe the different techniques used to deposit these thin films. In Chapter 10 the techniques used to etch thin films are discussed.

We begin with several issues related to thin films and their deposition. These issues are important when evaluating deposited films and choosing deposition methods. The first issue is the "quality" of the deposited film. By quality we mean composition, contamination levels, defect density, and mechanical and electrical properties. The composition of films may vary depending on the deposition conditions and it is usually important to achieve a specific composition. This is especially important when a range of compositions is possible, such as with TiN films which can be Ti or N rich, or with alloys of metals like Al/Cu. Generally there needs to be a low level of contaminants, such as unwanted metals, water, oxygen, or halogens. There should also be a minimum of pin holes or other structural defects. These often result from particle contamination on the surface and are a common problem in thin film deposition (recall Figure 4–1). There should also be a minimum of stress in the films, and the films should be mechanically stable during subsequent processing. Finally, good adhesion to underlying films is important.

A second issue is that of uniform thickness across a wafer, from wafer to wafer, and as the film crosses nonplanar topography. Figure 9–1 shows a metal layer deposited over nonplanar topography in which there is a step in the underlying oxide. In Figure 9–1(b) the film thickness decreases as it crosses a step, which can lead to high electrical resistance in metal lines and a greater chance of mechanical cracking and failure. Coverage on the side of a step in topography is called step coverage. It is often defined quantitatively as the minimum thickness deposited on the side of a step divided by the thickness deposited on the top horizontal surface. Conformal step coverage, or conformal coverage, refers to uniform film thickness on both horizontal and vertical surfaces, or a step coverage of one.

A related issue is that of filling spaces between or within topographical structures. This includes filling a via or contact hole with metal, as illustrated in Figure 9–2(a), and filling spaces or gaps in shallow trench isolation structures or between metal lines with an oxide (often called gap filling), as illustrated in Figure 9–2(b). In this figure incomplete filling is shown, leading to a void in the dielectric between the lines. A void in a metal layer can lead to high-contact or -sheet resistance, and in a dielectric layer can

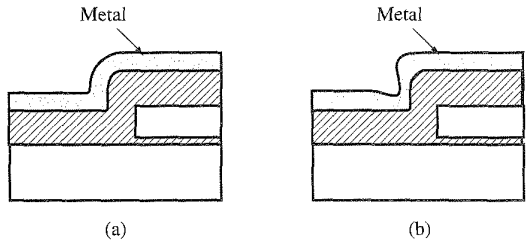


Figure 9–1 Step coverage of metal over nonplanar topography. (a) shows conformal step coverage, with constant thickness on horizontal and vertical surfaces. (b) shows poor step coverage.

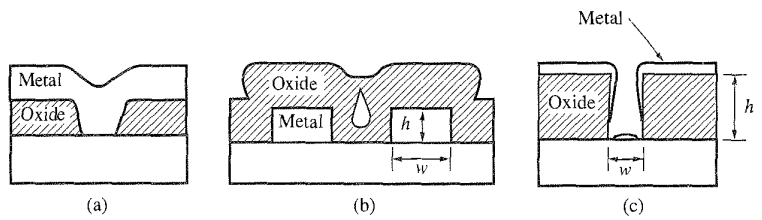


Figure 9–2 Thin film filling issues. (a) shows good metal filling of a via or contact hole in a dielectric layer. (b) shows silicon dioxide dielectric filling the space between metal lines, with poor filling leading to void formation. (c) shows poor filling of the bottom of a via hole with a barrier or contact metal.

result in cracking problems. In addition, a void may trap processing chemicals or moisture, leading to reliability problems. Even if complete filling of a space is not needed, good coverage or filling at the bottom of a space is often required. An example of this is the case of depositing contact and barrier layers, in which good filling of the bottom of contact or via holes is important. Figure 9–2(c) shows poor bottom coverage of a deep contact hole. The sides may also need adequate coverage of an adhesion layer.

An important parameter that can affect filling and bottom coverage is the Aspect Ratio (AR) of a feature, defined as the ratio of the height of a feature to its width:

$$AR = \frac{\text{height of feature}}{\text{width of feature}} = \frac{h}{w} \tag{9.1}$$

The feature could be a metal line, for example, or a space, such as a gap between two metal lines or a contact hole, as illustrated in Figure 9–2(b) and (c). A deep, narrow contact hole would have a large aspect ratio and would be harder to fill.

Two examples of poor filling and coverage are shown in Figure 9–3. Figure 9–3(a) shows poor step coverage of a TiW/Al/TiW metal stack layer over an oxide step. Figure 9–3(b) shows voids in an oxide layer for narrow spaces between metal lines. We will see that in filling or depositing in very deep contacts or vias, conditions that would lead to good step coverage may actually hinder adequate filling of the bottom of the contacts or vias.

We saw in the CMOS process flow in Chapter 2 that as more interconnect levels are used, another property that is desired for thin films is a smooth and flat top surface, as op-

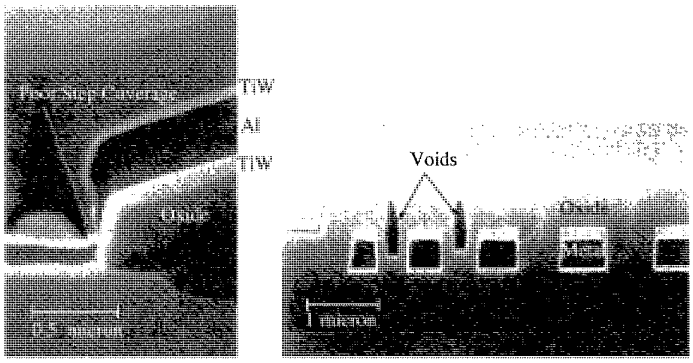


Figure 9–3 SEM images of coverage and filling problems. (a) shows poor step coverage of a TiW/Al/TiW metal stack layer, deposited by sputter deposition, over an oxide step. (b) shows voids in a Chemical Vapor Deposition (CVD) oxide layer for narrow spaces between metal lines. Photos courtesy of VLSI Technology, Inc.

posed to one that just follows the underlying topography. Figure 9–3(b) shows a relatively flat, planarized top surface of the oxide over the metal lines. While planarized topography can be achieved by postdeposition processing using etching or polishing techniques, we will also see how certain deposition techniques can also help attain this feature.

9.2 Historical Development and Basic Concepts

Methods of thin film deposition are usually separated into two main categories: Chemical Vapor Deposition (CVD) and Physical Vapor Deposition (PVD). In each case the silicon wafer is placed in a deposition chamber, and the constituents of the film are delivered through the gas phase to the surface of the substrate where they form the film. In the case of CVD, reactant gases are introduced into the deposition chamber, and chemical reactions between the reactant gases on the substrate surface are used to produce the film. In the case of PVD, physical methods are used to produce the constituent atoms which pass through a low-pressure gas phase and then condense on the substrate. The methods used to produce the atom flux in PVD include (1) heating a solid or molten source until it vaporizes, called "evaporation," and (2) bombarding a solid source with energetic ions formed in a plasma, called "sputter deposition." PVD is sometimes called "vacuum deposition" since very low-pressure environments are required for the transport of the gaseous species from the source to the film surface. (As we will see, low-pressure CVD is commonly used, but not at nearly as low pressures as in PVD.) CVD has historically been used in the integrated circuit industry mainly for silicon and dielectric deposition, primarily due to its good quality films and good step coverage. PVD has been used historically for metal deposition, mostly due to its ability to deposit a variety of metals and alloys not easily deposited by CVD.

A third category for thin film deposition is the technique of coating a wafer with a liquid film that forms a solid film when heated. Spin-On-Glasses (SOGs) are examples of this. This method, which is used for topography smoothing or "planarization," will be discussed in Chapter 11. A fourth category includes electrolytic deposition techniques. These have historically been used for printed circuit board fabrication but have recently been extended to the deposition of Cu interconnect layers.

9.2.1 Chemical Vapor Deposition (CVD)

In CVD, gases are introduced into the deposition chamber that react and form the desired film on the surface of the substrate. The silicon substrate will usually have other films deposited and patterned on it which then form a complex topography on which the new film must be deposited. Either a single gas is used that will decompose when heated to supply the necessary component or components for the film, or multiple gases are used which will react to produce the film. In either case, the film-producing reaction should take place on the surface of the substrate as opposed to in the gas stream. Gas stream reactions can create particulates that form above the wafer which can drop down onto the surface and cause the films to have pinholes or low density. Sometimes appropriate reactants are not available in gaseous form, in which case a liquid source may be used. A carrier gas such as hydrogen is bubbled through the liquid source and carries the source vapor to the reaction chamber. Nitrogen or argon may be used as a diluent or carrier gas.

The simplest CVD process uses an atmospheric deposition chamber, and no plasma-enhanced processes are used. This is appropriately called Atmospheric Pressure Chemical Vapor Deposition, or APCVD. The configuration shown in Figure 9–4(a) is a common type that has been used for APCVD, especially for epitaxial silicon deposition, but several other systems in a variety of configurations such as barrel-shaped chambers, have been used. The walls of the chamber in this case are not heated—a so-called "cold-wall" reactor. The wafers are heated by using a graphite susceptor which is heated by RF induction. This minimizes deposition on the reactor walls since only the susceptor and wafers are hot. APCVD has been used for many films, especially epitaxial silicon and silicon dioxide—both undoped and doped with boron and phosphorus. However, the increasing use of Low-Pressure CVD, or LPCVD, and Plasma-Enhanced CVD, or PECVD, for reasons to be discussed in the following sections, has made the use of APCVD reactors much less common today.

Figure 9–4(b) shows the low-pressure CVD configuration. In this case the wafers are stacked upright, and the deposition is performed at a reduced pressure. Heating is accomplished using resistive heating elements wrapped around the tube (a "hot-wall" reactor), which heats up the tube and everything inside. This is very similar to the oxidation system described in Chapter 6. Very uniform temperatures can be obtained in this type of system.

There are other types of CVD configurations and techniques, but all are based on the idea of supplying gaseous chemical sources and energy to the surface of the wafers. The energy is usually thermal energy, but one can greatly reduce the amount of thermal energy needed by using a plasma to provide for energetic ions and very reactive free radicals to stimulate the deposition.

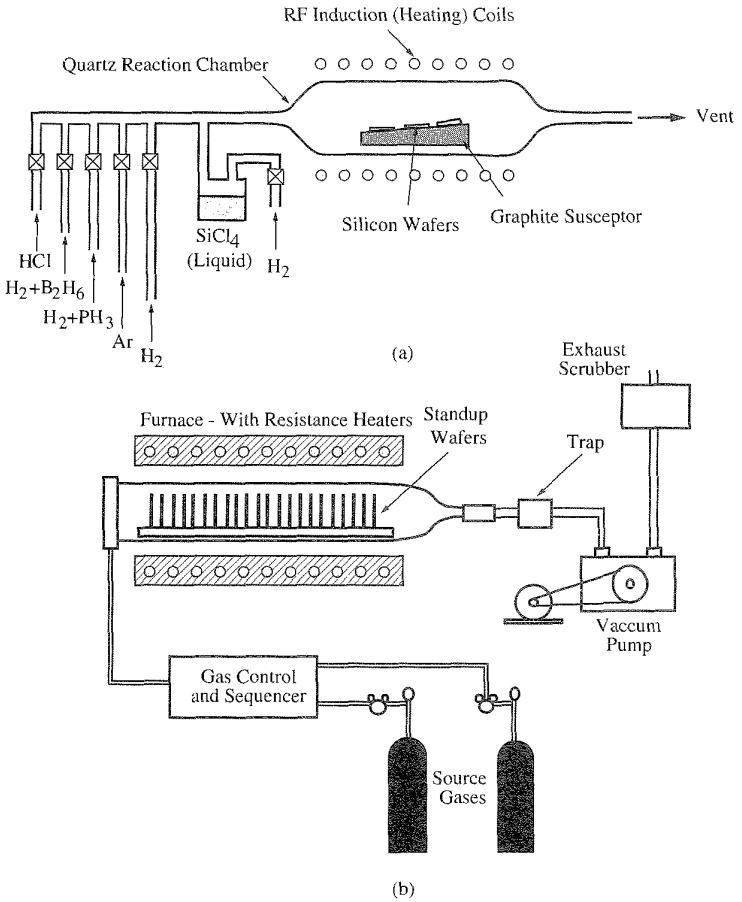


Figure 9–4 Chemical Vapor Deposition (CVD) systems. (a) is an atmospheric cold-wall system used for deposition of epitaxial silicon. (b) is a low-pressure hot-wall system used for deposition of polycrystalline and amorphous films, such as polysilicon and silicon dioxide, respectively.

9.2.1.1 Atmospheric Pressure Chemical Vapor Deposition (APCVD)

Although APCVD systems are not used as commonly today, we will begin with an APCVD system to describe some of the basic concepts of CVD. The steps involved in a CVD process, schematically illustrated in Figure 9–5, are

1. Transport of reactants by forced convection to the deposition region.
2. Transport of reactants by diffusion from the main gas stream through the boundary layer to the wafer surface.

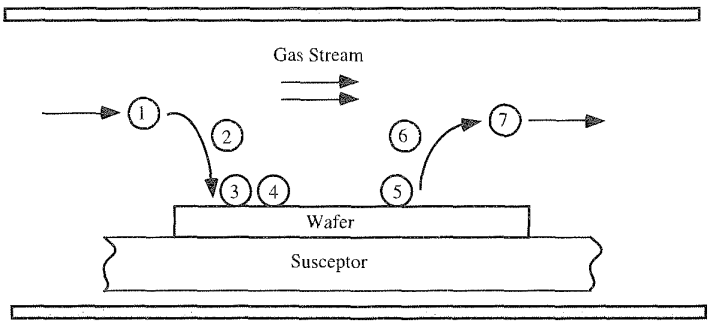


Figure 9–5 Steps involved in a CVD process. Numbered steps are explained in text.

3. Adsorption of reactants on the wafer surface.
4. Surface processes, including chemical decomposition or reaction, surface migration to attachment sites (such as atomic-level ledges and kinks), site incorporation, and other surface reactions.
5. Desorption of byproducts from the surface.
6. Transport of byproducts by diffusion through the boundary layer and back to the main gas stream.
7. Transport of byproducts by forced convection away from the deposition region.

A process not explicitly shown is that of desorption or "reemission," often just called "emission," of the reacting species before any reaction can occur. This is related to the sticking probability or coefficient, which will be discussed in more detail later. This process becomes important in coverage and filling issues, particularly in low pressure systems, since the desorbed species may deposit elsewhere such as on the sidewall of a trench or contact hole. In CVD processes the emission and redeposition process may occur quite readily, corresponding to a low sticking coefficient and can lead to better coverage. Surface migration or diffusion can also occur on a more long-range scale on nonflat topographies, due to surface curvature and surface energy reduction considerations for example. This can affect coverage and filling in a similar manner to reemission and redeposition. In the analysis here we are mainly concerned with deposition kinetics of thin films for relatively large-scale depositions on relatively flat surfaces. Local topography and coverage issues will be examined later focusing on low pressure CVD and PVD systems. We will also not consider gas-phase reactions, which sometimes produce intermediary vapor-phase species.

A full treatment of the kinetics of CVD thin film deposition, or growth, would involve all of the steps in Figure 9–5. However, steps 2–5 are usually the most important as far as determining the overall growth rate and thus we consider only those processes here. Step 2 is the mass transfer of reactants through the boundary layer. To simplify matters, we group steps 3–5 and classify them together as surface reactions. In our simple treatment

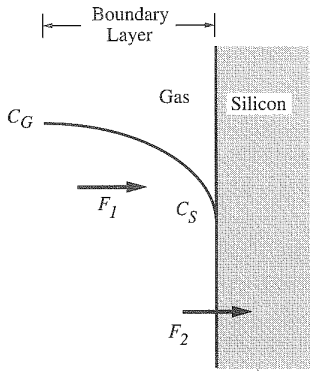


Figure 9–6 Wafer surface region in CVD process showing concentrations and fluxes of reactant species.

we consider the fluxes for the two important processes (mass transfer and surface reactions) and equate them under steady-state conditions. The treatment is very similar to that used for oxidation kinetics in Chapter 6 and for photoresist development in Chapter 5. Figure 9–6 shows the surface region of the depositing films, with the appropriate concentrations and fluxes in the gas-phase boundary layer and at the wafer surface.

The flux of reactant species from the gas phase to the wafer surface through the boundary layer, F_1 (in molecules cm^{-2} sec^{-1}), can be described by

$$F_1 = h_G(C_G - C_S) \tag{9.2}$$

The $(C_G - C_S)$ term is the difference in concentration of the reactant species (in molecules cm^{-3}) between the main gas flow and the wafer surface, and h_G is the mass transfer coefficient (in cm sec^{-1}). This flux represents the gas-phase diffusion through the stagnant boundary layer that forms between a flowing gas and a solid object. F_2 is the flux of reactant consumed by the reaction at the surface. Assuming first-order reaction kinetics, this can be written as

$$F_2 = k_S C_S \tag{9.3}$$

where k_S is the chemical surface reaction rate (in cm sec^{-1}) and C_S is the concentration of the reacting species at the surface (in molecules cm^{-3}). All of the processes involved with the surface chemical reactions and surface diffusion that occur during the actual film formation, steps 3–5 above, are lumped into this one parameter. This is very similar to the strategy used in the Deal-Grove oxidation model. [Recall Eq. (6.17).]

Assuming steady-state deposition conditions, these two processes that act in series must be equal to each other, and the overall process will proceed at the rate of the slower process. Thus

$$F = F_1 = F_2 \tag{9.4}$$

and equating Eqs. (9.2) and (9.3) leads to

$$C_S = C_G\left(1 + \frac{k_S}{h_G}\right)^{-1} \tag{9.5}$$

The growth rate of the film is now given by

$$v = \frac{F}{N} = \frac{k_S h_G}{k_S + h_G}\frac{C_G}{N} \tag{9.6}$$

where v is the deposition rate or velocity in cm sec^{-1}, and N is the number of atoms incorporated per unit volume in the film, or its density, in cm^{-3} (5×10^{22} cm^{-3} for the case of Si deposition).

We now define Y to be the mole fraction of the incorporating species in the gas phase:

$$Y \equiv \frac{C_G}{C_T} \tag{9.7}$$

where C_T is the concentration of all molecules in the gas phase. One example is the case of epitaxial Si deposition by the reaction of $SiCl_4$ and hydrogen:

$$SiCl_{4(g)} + 2H_{2(g)} \Leftrightarrow Si_{(s)} + 4HCl_{(g)} \tag{9.8}$$

Here the incorporating species is Si, C_G is the number of molecules of $SiCl_4$ per cm^3 in the gas phase and C_T would correspond to the total number of $SiCl_4$ and H_2 molecules (plus any other species) per cm^3 in the gas phase. Y is also equal to the partial pressure of the incorporating species, P_G, divided by the total pressure in the system such that

$$Y \equiv \frac{C_G}{C_T} = \frac{P_G}{P_{\text{Total}}} = \frac{P_G}{P_G + P_{G'} + \ldots} \tag{9.9}$$

where P_G is the partial pressure of $SiCl_4$ and $P_{G'}$ corresponds to the partial pressure of H_2. The partial pressures of any other gas species present, such as N_2 or Ar used as carrier or dilutant gases and any byproduct gases, would also have to be added to the total pressure term. The equation for the deposition velocity is now

$$v = \frac{k_S h_G}{k_S + h_G}\frac{C_T}{N}Y \tag{9.10}$$

Example Calculate the deposition rate for a CVD system in which

$h_G = 1.0$ cm sec^{-1}

$k_S = 10$ cm sec^{-1}

Partial pressure of incorporating species $= P_G = 1$ torr

Total pressure $= P_T = 1$ atm $= 760$ torr

Total concentration in gas phase $= C_T = 1 \times 10^{19}\ \text{cm}^{-3}$
Density of depositing film $= N = 5 \times 10^{22}\ \text{cm}^{-3}$

Answer By Eq. (9.10):

$$v = \frac{k_S h_G}{k_S + h_G}\frac{C_T}{N}Y = \frac{1}{\left(\frac{1}{h_G} + \frac{1}{k_S}\right)}\frac{C_T}{N}Y$$

$$v = \frac{1}{\left(\frac{1}{1.0\ \text{cm sec}^{-1}} + \frac{1}{10\ \text{cm sec}^{-1}}\right)}\frac{1 \times 10^{19}\ \text{cm}^{-3}}{5 \times 10^{22}\ \text{cm}^{-3}}\frac{1}{760}$$

$$= 2.4 \times 10^{-7}\ \text{cm sec}^{-1} = 0.14\ \mu\text{m min}^{-1}$$

From Eq. (9.10) one sees that the deposition velocity is determined by the smaller of k_S or h_G, leading to two limiting cases. If $k_S << h_G$, then

$$v \cong \frac{C_T}{N}k_S Y \tag{9.11}$$

This is the surface reaction controlled case. The mass transfer through the gas boundary layer is relatively fast, while the surface reaction is sluggish. From Eq. (9.5) one sees that the surface concentration, C_S, approaches C_G. If $h_G << k_S$, then

$$v \cong \frac{C_T}{N}h_G Y \tag{9.12}$$

This is the mass transfer, or gas-phase diffusion, controlled case. Here the surface reaction is quite fast compared to the mass transfer process. C_S approaches zero because the reactant species are used up at the surface as soon as they arrive. In the above example, $h_G << k_S$ and so that system is operating in the mass transfer controlled regime.

These two limiting cases are analogous to the oxidation cases in Figure 6–15. The analysis in Chapter 6 showed how thermal oxidation leads to two kinetic regimes, one described by a linear rate constant (interface reaction limited or controlled) and the other described by a parabolic rate constant (solid-state diffusion limited). In the case of CVD deposition, there are also two limiting cases: one reaction controlled and one diffusion controlled. But the growth rate in both regimes is linear with time. That is, both regimes have constant growth velocities, and neither is parabolic. This is because the reaction always occurs at the growing surface, and the diffusion-related process is mass transport through a gas region of constant thickness, not a growing solid region as in the oxidation process. The two CVD regimes are therefore not differentiated by different growth laws. However, the two processes dominate in different temperature regimes.

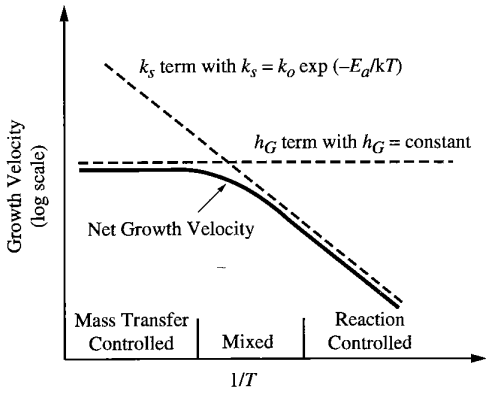


Figure 9–7 Arrhenius plot of growth velocity (or deposition rate) vs. $1/T$ for CVD process. The net growth velocity is the result of the surface reaction and gas-phase mass transfer processes acting in series so that the slower of the two dominates at any temperature.

The temperature dependence of the surface reaction controlled case is Arrhenius just as in the oxidation case. On the other hand, the mass transfer coefficient, h_G, is rather insensitive to temperature. Figure 9–7 plots the growth velocity versus reciprocal temperature. The k_S and h_G components are plotted separately showing their different temperature dependencies. When the two components are combined so that the smaller of the two components dominates (two processes acting in series), as described by Eq. (9.10), the resulting growth velocity behaves as shown in the figure. In practice, the cross-over temperature (where the two processes are equal) and the temperature ranges for the different regimes depend on a variety of factors, such as the specific gases used, the reaction activation energy, wafer configuration, the flow conditions in the reactor, and the pressure.

This basic model for CVD deposition has important implications for deposition conditions and equipment configuration. When the deposition occurs in the mass transfer controlled regime (higher temperatures), the deposition rate is relatively constant with temperature. However because it is limited by mass transfer of species through the gas-phase boundary layer, the flow of the gas over the wafers and transport of the reactants to the wafer surfaces are very important and can place major restrictions on the configuration of the equipment and the placement of the wafers. For example, the wafers must usually be placed face up and edge to edge in the chamber [as in Figure 9–4(a)] to ensure that equal gas flows reach each wafer surface and that equal deposition rates are achieved. By contrast, when the deposition occurs in the surface reaction controlled regime (lower temperatures), the process is very sensitive to temperature. However, the mass transfer through the boundary layer is not as important, leading to fewer restrictions on the gas flow and wafer placement. For example, the wafers may be stacked face-to-face [as in Figure 9–4(b)] without much of a penalty in wafer-to-wafer deposition uniformity.

Figure 9–8 shows typical experimental data for the deposition of silicon at atmospheric pressure. All four silicon sources exhibit similar types of behavior, with surface reaction limited regimes showing Arrhenius behavior, and mass transfer (or diffusion) limited regimes showing little temperature dependence. For all sources, the slopes of the curves in the reaction limited regimes are approximately the same, with an activation

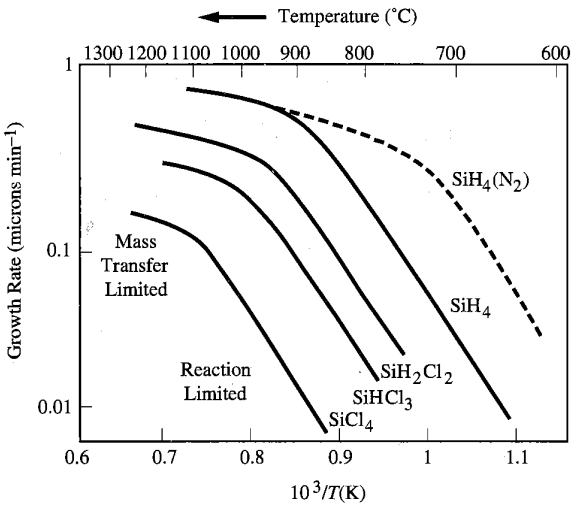


Figure 9–8 Growth or deposition rate versus $1/T$ for silicon deposited by CVD at atmospheric pressure using various silicon source gases. The partial pressure of the reactant gas is 0.8 torr. Hydrogen is used as the carrier or diluent gas for the solid curves. (After [9.1].) Using an inert gas such as N_2 for the diluent shifts the SiH_4 curve to the right, as indicated by the dashed curve and as discussed later in Section 9.3.2.

energy of about 1.6 eV. This suggests that the rate limiting surface reaction process is the same for each of the different sources and may be related to the desorption of hydrogen atoms from the Si surface. The choice of source gas does affect the overall growth velocity. The growth is fastest for a silane (SiH_4) source and slowest for a silicon tetrachloride ($SiCl_4$) source. The growth rate in the mass transfer limited regime is found to be approximately inversely proportional to the square root of the source gas molecular weight. There is also a substrate orientation effect on the growth kinetics. As expected, this is only important at lower temperatures where the surface reaction dominates. The (111) Si surface has a lower growth rate, presumably due to the fewer kink or attachment sites available for atom incorporation.

Epitaxial Si deposition has traditionally been done at higher temperatures to ensure that all the Si atoms being deposited are incorporated into lattice sites in order to obtain a single-crystal thin film. This means that epitaxial Si in these systems is deposited in the mass transfer regime where gas-phase processes are important. On the other hand, polysilicon is usually deposited at lower temperatures and in the surface reaction regime. Here the gas-phase processes are not important (so that stacking of wafers is possible), but the deposition is very temperature sensitive.

Earlier we assumed that the flux of reactant from the main gas flow to the wafer surface could be described simply by

$$F_1 = h_G(C_G - C_S) \tag{9.2}$$

where h_G is a constant. But gas flow in deposition chambers is more complicated than this, and it is worthwhile to take a little deeper look at it. We start with the configuration of Figure 9–4(a) where the wafers lie flat on a susceptor. The simplest model assumes that a completely stagnant boundary layer δ_S exists next to the surface of the wafer. This means that for some distance, away from the surface, the gas is not moving

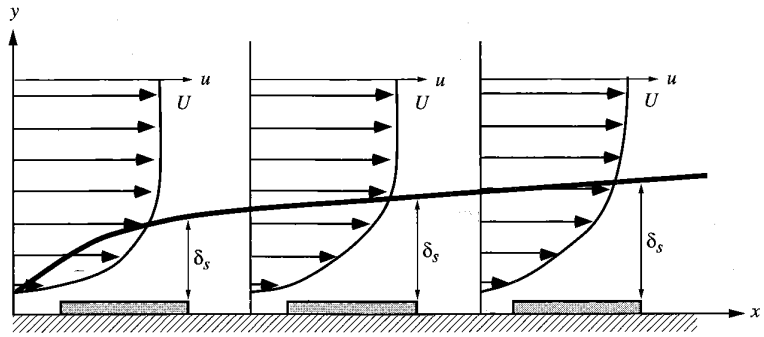


Figure 9–9 Boundary layer velocities along susceptor. δ_S is the thickness of the boundary layer. The boundary layer increases with distance in the direction of the gas flow from Newton's second law. (After [9.2].)

at all. Beyond δ_S, the gas is well mixed and moving at a constant velocity, U, parallel to the substrate surface. Figure 9–9 illustrates this and also the fact that δ_S increases in thickness along the susceptor. The transfer of gas across the boundary layer is assumed to occur via diffusion.

According to Fick's first law, the flux across the boundary layer at steady state is

$$F_1 = -D_G \frac{\partial C}{\partial x} = \frac{D_G}{\delta_S}(C_G - C_S) \tag{9.13}$$

where D_G is the diffusion coefficient of the reacting species across the boundary layer in the gas phase, and C_G and C_S are defined in Figure 9–6. From Eq. (9.2) we see that h_G corresponds to

$$h_G = \frac{D_G}{\delta_S} \tag{9.14}$$

If D_G and δ_S are constant, then h_G is constant. But δ_S is generally not constant with position in the chamber.

Right next to the wafer surface, the gas velocity is zero, and the velocity changes from zero to U within the boundary region. The boundary layer is caused by the frictional force of the gas stream on the stationary susceptor and wafers. If the gas velocity right next to the wafer surface were not zero, an infinitely large frictional force would exist right at the surface which would bring the velocity to zero there. Using Newton's second law ($F = ma$), where F is the friction force, one can determine the position of the boundary layer as a function of x to be [9.2]

$$\delta_S(x) = \left(\frac{\mu x}{\rho U}\right)^{1/2} \tag{9.15}$$

where ρ is the gas density, μ is the gas viscosity, and U is the bulk gas velocity as before. Since δ_S increases along the length of the susceptor, then the effective mass transfer coefficient, h_G, decreases along the length of the susceptor according to Eq. (9.14). Therefore, if the deposition is limited by mass transfer through the boundary layer, then the deposition rate would decrease going from the front of the susceptor to the back.

In addition, source gas depletion occurs down the length of the susceptor. As reacting gases are consumed, their concentrations decrease with distance along the susceptor. Since the growth rate is proportional to the partial pressure of the reacting species, the growth rate will decrease going downstream. While the growth rate decrease due to the increasing boundary layer is important only in the mass transfer regime, this decrease due to reactant depletion occurs in both the mass transfer and surface reaction controlled regimes. This is because Y is important in both regimes [Eqs. (9.11) and (9.12)].

To compensate for both the boundary-layer variation and depletion effects, a change in reactor geometry is needed. The wafer susceptor is tilted, as shown in Figure 9–10. This decreases the cross sectional area of the chamber along the susceptor, which causes the gas velocity to increase. This in turn causes the boundary layer to decrease along the susceptor. This increases the growth rate downstream, compensating for the friction-induced boundary-layer increase as well as the depletion effects. A way to compensate for the depletion effect when operating in the surface reaction limited regime is to impose a 5–25° temperature gradient along the tube or chamber, increasing from front to back. This is often done in LPCVD tubes, which normally operate in the surface reaction limited regime and where depletion is the main problem. For SiO_2 deposition by APCVD, commonly done in the mass transfer limited regime, a different gas flow configuration has been used to help overcome both boundary layer nonuniformities and depletion effects. The wafers still are placed side by side, facing up, but the gas is injected straight down from above. In addition the wafers can be moved horizontally on a conveyor belt. This helps ensure that uniform fluxes of reactant gases are delivered to each wafer.

As we have seen, chemical vapor deposition can be used to deposit compounds, such as SiO_2 and Si_3N_4, as well as elements. However alloys, such as Al(2% Cu), are much

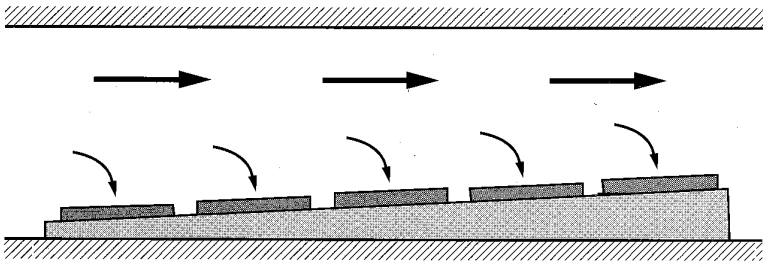


Figure 9–10 The susceptor in a horizontal epitaxial reactor is tilted so that the cross-sectional area of the chamber is decreased, increasing the gas velocity along the susceptor. This compensates for both the boundary layer and depletion effects.

more difficult. The problem is usually that separate reactions are required for each constituent (Al and Cu in our example) and it is often difficult to avoid unwanted reactions between the different reactant species, or precursors. In addition it can be hard to control the exact composition of the alloy, because unlike compounds where limited compositions result (very close to 2:1 oxygen-to-silicon in SiO_2, for example), a wide range of alloy compositions are possible. Physical vapor deposition methods are usually better suited for alloy deposition.

Thin films deposited by CVD may be doped during the deposition process. This is common for oxides, polysilicon, and epitaxial silicon. A gaseous source of the desired dopant is mixed with the other source gases, often along with hydrogen which acts as a carrier gas and a diluent. Common dopant source gases are arsine (AsH_3), phosphine (PH_3), and diborane (B_2H_6). The same steps occur in the incorporation of the dopants as occur in the thin film deposition, including mass transport and surface reactions. The surface reactions include (1) dissociation of the hydride gas, (2) lattice site incorporation, and (3) burying of dopants atoms by the other atoms in the film. It is believed that these are the rate limiting steps for doping of Si during epitaxial growth, and not the mass transport through the boundary layer. Models have been developed to predict the doping concentration in the film as a function of gas partial pressures, temperature, flow rate, and so on. However the models have not been very successful since many factors, especially in regard to the reactor geometry, are not well characterized and not adequately incorporated into the models. In general, for low growth rates the dopant concentration in the growing film C is proportional to the dopant gas-phase concentration (which equals the partial pressure of the dopant species, P_i) at the reactor input. For higher growth rates, C is inversely proportional to v, the deposition velocity.

$$C \propto P_i \text{ for low growth rates} \tag{9.16a}$$

$$C \propto \frac{P_i}{v} \text{ for high growth rates} \tag{9.16b}$$

In practice, the relationship between dopant concentration in a deposited film and the deposition conditions is determined experimentally for each particular reactor.

Unintentional doping of deposited layers can also occur. This is especially common when depositing a lightly doped epitaxial Si film on a highly doped Si substrate—a common occurrence in CMOS technology (Figure 2–20). One way unintentional doping can occur is through normal solid-state diffusion. During the deposition process, dopants from the highly doped substrate can diffuse into the newly deposited, lightly doped region. This is often termed outdiffusion. A large concentration gradient exists and at normal deposition temperatures (800–1100°C), significant diffusion can occur. Another way unintentional doping can occur is when dopant atoms get added to the gas stream and are then subsequently reintroduced into the growing film. This is termed "autodoping." The dopant may be added to the gas during film deposition (or even during an HCl in situ preclean) by evaporation from the wafer frontside, from the wafer backside or edges, other wafers, or from the susceptor.

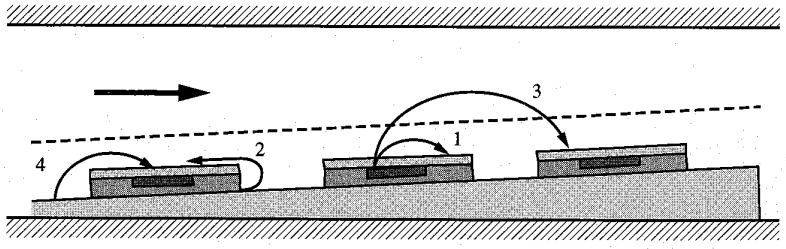


Figure 9–11 Autodoping processes in an epitaxial reactor. Illustrated are evaporation from 1, the wafer frontside; 2, the wafer backside or edges; 3, other wafers; and 4, the susceptor.

Figure 9–11 illustrates these processes. The autodoping effects add to the outdiffusion problem, often making the doping profile in the deposited layer quite complicated. The outdiffusion can be modeled with standard diffusion models. Since the growth of the layer is usually faster than the diffusion from the surface (that is, $vt \gg \sqrt{Dt}$), one can assume this is just diffusion from a doped region into a semi-infinite medium. This is described by a simple error-function profile, as discussed in Chapter 7 and given by Eq. (7.26):

$$C(x,t) = \frac{C}{2}\left[\text{erfc}\,\frac{x}{2\sqrt{Dt}}\right] \tag{9.17}$$

Here $x = 0$ at the substrate/epitaxial layer interface, and x is positive in the epitaxial layer.

Autodoping from the frontside can usually be modeled with the empirical expression

$$C_{\text{autodoping}} = C_S^* \exp\left(-\frac{x}{L}\right) \tag{9.18}$$

Here C_S^* is the effective substrate surface concentration and L is an experimentally determined parameter, or diffusion length. This decrease in dopant concentration toward the surface arises from the fact that as the film grows in thickness, the dopant must diffuse through more and more film, and less dopant enters the gas phase. Autodoping from the backside, edges, or other sources usually results in a relatively constant level in the deposited film since the source of dopant does not diminish as quickly but is at a much lower level.

Figure 9–12 illustrates these effects in a Si epilayer. The top part of the curve, closest to the substrate, is due to outdiffusion, as described to first order by Eq. (9.17). (In fact, a concentration-dependent diffusivity would model the diffusion profile better than a simple error function.). The next segment, the nearly straight line on the plot, is from frontside autodoping, described by Eq. (9.18). The background portion is due to backside autodoping, and the concentration level is constant. These effects can significantly modify desired doping profiles in epitaxial layers.

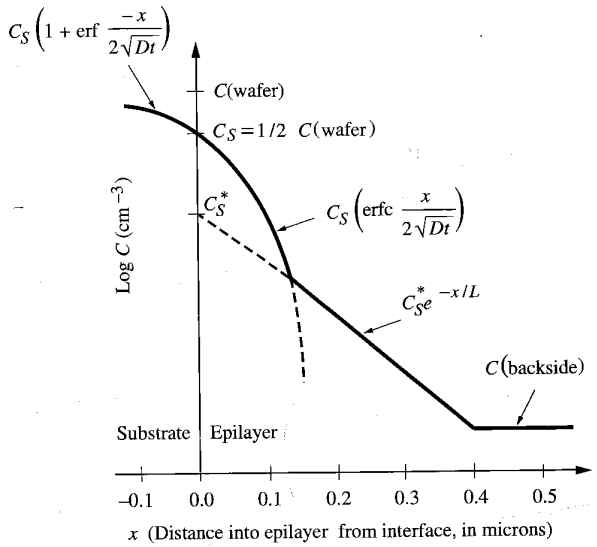


Figure 9–12 Dopant profile in a Si epilayer due to outdiffusion and autodoping. (After [9.3].)

Example Calculate the dopant concentration due to outdiffusion and autodoping in an otherwise undoped epilayer at 0.2 μm from the substrate/epilayer interface ($x = 0.2$ μm). The deposition time is 1 hour with a deposition rate of 1 μm hour^{-1}. The substrate doping, C(wafer), is 1×10^{19} cm^{-3}, C_S^* is 1×10^{18} cm^{-3}, L is 0.25 μm, D is 6×10^{-15} cm^2 sec^{-1}, and the backside doping concentration, C(backside), is 1×10^{15} cm^{-3}.

Answer The dopant concentration profile looks like Figure 9–12, with the concentration at each depth equal to the sum of the contributions from the outdiffusion, frontside autodoping and backside autodoping. For the outdiffusion term, we first check vt versus $\sqrt{Dt}$:

$$vt = (1 \times 10^{-4}\ \text{cm hr}^{-1})(1\text{hr}) = 1 \times 10^{-4}\ \text{cm}$$

$$\sqrt{Dt} = \sqrt{(6 \times 10^{-15}\ \text{cm}^2\ \text{sec}^{-1})(3600\ \text{sec})} = 4.6 \times 10^{-6}\ \text{cm}$$

Since $vt \gg \sqrt{Dt}$, we can assume the diffusion is occurring into a semi-infinite medium at all times:

$$\therefore C(x,t) = \frac{C}{2}\left[\text{erfc}\frac{x}{2\sqrt{Dt}}\right] + C_S^* \exp\left(-\frac{x}{L}\right) + C(\text{backside})$$

The outdiffusion and autodoping occur throughout the deposition, so $t = 3600$ sec. We wish to know the doping concentration at $x = 0.2$ μm, so

$$C(0.2 \text{ micron}, 3600 \text{ sec}) = \frac{1 \times 10^{19}\ \text{cm}^{-3}}{2}\left[\text{erfc}\frac{0.2 \times 10^{-4}\ \text{cm}}{2\sqrt{(6 \times 10^{-15}\ \text{cm}^2\ \text{sec}^{-1})(3600\ \text{sec})}}\right]$$

$$+ 1 \times 10^{18}\ \text{cm}^{-3} \exp\left(-\frac{0.20\ \mu\text{m}}{0.25\ \mu\text{m}}\right) + 1 \times 10^{15}\ \text{cm}^{-3}$$

$$= 0.5 \times 10^{19}\ \text{cm}^{-3}[\text{erfc}(2.15)] + 1 \times 10^{18}\ \text{cm}^{-3} \cdot 0.45 + 1 \times 10^{15}\ \text{cm}^{-3}$$
$$= 1.2 \times 10^{16} + 4.5 \times 10^{17} + 1 \times 10^{15}\ \text{cm}^{-3}$$
$$= 4.6 \times 10^{17}\ \text{cm}^{-3}$$

The major contribution at this depth is from the frontside autodoping.

9.2.1.2 Low-Pressure Chemical Vapor Deposition (LPCVD)

We saw that when operating CVD reactors in the mass transfer limited regime, one has to ensure that equal fluxes of reactant gases must reach every location in order to achieve uniform depositions across the wafers. This generally implies placing the wafers side by side as in Figure 9–4(a). Otherwise the variability in the boundary layer thickness and transport through that layer would result in nonuniform depositions. Operating in the surface reaction limited regime where transport is not important would avoid these problems and allow stacking the wafers as in Figure 9–4(b). This generally provides higher throughput. This can be achieved by going to a lower deposition temperature (Figure 9–7). However as the temperature is lowered, the deposition rate can decrease significantly. In addition the film quality may suffer.

Lowering the total pressure of the gas stream increases the diffusion and extends the reaction controlled regime to higher temperatures. First recall that the CVD film deposition rate is given by

$$v = \frac{k_S h_G}{k_S + h_G}\frac{C_T}{N}Y \tag{9.10}$$

where k_S is the surface reaction coefficient and h_G is the mass transfer coefficient. The smaller of the two terms dominates and controls the overall deposition process. Diffusion through the boundary layer is given by

$$h_G = \frac{D_G}{\delta_S} \tag{9.14}$$

where D_G is the diffusivity of the reacting species through the boundary layer and δ_S is the boundary layer thickness. The key new point is that D_G is inversely proportional to P_{total},

$$D_G \propto \frac{1}{P_{\text{total}}} \tag{9.19}$$

since the diffusivity is inversely proportional to the number of collisions that the gas species experience. So decreasing the total pressure, which also means decreasing the total concentration of gases in the gas stream but not necessarily decreasing the concentrations or partial pressures of the reactant gases, causes D_G to increase.

Decreasing P_{total} from 1 atm (760 torr) to 1 torr, a typical LPCVD pressure, increases D_G by 760 times. Decreasing P_{total} also increases the boundary layer thickness, δ_S, but not proportionally. Reducing the pressure to 1 torr only increases δ_S by 3 to 10 times. The net effect on h_G is that it increases by about 100 times. In the earlier APCVD numerical example, if h_G were increased by 100 times, with everything kept the same, the deposition rate would increase to 1.4 μm min^{-1} and system operation switches to the surface reaction controlled regime (problem 9.6).

Since h_G is much larger at the lower pressure, mass transport through the boundary layer becomes much less important compared to the surface reactions. Figure 9–13 illustrates this. Now that the deposition velocity is not limited by the mass transfer of reactants through the gas-phase boundary layer, the wafers can be placed much closer together without affecting the deposition uniformity nearly as much. The configuration

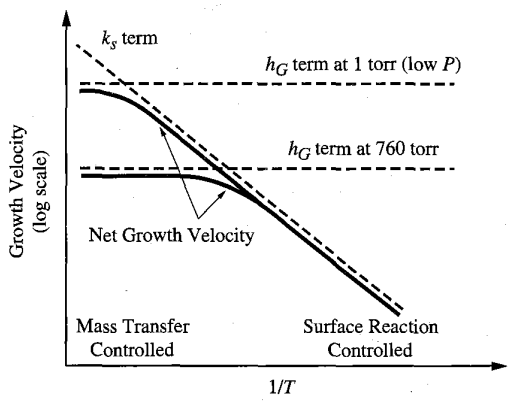


Figure 9–13 Growth velocity vs. $1/T$ for APCVD (760 torr) and LPCVD (1 torr) systems. The lower total pressure (with P_G and C_G remaining fixed) shifts the h_G curve upward, extending the surface reaction regime to higher temperatures.

shown earlier in Figure 9–4(b) is typical for LPCVD reactors. The wafers are stacked vertically. A vacuum pump is needed to reduce and control the pressure, usually in the 0.25–2.0 torr range. Deposition temperatures in LPCVD systems range from 300–900°C. Since these systems are operated in the surface reaction regime, temperature must be controlled typically to ± 1°C, but this is easily achieved in hot-walled furnaces. Depletion of reactant gas can still be a problem, as discussed earlier, so a 5–25°C temperature gradient (higher temperature downstream) is often used. This speeds up the reaction rate toward the back of the tube to compensate for the depletion. An alternate method to avoid the depletion problem is to use a distributed feed setup in which the reactant gases are injected at regular intervals down the tube.

There are additional advantages to using a low-pressure system. With the lower pressure and much faster diffusion out through the boundary layer to the main gas stream, less autodoping occurs. In addition, little or no carrier or dilutant gas is needed and gas consumption is much lower. Also, at lower pressures there are usually fewer gas-phase reactions and hence fewer particulates form that can deposit on the wafers. There is of course a limit to how much the pressure can be lowered. At some point the reactant gas concentration, C_G, or pressure, P_G, would have to be lowered to reduce the total gas pressure, which would then decrease the deposition rate.

It should be pointed out that even though we describe operating in the surface reaction limited regime at low pressures to achieve uniform deposition, the transport of the depositing material through the gas phase can still be important, especially on a local level near the surface. It is true that the diffusional flow through a boundary layer is no longer the limiting step. One might assume that under such conditions there would always be enough species available to react and deposit at each position along the surface, resulting in perfectly conformal coverage for local topographical features. However, at these low pressures virtually no gas-phase collisions occur in the near-surface region within the topographical features. Therefore line-of-sight transport occurs over this distance, as opposed to more randomly directed diffusional transport that occurs within the surface features at atmospheric pressure. As a result, shadowing can occur where the flux from the gas phase is impeded locally by surface topography features. This shadowing can directly affect step coverage and filling. Reemission processes may not always occur readily enough (or equivalently, the sticking coefficient low enough) or surface diffusion fast enough to allow for constant flux to reach each point on the surface from the gas phase. Thus nonconformal profiles can still result, as we shall see. These issues will be even more important for PVD systems—not only do they operate at very low pressures, but the deposition processes are usually assumed to occur with very little surface diffusion or reemission.

9.2.1.3 Plasma-Enhanced Chemical Vapor Deposition (PECVD)

Sometimes there are restrictions on the temperature that the substrate can be exposed to when depositing a film. The most obvious example is when depositing a silicon dioxide or silicon nitride film when an aluminum metal level is already in place. The melting point of Al is 660°C, and any subsequent processing temperature must be less than about 450°C. APCVD or LPCVD films can be deposited at these lower temperatures,

but the deposition rate is usually quite low because k_S decreases exponentially with temperature. In addition the film quality can be poor, with the films being rather porous and susceptible to moisture absorption. Good film quality is especially important for multilevel metallization schemes where deposited dielectric layers are used to provide electrical and physical barriers between metal interconnect layers. Conformal and void-free films are also important. Therefore a dielectric deposition technique is needed that provides reasonable deposition rates, good film quality, conformality, and low deposition temperatures.

A method developed for these needs is Plasma-Enhanced Chemical Vapor Deposition, or PECVD. In addition to using a thermal source to provide the energy needed for the chemical reactions to occur during the deposition, a plasma source is used. Plasmas are highly ionized gases and will be discussed in more detail later in Section 9.2.2.2. By supplying additional energy from the plasma to the reactant gases, the reactions needed for deposition can occur at temperatures much lower than those needed when only thermal energy is provided. In fact, room temperature depositions are even possible with this method. Usually, though, PECVD depositions of dielectrics are done in the 200–350°C range. In addition to allowing the depositions to occur at lower substrate temperatures, PECVD also allows more easily altering the film's properties (composition, density, stress, etc.) and the tailoring of those properties for particular applications. This is due mainly to the more nonequilibrium nature of the deposition process compared to the nonplasma CVD methods. However, this can also lead to undesirable compositions or properties in the films, such as the incorporation of byproducts or gas molecules into the film.

A typical equipment configuration for PECVD is shown in Figure 9–14. This is similar to the parallel plate plasma etching equipment to be described in Chapter 10. In plasma etching, the energetic species created in the plasma are used to etch materials on

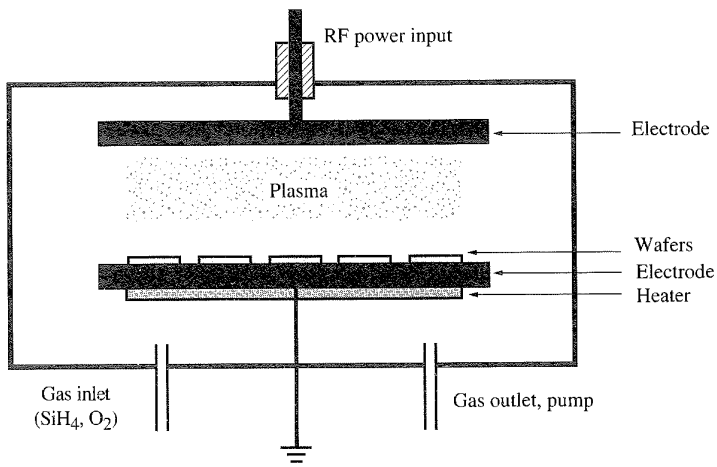


Figure 9–14 Typical PECVD equipment configuration.

the wafer. In plasma-enhanced deposition, the energetic species are used to help form the films on top of wafers. The wafers are placed on the lower plate electrode. Heat can be supplied to the wafers through a heater beneath the electrode. The reactant gases, such as silane and oxygen for the deposition of silicon dioxide, are fed through inlets. A glow discharge, or plasma, is sustained between the lower and upper electrodes. This is formed by applying a high electric field, often at 13.56 MHz, to a low pressure gas (between 50 mtorr and 5 torr), creating ions and free electrons. The plasma is sustained when high-energy electrons strike and ionize atoms and molecules.

In the plasma, interactions with the high-energy electrons cause the reactant gases to dissociate and ionize into a variety of species. These include ionized and excited molecules (or atoms), neutral molecules and neutral and ionized fragments of broken-up molecules, including free radicals. Free radicals are electrically neutral species that have incomplete bonding; that is, they have unpaired electrons. Examples are SiO, SiH_3, and F (the first two are important in plasma deposition and the last one in plasma etching). These species are extremely reactive, and with other species are absorbed onto the wafer surface, migrate, interact, rearrange, and chemically recombine to form the film. In addition, ion and electron bombardment from the plasma onto the wafer surface can occur. As we will see later in the description of RF sputtering systems, a potential difference will exist between the plasma and the wafer surface in RF plasma systems, with the wafer surface electrically negative with respect to the plasma. This will accelerate positive ions to the wafer surface. Even some of the high-energy electrons can escape the plasma to strike the wafers. This bombardment of ions and electrons transfers even more energy to the species on the surface, breaking chemical bonds for example, and further enhances the different surface processes and reactions. The net result from the plasma-induced fragmentation, free radical generation, and ion bombardment is that the surface processes and deposition can occur at much lower temperatures than in nonplasma systems. The byproducts of the deposition reactions are desorbed and exhausted through the outlet ports.

The reactions and processes that occur in a PECVD system are numerous and complicated and not easy to predict or model. The complexity and nonequilibrium nature of the reactions can lead to nonstoichiometric compositions of the films (Si- or O-rich SiO_2, for example), as well as the incorporation of byproducts into the films. Incorporation of H_2, O_2, or N_2 is common and can result in outgassing, peeling, or cracking of the film during subsequent processing. Film density and stress may also vary, depending on the conditions of the deposition.

PECVD depositions can result in fairly good coverage and filling of non-planar topography in some cases (such as TEOS SiO_2 depositions), better than what might be expected at these low temperatures and often similar to higher temperature depositions in non-plasma systems. This has often been attributed in the past to long-range surface diffusion, enhanced by the more energetic species and by ion bombardment. However, as we will discuss later, for oxide deposition this may be instead due to muliple near-surface emission and redeposition events occurring, characterized by a low effective sticking coefficient for the neutral species even at low temperatures, along with the directional ion component of the deposition to help in filling spaces [9.4, 9.5]. While standard PECVD depositions may be adequate for covering and filling larger dimension

structures, for smaller dimensions and higher aspect ratio spaces (greater than about 1) other methods are required.

9.2.1.4 High-Density Plasma Chemical Vapor Deposition (HDPCVD)

A newly developed version of PECVD utilizes a very high-density plasma and a separate RF bias applied on the substrate. Called HDP (High-Density Plasma) CVD, this technique combines PECVD deposition with bias sputtering to obtain very good filling of narrow gaps. It is used primarily for silicon dioxide depositions. The high-density plasma can be generated by a variety of sources, including Electron Cyclotron Resonance (ECR) and Inductively Coupled Plasma (ICP). This high-density plasma results in a dense CVD silicon dioxide film at low temperatures, typically 150°C down to room temperature, and with a very low chamber pressure in the 1–10 mtorr range. In fact, there is usually no intentional heating in these CVD systems. The ion bombardment supplies enough energy to raise the substrate temperature and cooling of the wafers is often required to keep the temperature below 400°C.

With a separate RF bias applied on the substrate in HDP systems, the angular dependence of ion sputtering is exploited. Sputtering—the knocking off of atoms from a solid by incident ions when the solid is electrically biased—occurs preferentially on sloped surfaces rather than on vertical or horizontal surfaces. This simultaneous sputtering of the film during its deposition can result in planarized and void-free films. This is the same idea behind the bias sputtering deposition method. Bias-sputter deposition and angle-dependent sputtering will be more fully explained in the next section. The strong directed ion component of the HDP deposition process also helps in the filling of holes, and the ion bombardment helps make the film denser.

9.2.2 Physical Vapor Deposition (PVD)

Rather than relying on chemical reactions to produce the reacting species and to form the film as in chemical vapor deposition, physical vapor deposition methods use mainly physical processes to deposit the films. PVD techniques are generally more versatile than CVD methods, allowing for the deposition of almost any material. The constituent species, individual atoms or molecules, are produced by either evaporation of a solid or molten source, or by using energetic gaseous ions in a plasma to knock off, or sputter, the atoms from a source "target." These atoms or molecules then travel though a vacuum or a very low pressure gas phase, impinge on the wafers, and condense on the surface to form the film. Very few, if any, chemical reactions occur. An exception to this is in reactive sputtering, in which a species is sputtered in the presence of a reactive gas (such as N_2) and a compound is formed and deposited. But in general, physical processes dominate in PVD.

Because of the very low pressures in these systems, very few gas-phase collisions occur. In addition, the surface reactions occur very rapidly and very little rearrangement of atoms usually occurs on the wafer surface. As a result, thickness uniformity, shadowing by surface topography, and step coverage can be very important issues in PVD depositions.

Both PVD techniques, evaporation and sputter deposition, have a long history. Both have been used for over 100 years to coat objects with metal thin films [9.6]. Today sputtering is the dominant PVD technique, but evaporation is still used for some special applications.

9.2.2.1 Evaporation

A schematic of a simple evaporator is shown in Figure 9–15. In this process the source material is heated in a vacuum chamber which has initially been pumped down to less than 10^{-5} torr. Evaporated atoms from the source condense on the surfaces of the wafers. The heater can be of the resistance type, using a tungsten filament which heats up when current passes through it. More popular for microelectronics use is an e-beam heater, in which a high-energy electron beam is focused onto the source material in a crucible using magnetic fields. e-beam heaters can achieve higher temperatures so that a wider range of materials can be evaporated. In addition, the process is cleaner since no metal filaments are used. (Recall our discussion in Chapters 1 and 4 regarding alkali ions as sources of electrical instability in early MOS devices. One source of these contaminants was found to be Al evaporation systems where sodium and potassium were used in the production of the tungsten filaments.) In addition, since only the top of the source in e-beam systems is usually melted, no contamination from the crucible occurs. As a result, purer films are deposited with e-beam evaporation systems. A downside of e-beam systems is that X-rays can be emitted when an e-beam strikes Al. These X-rays can create trapped charges in gate oxides. Annealing of the film is required to remove this damage.

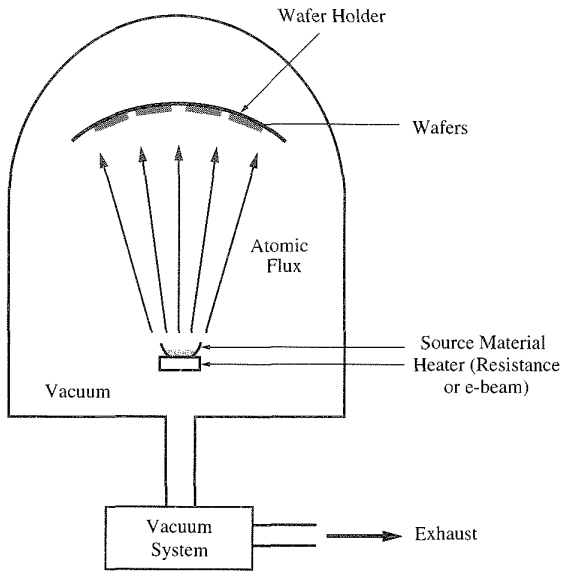


Figure 9–15 Schematic diagram of evaporation equipment.

Since the evaporated atoms are transported through a high-vacuum atmosphere, very little gas-phase scattering occurs. Therefore, unlike CVD systems, the mean free path is very long and the atoms travel in essentially straight lines all the way from the source to the surface of the wafers. This is illustrated in Figure 9–15. Since the source is small, we can first consider it as a point source. For uniform or isotropic emission from a point source, that is, an equal evaporation rate in every direction, the contours of equal flux (atoms per area per time) would be circular in two dimensions or spherical in three dimensions.

For a point source one can derive an expression for the deposition rate or growth velocity, v, of a film deposited on a flat surface as a function of distance along the surface. For this we use the coordinate system of Figure 9–16. We are interested in the amount of material striking a small area on the surface, A_k, per unit time. But first we consider the flux being emitted from the source. If R_{evap} is the evaporation rate from the point source (in either atoms or grams per second), r is the distance from the source to the spot on the surface, and A_k^p is the projected area of A_k facing the source, then the flux F_k^p that strikes A_k^p is simply

$$F_k^p = \frac{R_{evap}}{\Omega r^2} \tag{9.20}$$

Here Ω is the solid angle over which the source emits evaporated material. For a source emitting in all directions (in three dimensions) this would equal 4π. For a source emitting only upward, this would equal 2π.

Eq. (9.20) describes the flux emitting outward from a point source and it is independent of emission angle θ_i for any radius r. Since the deposition rate on A_k depends on the component of the incoming flux that is normal to the surface A_k, we must then mul-

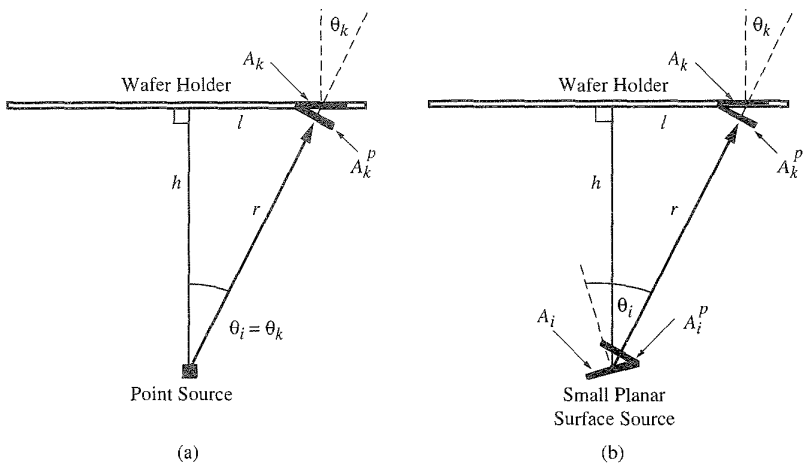


Figure 9–16 Geometries of flux and deposition of small areas on a flat wafer holder for (a), a point source, and (b), a small planar surface source.

tiply this by $\cos\theta_k$. Here θ_k is the angle between the surface normal and the direction of the source to a spot on the surface, as indicated in the figure. We designate the normal component of the incoming flux with respect to A_k to be F_k. This corresponds to the amount of material deposited per time per unit surface area. This leads to

$$F_k = \frac{R_{evap}}{\Omega r^2}\cos\theta_k \tag{9.21}$$

Another way to look at this is that for a larger angle, the material that is coming through the cone to strike the area A_k^p must be deposited over a larger surface area A_k, and the amount striking A_k per unit area is thereby decreased (by $\cos\theta$). Thus the deposition rate on a surface area will decrease as the area is tilted from an orientation facing the source.

The deposition rate or velocity, v, on area A_k is then just the flux F_k as expressed by Eq. (9.21) divided by the density of the material being deposited, N, so that for a point source:

$$v = \frac{R_{evap}}{\Omega N r^2}\cos\theta_k \tag{9.22}$$

Eq. (9.22) can be converted to a plot of the relative or normalized deposition rate versus ℓ/h, as is done in Figure 9–17, where ℓ is the distance from the center of the wafer holder and h is the length of the surface normal. Note that $\cos\theta_k = h/r$ and that $r^2 = \ell^2 + h^2$. The thickest deposit is at $\ell = 0$ and leads to v_o (v at $\ell = 0$) being equal to $R_{evap}/\Omega N h^2$. The nonuniform deposition is obvious. As ℓ increases, not only does r increase, but also the $\cos\theta_k$ term decreases.

Evaporation is usually modeled better as a small surface area source rather than a point source. Evaporation from a small surface area behaves similarly to effusion of gas

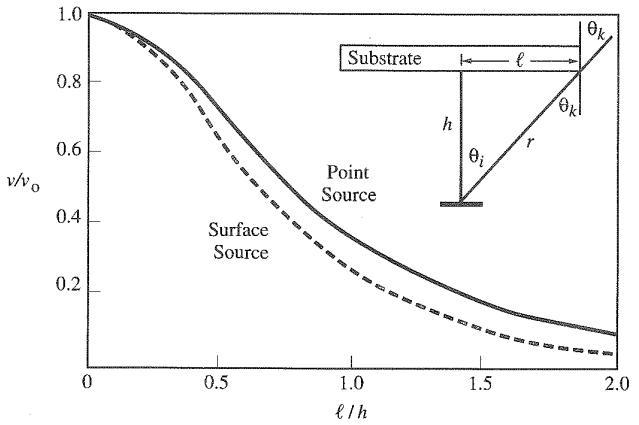


Figure 9–17 Deposition rate of evaporated film as a function of position on substrate for point and surface sources. $\theta_i = \theta_k$ in this configuration for both point and surface sources. (After [9.6].)

out a small opening from a container, as in a Knudsen cell [9.7]. The outward or emitted flux is not uniform in all directions as with an ideal point source but is more directed and dependent on the projected area of the small source area toward the flux direction. Thus the emitted flux is largest in the direction perpendicular to the source surface and less to the sides. The orientation of the 2D surface source to the area being deposited on is defined by the angle θ_i, as shown in Figure 9–16(b), which may be different from θ_k. As θ_i becomes larger, the projected area of the surface source, A_i^p, facing the deposition area, becomes smaller and there is less flux in that direction. Hence the emitted flux for any distance r is greatest in the direction normal to the source, $\theta_i = 0$, falling off as a cosine function toward either side. This is different than the $\cos\theta_k$ factor above used to calculate the deposition rate on a flat surface and instead describes how a certain type of source is seen to emit material differently in different directions, following a cosine relation. This emission behavior from a small planar source is known as the cosine distribution law for evaporation from a planar source, or ideal cosine emission.

An extra $2\cos\theta_i$ factor due to this cosine emission now comes into the expressions for the emitted flux [Eq. (9.20)], the normal component of the incoming flux striking A_K, [Eq. (9.21)], and the deposition rate [Eq. (9.22)]. (The factor of 2 normalizes the total evaporation rate with the $\cos\theta_i$ term.)The deposition rate, v, for a small surface area source becomes

$$v = \frac{R_{\text{evap}}}{\pi N r^2} \cos\theta_i \cdot \cos\theta_k \tag{9.23}$$

Eq. (9.23) is valid for any small area emitting material, with ideal cosine emission behavior, to any other small area in any relative orientation to each other, with r the distance between them and the angles as defined above.

While it may not be intuitively obvious that this Knudsen-cell-like behavior, resulting in cosine emission behavior, should apply to evaporation from a small surface, it has been verified experimentally by Knudsen himself [9.8] and others for many materials. (In radiation theory, this type of emission is called "diffuse" emission or more specifically "perfectly diffuse" emission from a surface, and very similar equations express the radiation behavior from a blackbody [9.9].) Cosine emission behavior not only describes source emission in most evaporation systems, but also describes the reemission of deposited material from a surface area in many cases, as we will discuss later. The contours of equal flux for an evaporation system with a small planar source would not be hemispherical as in the point-source case in Figure 9–15 but would fall off more quickly from the center of the wafer holder to each side.

We can now plot the deposition rate versus ℓ/h curve for a small surface source in Figure 9–17, as we did for a point source. We do this for the special case when the small surface source is parallel to the wafer holder. In this particular situation with the surface source facing the wafer holder, θ_i equals θ_k. The deposition curve for the planar surface source case is similar to the point-source case, except that the deposition is even more directed toward the region directly above it. Not only is r increasing and the $\cos\theta_k$ term decreasing as ℓ gets larger, as with the point source, but due to the $\cos\theta_i$ emission from the planar source, even more deposition occurs upward.

To achieve thickness uniformity over many wafers for either type of source, the wafers obviously cannot be placed on a flat holder. Instead, spherical or hemispherical

holders are used, with the wafers all facing inward toward the center. For a point source at the center of the sphere, θ_k and r will be constant giving a uniform deposition rate over the whole wafer holder. (Across each individual wafer, θ_k and r will vary somewhat, giving less than uniform across each wafer, but from wafer to wafer the deposition will be uniform.) It can be shown (problem 9.8) that to obtain uniform deposition around a sphere for a small surface source, the source itself should be placed on the inside surface of the sphere holding the wafers, rather than the sphere's center. Figure 9–18 illustrates these configurations. Since most evaporation sources act more like small surface sources than point sources, it is common practice to place the source at the inside surface of a spherical wafer holder in an evaporation system.

Deviations from ideal point-source or small-area surface source emissions can occur and sometimes even more directed and narrower emission distributions are observed. For small planar surface sources, the flux and deposition rate expressions can be described with a $\cos^n\theta_i$ term rather than a $\cos\theta_i$ term to take this nonideal behavior into account. Values of n greater than 1 correspond to angular distributions of evaporant flux from the source being narrower than from ideal sources. This is conceptually illustrated in Figure 9–19 for a planar source, where a value of $n = 1$ leads to ideal cosine emission

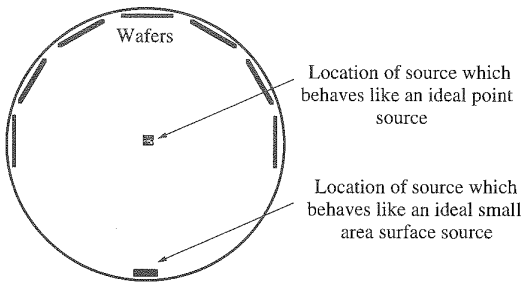
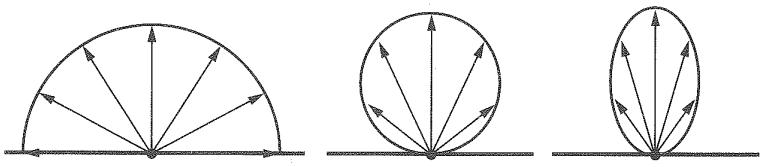


Figure 9–18 Positions of wafers and sources to achieve uniform deposition in evaporation system on all the wafers. Most evaporation sources behave more like a small area surface source and are in that configuration.



(a) Uniform (isotropic) emission from a point source

(b) Ideal cosine emission from a small planar surface source. ($n = 1$ in $\cos^n\theta$ distribution)

(c) Non-ideal, more anisotropic emission from a small planar surface source. ($n > 1$ in $\cos^n\theta$ distribution)

Figure 9–19 Emitted fluxes for point sources and small surface areas sources.

behavior in the emitted angle distribution, and a value of $n > 1$ leads to even more anisotropic behavior.

This nonideal behavior may be due to crucible geometry, the ratio of melt depth to the melt surface area, and other factors. In addition, a large density of evaporated atoms right above the source sometimes causes the gas phase to act more like a viscous medium rather than the preferred near-vacuum condition. (In the latter case, molecular flow, or free molecular flow, is said to occur, with virtually no gas-phase collisions or interactions.) This viscous behavior near the source also causes nonideal, nonuniform behavior. To counter these problems another solution is employed. Not only are hemispherical wafer holders used, but they are rotated. And not only is the entire planetary fixture rotated about the vertical axis, but each group of wafers is also rotated about a second axis. This use of rotating planetaries helps compensate for nonideal deposition behavior and results in more uniform deposition.

The source evaporation rate, R_{evap}, is found to be proportional to the equilibrium vapor pressure (P_e) of the source material at the evaporation temperature. From Langmuir-Knudsen theory, the evaporation rate in g sec^{-1} is estimated to be

$$R_{evap} = 5.83 \times 10^{-2} A_s \left(\frac{m}{T}\right)^{1/2} P_e \tag{9.24}$$

where A_s is the area of the source in cm^2, m is the gram-molecular mass, T is the temperature in Kelvin, and the vapor pressure, P_e, is in torr. The vapor pressure of the material is a very strong function of temperature, and Figure 9–20 shows the vapor pressure of several metals commonly evaporated. A partial vapor pressure of the element of about 1–10 mtorr or more is required to achieve reasonable deposition rates (on the order of 0.1–1 μm per minute), and one can see that a temperature of about

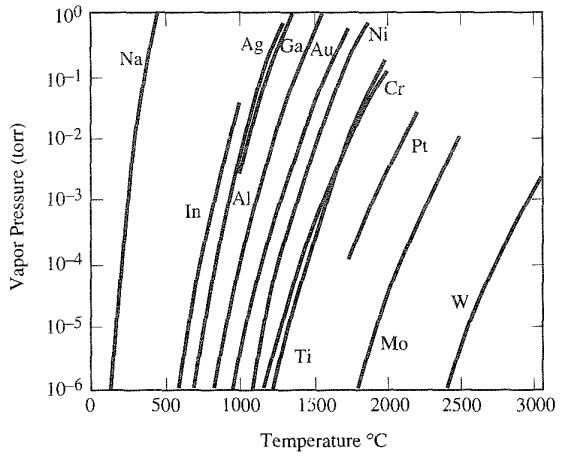


Figure 9–20 Vapor pressure as a function of temperature of commonly evaporated metals. (After [9.10].)

1100°C is needed for the evaporation of aluminum. This means that the aluminum is evaporated from the liquid phase. Tungsten must be heated to a much higher temperature of about 3100°C, where it evaporates or "sublimes" from the solid phase.

Because the vapor pressures of different elements vary so much, evaporating alloys presents a special challenge. Alloying one element with another is common in microelectronics to achieve the desired thin film properties. Alloys of Al with small amounts of Si and Cu for interconnects are examples, as we will see in Chapter 11. But with different vapor pressures at a given temperature (the vapor pressure of Cu is about 5 times lower than that of Al at 1100°C, while that of Si is 10^4 times lower!), the evaporation rates will be different, and the deposited film will have a different composition than the source. When evaporating films from a single, multicomponent source, the differences in evaporation rates of the constituent elements must be taken into account, and the sources must be enriched in the element with the lower evaporation rate. But the source itself will change in composition as one element evaporates faster than the other (called fractionation). This will in turn change the composition of the deposited film, resulting in a deposited film with a composition that varies throughout the thickness of the film. This problem is very similar to the dopant segregation problem during Czochralski crystal growth that we considered in Chapter 3.

The most common solution to this problem is two or more separate e-beams used on separate sources, each containing just one of the elements in the alloy. The e-beam energies, and hence the source temperatures, can be different so that the evaporation rates can be controlled to give the desired deposited film composition. Differences in evaporation rates and changing source compositions are not a problem here. But spatial compositional uniformity when using dual sources is a problem, and it is difficult to obtain constant compositions across the wafer and across the wafer holder.

Compounds, such as SiO_2, are just as hard, if not harder, to evaporate as alloys. The composition of the vapor phase is usually not the same as the source compound, and chemical reactions and molecular changes often occur as the material is evaporated and deposited. Sputter deposition generally does a much better job in depositing compositionally uniform films of both alloys and compounds. This is partly because the sputter rates of elements are usually much closer to each other than their vapor pressures are.

There are advantages in using evaporation to deposit thin films for microelectronics applications. First, there is little damage caused to the wafer, since the wafers are not subjected to energetic particles. Second, the deposited films are usually very pure. Because the deposition is done in a high vacuum, there are no residual gases or particles to get incorporated in the film.

However, there are also disadvantages to this method which have all but eliminated its use from mainstream silicon fabrication. First, metals with low vapor pressures, such as W, and films of alloys or compounds with precisely controlled composition are difficult to evaporate, as discussed above. Second, there is no *in situ* precleaning method available as there is for sputter deposition. The third problem with evaporation has to do with step coverage. Because of the very low chamber pressure which makes the mean free path of the gaseous species very large, the evaporated deposition species travel essentially in straight lines from the source to the wafer surfaces. This and the fact that the source is small means that the atoms arrive at a limited range of angles, leading

to poor step coverage. Rotating the planetary and wafers does help to widen the range of arrival angles, but shadowing still occurs.

Step coverage also depends on another property or characteristic of the deposition species, the sticking coefficient. This can be even more important than the line-of-sight issues in regards to step coverage and filling ability. The sticking coefficient is the ratio of the number of species that actually stay or "stick" on the surface relative to the number of incident species. We define the sticking coefficient S_C, or more accurately the effective sticking probability, specifically in terms of incident and reacting fluxes, where a deposition species that has "reacted" on the surface is one that properly deposits and stays there:

$$S_c = \frac{F_{\text{reacted}}}{F_{\text{incident}}} \tag{9.25}$$

The idea of a sticking coefficient is important for step coverage since the deposition species that do not stick or react are desorbed and may deposit elsewhere—on the sidewall of a stop, a trench, or a contact hole, for example. This is illustrated in Figure 9–21. Therefore, deposition processes with a sticking coefficient close to one, in which the species react and stick locally where they first hit without desorbing, will tend to have poorer step coverage and filling ability. Those with a sticking coefficient much less than one, in which the species bounce around before they stick and deposit, such as is generally the case in CVD depositions, will have better step coverage and better filling. In evaporation systems and other PVD methods, the sticking coefficient of the depositing species is usually very close to one, meaning that they deposit where they hit and there is little or no redepositing of desorbed species to other regions. Surface diffusion is also very slow at low temperatures. In addition, there is no resputtering of material in evaporation systems, as can be in the case in plasma CVD or sputter deposition systems where there is significant ion bombardment. As we will see in more detail later, deposition fluxes that have narrow arrival angle distributions and high sticking coefficients with little or no resputtering or diffusion lead to very poor step coverage. Deposition will occur primarily on horizontal surfaces, not on the sides of features. Heating the substrate during the evaporation process increases the surface diffusion and allows for

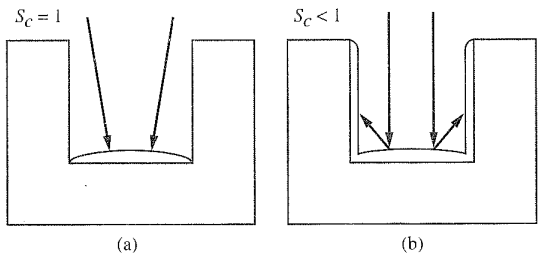


Figure 9–21 In (a) the depositing species have a high sticking coefficient (close to 1) so that they are deposited where they first strike. In (b) the depositing species have a low sticking coefficient ($<< 1$) so that many are reemitted and deposited elsewhere on the topography, such as on the sidewalls.

some coverage on sidewalls. But heating can also affect the film composition and microstructure.

For all these reasons—difficulty in depositing alloys, compounds, and high-vapor pressure elements, lack of an in situ preclean, and poor step coverage—evaporation is rarely used today for standard silicon microelectronics technology. CVD and sputter deposition, described next, are now used instead. Nevertheless, we will make extensive use of the basic concepts of emission distribution laws and deposition rates developed in this section.

9.2.2.2 Sputter Deposition

Because the vacuum requirements for sputter deposition are less severe than for evaporation (in the 1–100 mtorr range for sputtering compared to $<10^{-5}$ torr for evaporation), sputter deposition was the preferred method in a variety of technologies prior to the microelectronics age. However, because the higher pressures in sputtering systems allowed significant contamination, evaporation was generally preferred in the early years of silicon microelectronics technology. But with the availability of ultrahigh purity source gases and the use of an initial low-pressure pumpdown, gas incorporation in sputtered films was found to be greatly reduced and contamination became much less of a problem with sputtered films. Because of the problems with evaporation discussed above, as well as the inability to develop manufacturable CVD methods for some materials, sputter deposition is commonly used today in semiconductor technology.

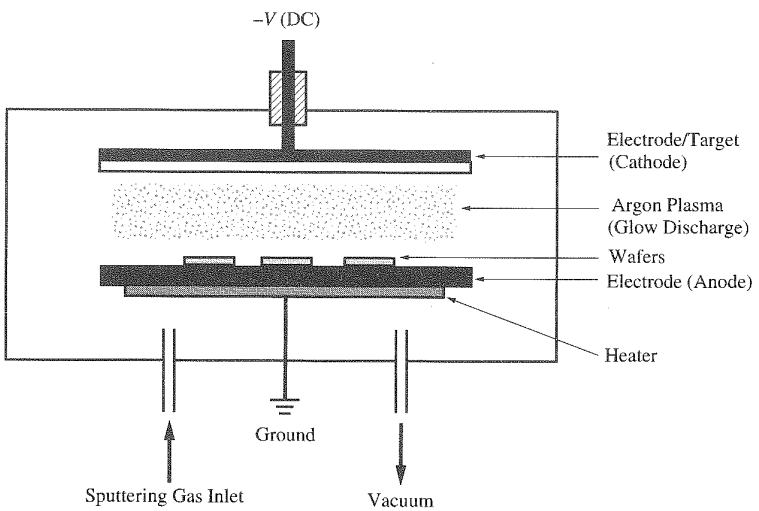


Figure 9–22 Schematic diagram of DC-powered sputter deposition equipment.

DC Sputter Deposition Figure 9–22 illustrates schematically a sputter deposition system. Note the similarity to the PECVD system in Figure 9–14. In the sputter system an inert gas, such as argon, is fed into the sputtering chamber at low pressure. A voltage is applied across the two electrodes and a plasma is created. The plasma contains neutral argon atoms, and roughly equal numbers of positive ions and free electrons, and is a conducting medium. (No free radical species are present here since Ar does not form free radicals in a plasma as O, H, and F species do in plasma deposition and etching systems.) The top electrode, where a negative DC voltage is applied, is actually the source material to be deposited, a plate or "platen" of Al, for example. This electrode becomes the cathode and is called the target. The bottom electrode is a metal platen, which is grounded and becomes the anode. The wafers sit upon the anode electrode. The positive ions in the plasma are accelerated to the negatively biased target, which can be several hundred to several thousand volts negative relative to the plasma. These energetic ions strike the target and dislodge, or sputter, the target atoms. These atoms are then free to travel through the plasma as a vapor and strike the surface of the wafers, where they condense and form the deposited film.

Since the target acts as an electrode in the DC mode of sputter deposition, the target or source material must be conductive. Therefore Al, W, Ti, silicides, and other metals can be sputtered this way. To deposit nonconductors—oxides, nitrides, and lightly doped silicon—another method must be used. We will discuss a technique to do that, the

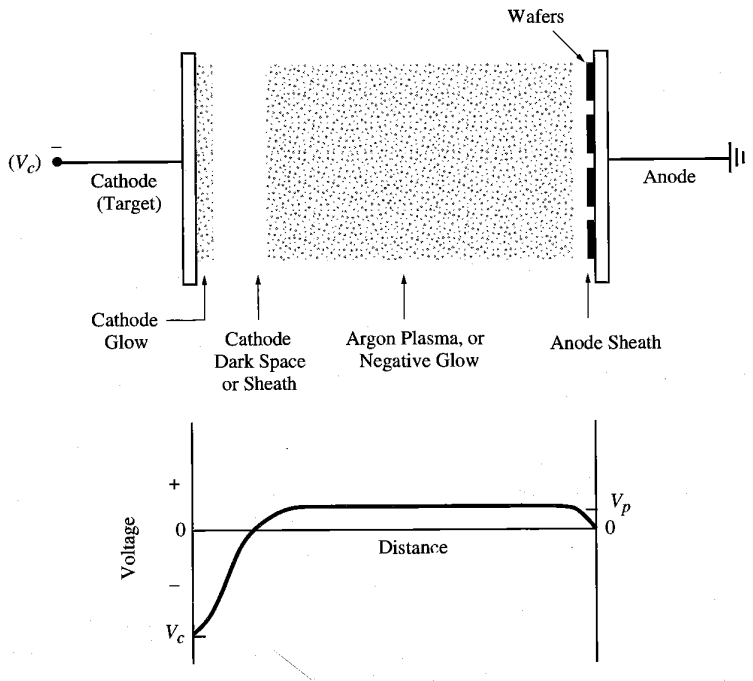


Figure 9–23 Plasma structure and voltage distribution in DC sputter system. (After [9.11].)

RF or AC method, a little later.

The structure and voltage distribution of the plasma in a DC system are shown in Figure 9–23. Most of the space between the electrodes is made up of the argon plasma, called the negative glow due to photon emission from excited Ar atoms. The voltage is fairly constant throughout this region since it has a fairly low resistance. Near the negative electrode, where there is an excess of positive ions and a shortage of electrons, as discussed below, there exists a space charge region that does not glow. This is called either the cathode dark space, the Crookes dark space, or the cathode sheath and is on the order of 0.1 to 10 mm thick. It is across this space charge region that most of the voltage drop in the system occurs. Because of the low density of electrons in the cathode sheath, few collisions between Ar atoms and electrons occur and that region is dark. Right next to the negative electrode is the thin cathode glow region; the light produced in this case is due to the interaction between the incoming ions and the cathode material. In this region incoming argon ions and sputtered target ions are neutralized and secondary electrons created by the ion bombardment are accelerated away from the cathode. These electrons then pass through the cathode sheath and collide with the Ar atoms in the negative glow region, ionizing some of them and sustaining the plasma.

While the electrons are accelerated in one direction across the high electric field in the cathode dark space, the positive argon ions are accelerated in the opposite direction, from the negative glow region to the target. They strike the target surface and sputter off the source atoms, which can then transport to the wafers and form the film.

Additional glow regions and dark spaces can form closer to the anode, such as the Faraday dark space, the positive column region, and the anode sheath [9.11]. But these are not useful in the sputter deposition process, and all but the anode sheath are eliminated or minimized by having the positive electrode close to the negative glow, resulting in the plasma structure and voltage distribution shown in Figure 9–23.

One interesting feature is that the plasma potential, V_p, is positive with respect to the anode (on the order of 10V), as shown in Figure 9–23. In fact, the plasma potential will be positive with respect to any surface within the plasma, with positive space charge regions, or sheaths, formed next to each surface. This is primarily due to the fact that both ions and electrons from the plasma will randomly strike the surface and escape the plasma. But the electrons with their lighter mass and higher mobility will strike the surface more often, causing an imbalance in the ion and electron currents. An electric field develops that opposes this imbalance by opposing the electron flow, raising the voltage (by V_p). Therefore the potential of the plasma will be positive with respect to both the anode and cathode, with a positive space charge region of excess ions and a shortage of electrons forming a sheath at both the anode and the cathode. With the applied negative voltage at the cathode, the voltage distribution shown in Figure 9–23 results, with a net DC current passing through the circuit. In the region near the cathode, this current is predominantly carried by the positive ion flux (which is also responsible for the sputtering). The large voltage drop there accelerates the secondary electrons from the cathode across the sheath and sustains the plasma. In the region near the anode an equal current flows, but is mostly carried by the electrons. Because of the exponential Maxwell-Boltzmann distribution of the electron energy, only a small perturbation of the zero-current plasma potential is needed to supply the net DC current at the anode. Because of the potential drop from the plasma to the anode, some sputtering of the anode

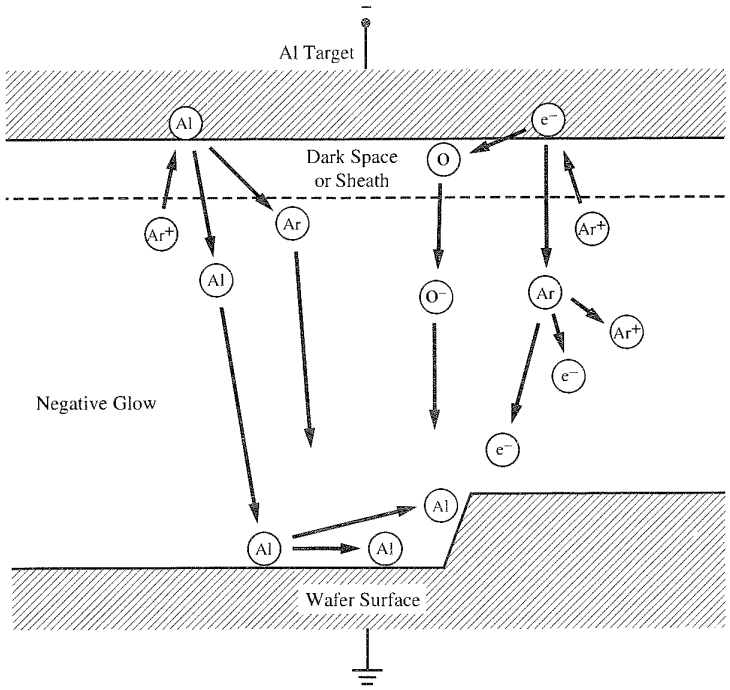


Figure 9–24 Important processes in sputter deposition.

can occur, but very little compared to that at the cathode.

Several of the important processes that occur during sputter deposition are illustrated in Figure 9–24. At the top left, an Ar ion from the negative glow accelerates across the dark space and sputters off an Al atom from the target. (A minimum energy, on the order of 10–20 eV, is needed to sputter an atom, and the sputter yield increases with ion energy.) This Al atom travels through the dark space and negative glow and lands on the wafer surface. If the argon atom is neutralized during gas phase collisions, it can retain enough kinetic energy to also travel to the wafer surface and become incorporated in the growing film. The Al atom now on the wafer surface may stay absorbed and become part of the growing film there. It may also surface migrate or diffuse to another location. Finally, it can leave the surface by either desorbing on its own (though this does not occur very often compared to CVD systems, as we will discuss later), or by being resputtered by another atom hitting the surface. The Al atom could then be redeposited at another location.

In the upper right, an Ar ion strikes the target, and a secondary electron is emitted. The secondary electron can accelerate across the dark space and strike another argon atom, exciting it or ionizing it, and sustaining the plasma. The electron can also ionize an impurity atom, oxygen in this case, which can then be accelerated across the dark space and travel toward the wafer surface. Finally, all the species moving toward the

wafer surface can be embedded in the wafer surface. Even a significant number of positively charged ions can be made to strike the wafer surface with high energies by independently biasing the wafer surface relative to the plasma, as will be described later in the sections on bias sputtering and ionized sputtering. Or the atoms or ions can strike the atoms on the wafer surface, imparting energy to them and enhancing their movements and reactions, and even heating the wafer surface. Thus physical reactions and processes, rather then chemical, dominate in PVD.

In evaporation, the deposition rate was seen to depend on the source material vapor pressure. In sputter deposition, it is the source, or target, sputtering rate that is important. The plasma gas ion strikes the target and physically dislodges the target atom. This is a physical, rather than a chemical, process, and is often explained using billiard ball models. Sputtering occurs when the incoming ion transfers enough energy to the target surface to break the bonds holding the target atom in place. This depends on the energy and mass of the ions, and the target material.

The sputter yields for argon sputtering of different target materials do not vary very much for the different target materials compared to the wide range of vapor pressures important in evaporation. Typically this yield, defined as the number of atoms or molecules ejected from the target per incident ion, is in the range of 0.1 to 3. It does depend on the energy of the incoming ion, determined mainly by the voltage across the sheath, and on the direction of incidence of the ions. An ion coming in at a low, glancing angle will not be as successful at dislodging an atom, with the ion itself being reflected off the surface. At the other extreme, ions arriving normal to the surface will push a surface atom further into the target but not necessarily sputter it off. The sputter yield peaks at some angle less than 90°. The angles that the sputtered atoms leave the surface of the targets are usually found to be diffusely distributed, with an ideal cosine-emitted angle distribution, or close to it, from each position on the target. This means that the sputtered atoms are ejected from each location on the target in a broad range of directions.

The targets used in sputter deposition come in a variety of shapes and sizes. Multiple targets may also be used in these systems for cosputtering of alloys and compounds or for multilayer depositions, where Ti and Al layers may be deposited one after the other in the chamber without breaking vacuum. Achieving good thickness uniformity in the different types of systems is accomplished in different ways: having the target larger than the wafer or wafer holder; having the wafers pass by or spin in front of a long rectangular target; or by using the cylindrical wafer holder around an oblong target (similar to having a spherical holder in front of a point source in evaporation.)

Having a large area target also means that the arrival angles of the atoms at the wafer surface are more widely distributed which generally improves step coverage. Having the wafers rotate or pass by a target is somewhat similar to having a large target in this regard and also helps increase the range of arrival angles in the direction of the rotation.

Let us first discuss how we describe or quantify the "arrival angle distribution." Here we are interested in the flux, as a function of arrival angle, arriving at any small area on the wafer surface. In Figure 9–25 on the left we show fluxes arriving at the surface from all angles with equal magnitude; that is, the number of particles or atoms, per time and per area normal to the flux direction, is equal for all incoming directions. We call this isotropic arrival. We could describe this arrival distribution as being constant since the flux relative to the flux direction is constant, similar to the way we characterized the

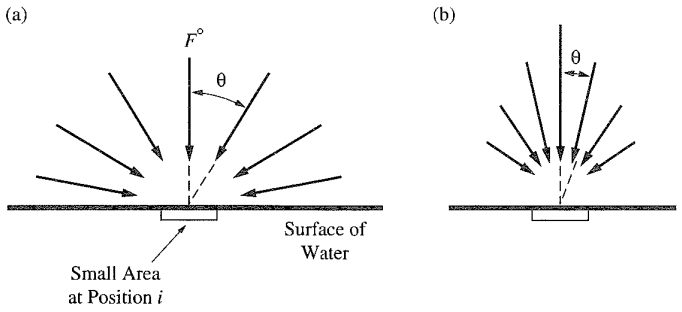


Figure 9–25 Distribution of arrival fluxes for (a) uniform or isotropic arrival distribution; and (b) directed or anisotropic arrival distribution. Arrival angle distribution ($\cos^n\theta$) is defined by arrival flux relative to unit surface area. This flux is equal to the normal component of incoming flux, relative to the vertical direction for a horizontal surface.

emission angle distribution. However, we are going to define the arrival angle distribution a little differently. We are ultimately interested in the deposition on a surface area, so we are concerned with the flux in terms of particles per time per area on the surface, or per unit surface area. This is just the incoming flux, which is constant for isotropic arrival, multiplied by the cosine of the arrival angle θ, where θ is the angle between the incoming flux and the normal direction to the surface. Just as in our treatment of a flux striking and depositing on a flat surface in an evaporation system, the deposition rate depends on the normal component, or the cosine of the angle, of the incoming flux to that surface. Using the cosine function in the arrival angle distribution also allows us to easily describe both isotropic and anisotropic arrival distributions. (This convention may be a little confusing since in our description of the emitted angle distribution for evaporation sources and emission, which uses a different convention, a "cosine distribution" from a small area is not uniform or isotropic as we describe it but is directed toward the surface normal. But to be consistent with normal terminology for emission, as well as with that used to describe arrival angle distributions in our deposition and etch modeling using the SPEEDIE simulator, we will use these two conventions.)

We thus describe the arrival angle distribution in terms of $\cos\theta$, or the normal component of the incoming flux. For isotropic incoming fluxes, the arrival angle distribution by this definition is therefore described as $\cos\theta$, with a maximum at $\theta = 0^\circ$ and decreasing as the angle gets larger, eventually going to 0 at $\theta = 90^\circ$. Another way to look at it is that for isotropic arrival, the atoms that strike the unit surface area from a particular direction, θ, must pass though the projected area of the unit surface area facing the flux direction. As θ increases, the projected area gets smaller. Therefore fewer atoms strike the unit surface area, and the flux relative to the surface area (the atoms per time per unit surface area) decreases. Ultimately, no atoms strike the surface area when $\theta = 90^\circ$, since the flux is parallel to the surface.

For incoming fluxes that are not isotropic, but anisotropic and more directed toward the surface [as shown in part (b) of the figure], a $\cos^n\theta$ distribution can be used. A value of n larger than 1 describes a more directed or anisotropic incoming flux distribution.

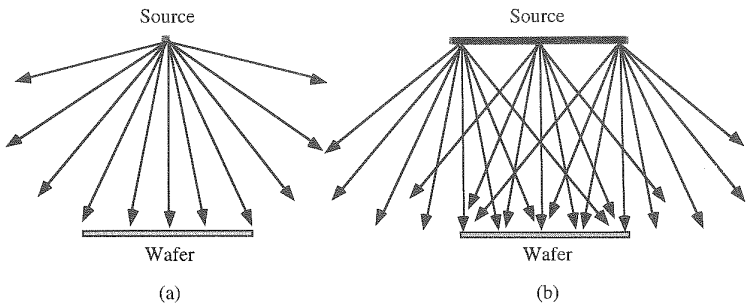


Figure 9–26 Flux lines of atoms from small area and extended sources. In (a) the flux lines from a point or small area source are shown, with a wide emitted angle distribution but a narrow arrival angle distribution. In (b) the flux lines from an extended source are shown. A wider arrival angle distribution is seen with the extended source.

We now return to illustrating the idea that a larger source region, or target area in the case of sputtering, results in a wider arrival angle distribution. In Figure 9–26(a), the flux lines of atoms from either an ideal small area surface source or an ideal point source are shown, in either case emitting in a rather wide angular distribution. While the atoms are emitted from the small source with a wide angular distribution, the distribution of angles arriving at each point or small area on the wafer is very small. This tight or narrow distribution can be described by an arrival distribution of $\cos^n \theta$ with n much greater than 1. Obviously a wide emission or departure distribution of angles does not necessarily mean a wide arrival distribution of angles.

In Figure 9–26(b) an extended source is used, such as a larger target. At each position on the wafer, atoms arrive at a range of angles, as determined by the geometry of the system, and not just from a narrow range of angles. Here n in the $\cos^n\theta$ arrival distribution would be closer to 1, meaning less anisotropic and more isotropic. This can cause more uniform step coverage on steps in the wafer topography, and would also improve thickness uniformity across the substrate.

As we described earlier, the arrival angle distribution for evaporation systems can be made wider by rotating the wafer holder around two axes, effectively making the source wider, and the step coverage can be improved. But there are other reasons that step coverage is usually still better in sputtering systems. Because of the higher pressure in sputtering systems, the atoms experience more gas-phase collisions in sputtering systems. One can estimate the average number of collisions by calculating the mean free path, λ, of a particle in a gas. From kinetic theory, the mean free path of a gas particle is

$$\lambda = \frac{kT}{\sqrt{2}\pi d^2 P} \tag{9.26}$$

where $k = 1.36 \times 10^{-22}$ cm^3 atm K^{-1}, T is the temperature in K, d is the collision diameter of the molecule in cm (approximately 4×10^{-8} cm, or about 0.4 nm, for most molecules of interest), and P is the pressure in atm. In a typical sputtering system with $P =$

5 mtorr, this results in a mean free path on the order of 1 cm, which is a little less than the physical path length the depositing atoms follow.

For evaporation systems, the number of collisions experienced between the source and wafer is virtually zero, while in sputtering systems the average number of collisions experienced is on the order of one to a few. This tends to make the arrival angle distribution somewhat wider. (Increasing the pressure or number of collisions too much to widen the distribution can result in poor film properties, partly due to less energetic species, and contamination problems.) Also, the number of resputtered atoms on the wafer surface is higher for sputtering due to the bombardment of energetic particles. This effectively increases the arrival angle distribution even more. For these reasons, step coverage in sputtering systems is usually better than in evaporation systems. But it is still usually much worse than in CVD depositions, which generally have an isotropic arrival angle distribution and lower sticking coefficients, as we will discuss later. In some cases, though, we actually want to narrow the arrival distribution—for example when filling a deep via with metal in a PVD system. We will see how that can be achieved when we discuss collimated and ionized sputtering systems.

Because the sputtering yields of different elements are much more similar to each other than their vapor pressures are, the deposition of alloys and compounds should be much more controllable using sputtering rather than evaporation. But there is another reason why the compositions of sputter deposited films are more controllable and uniform. Because sputtering occurs just from the top surface of the target and little mixing from the layers below takes place, a steady-state condition soon develops where the atoms are sputtered off at relative rates equal to the target composition, even when the sputtering rate of each type of atom is different. Initially the element with the higher sputtering rate does preferentially sputter off more from the surface. This leaves the top surface enriched, relative to the bulk of the target, with the element with the lower sputtering rate. But now more of the element with the lower sputtering rate is sputtered since the surface region is enriched with that element. A steady state is achieved where the elements leave the surface in a ratio that is the same as the bulk target composition—the difference in sputtering rates is exactly compensated for by the difference in the surface concentration.

The overall target composition thus remains at its original composition, with no depletion or fractionation, since the two elements are being sputtered off in proportion to the overall composition. This does not work for evaporation. Since the evaporation source material is often molten or near molten, the diffusion of atoms is very fast and mixing between the surface and bulk wipes out any concentration gradients. Therefore, the required enriched surface concentration cannot be maintained. But in sputtering, the temperature of the target is much lower—near room temperature—and little diffusion and mixing occur.

Compounds, such as SiO_2 and WSi_2, may be sputtered as well. In some cases, they are sputtered off as individual atoms, and sometimes they are sputtered as molecules. Both alloys and compounds may be sputtered from two different targets, called cosputtering (as opposed to the use of one "composite" target). For example, WSi_2 can be sputtered by simultaneously using a tungsten target and a silicon target. But it is often difficult to

control the stoichiometry, and composite targets are more commonly used. Even then the composition may vary.

Reactive Sputter Deposition Another way in which compounds can be sputter deposited is by reactive sputter deposition. Here a reactive gas is introduced into the sputtering chamber in addition to the Ar plasma, and the compound is formed by the elements of that gas combining with the sputtered material. For example, TiN can be deposited by sputtering Ti in the presence of nitrogen. The plasma can furnish energy to the N_2 to allow it to dissociate into atomic nitrogen, which then can easily react with the Ti. The reaction usually occurs either on the wafer surface or on the target itself, and not in the plasma. In the case of TiO_2 reactive sputtering, it is found that the oxygen reacts with the Ti on the surface of the Ti target to form the oxide there, and then it is sputtered onto the wafer surface. Controlling the stoichiometry of reactively sputtered films is sometimes difficult and this method is not used very often—sputtering from compound targets is more common. Sometimes a reactive gas such as O_2 or N_2 is used in conjunction with sputtering from a compound target. This could ensure that the composition of a gaseous constituent is maintained in a sputtered film, or it could be used to improve the properties of a film. Sputtering TiW in the presence of oxygen or nitrogen has been found to improve the barrier properties of these films, as discussed in Chapter 11.

RF Sputter Deposition DC sputter deposition is not suitable for insulator deposition. When applying the necessary DC voltage to the insulating target to sustain a plasma, difficulties arise. With a negative DC voltage applied, positive argon ions from the plasma strike the negatively charged insulator and positive charge would accumulate. The negative surface voltage would become less than that required to sustain the glow discharge, and the plasma would shut down. The time required for this to occur is only about 1–10 μsec.

The solution is to use a high-frequency alternating voltage instead of DC. 13.56 MHz is commonly used. RF voltages can be coupled capacitively through the insulating target to the plasma, so that conducting electrodes are not necessary, and the positive charge buildup is neutralized by electron bombardment over each cycle. The RF frequency is chosen to be high enough so that a continuous plasma discharge is maintained, that is, higher than the effective RC time constant of the system. The difference in mass and hence mobility between the electrons and ions allows for virtually continuous sputtering of the target throughout both half cycles of the RF voltage. The electrons have a high enough mobility to keep up with the changing electric field, but the heavier argon ions do not. During the first few complete cycles more electrons than ions are collected at each electrode, and this negative charge buildup on the electrodes is maintained during subsequent cycles. As a result, both electrodes maintain a self-biasing steady-state (DC) potential that is negative with respect to the plasma voltage, V_p. The value of V_p is set by the requirement that the DC current is zero. A positive V_p aids the transport of the slower positive ions and slows down the negative electrons. The spatial distribution of the voltage that develops is shown in Figure 9–27, with the solid curve for the symmetric RF case (that is, with equal area electrodes). This is similar to

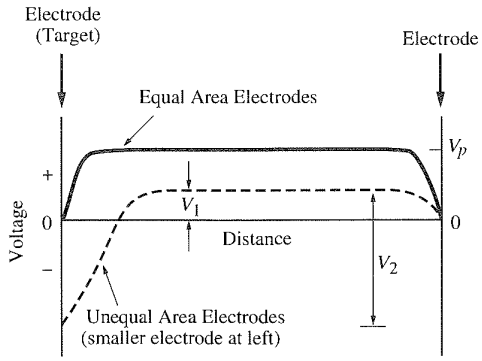


Figure 9–27 Steady-state voltage distribution in RF powered sputtering systems. The wafer sits on the right electrode.

the DC-powered case where electron depletion regions form and the plasma self-biases. However, V_p is generally larger and the net voltage drop at both electrodes would be equal in the symmetric RF powered case. The heavy ions, while they cannot travel across the sheath at the frequency of the RF signal, do respond to the average DC potential that develops across the sheath and strike the negatively biased target.

The induced negative biasing of the target due to the RF powering means that continuous sputtering of the target occurs throughout the RF cycle. But it also means that this occurs at both electrodes. The wafers will be sputtered at the same rate as the target since the voltage drops would be the same at both electrodes for a symmetric system. It would thus be very difficult to deposit any material this way. However, because the RF current must be continuous throughout the system, the voltage drops, V_1 and V_2, across the dark space next to the two electrodes are related to the respective electrode areas, A_1 and A_2, by [9.12]

$$\frac{V_1}{V_2} = \left(\frac{A_2}{A_1}\right)^m \tag{9.27}$$

The smaller electrode requires a higher RF current density to maintain the same total current as the larger electrode. As a result, the fields must be higher at the smaller electrode. Simple theory gives $m = 4$, but experimentally the exponent is found to be between 1 and 2, perhaps due to voltage saturation and sheath collision issues, but the effect is the same. By making the area of the target electrode smaller than the area of the other electrode, the voltage drop at the target electrode will be much greater than at the other electrode. V_p in the asymmetric case would also be smaller. The resulting voltage distribution for unequal electrodes is also shown in Figure 9–27, and closely resembles that for the DC sputter deposition case (Figure 9–23 shown earlier). The volt-

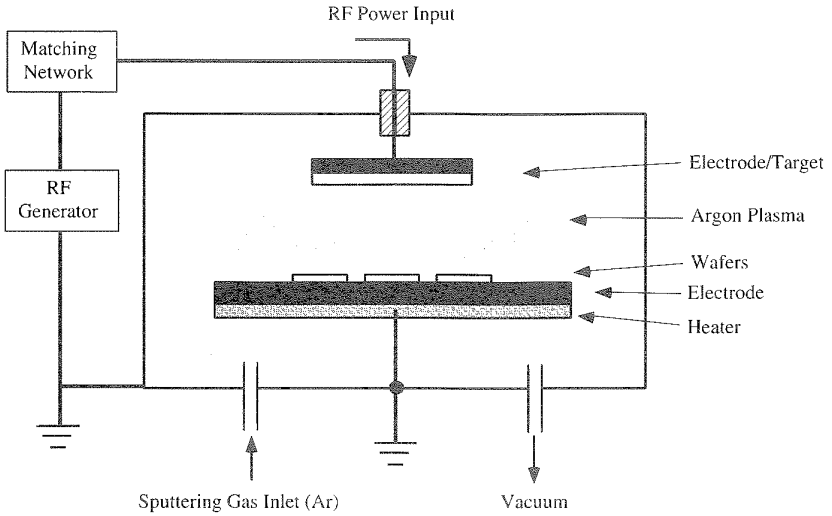


Figure 9–28 Schematic diagram of RF-powered sputter deposition system.

age drop from the plasma to the larger electrode is typically tens of volts while the voltage drop to the smaller electrode is several hundreds or even a thousand volts. Therefore almost all the sputtering will occur at the target electrode, as is desired.

Not only can the target electrode be made relatively small, but the wafer electrode can also be connected to the chamber walls (which would also be grounded), thus making the effective area of that electrode even larger. For typical RF-frequency power supplies (e.g., 13.56 MHz), an impedance matching network is also required. A typical setup for RF sputter deposition is shown in Figure 9–28.

In addition to allowing sputtering of nonconducting material, there are two other advantages of using RF powered systems. In RF plasmas, electrons gain energy directly from the oscillating RF fields because of low-loss elastic collisions with neutral atoms. Thus the plasma can in fact be sustained without the need for secondary electrons from the walls and electrodes. In addition, the efficiency of ionization is increased by these oscillating electrons, allowing for operation at somewhat lower pressures.

Bias Sputtering: Sputter Etching and Bias-Sputter Deposition Sometimes sputtering of the wafer is desirable. One application would be for precleaning the wafer before the actual deposition, or "sputter etching." Another application is in "bias sputtering," where deposition and sputtering of the wafer are done simultaneously. A negative bias relative to the plasma is applied to the wafer electrode, which is now electrically isolated from the chamber walls. Positive Ar ions from the plasma will now be accelerated to the wafers on the substrate carrier and sputter off atoms. The energy of the ions (and hence the sputtering yield) can be controlled by controlling the substrate bias. Usually an RF bias is used since the wafers often have insulating films on them, while the target is given either a DC or RF bias. The target bias is in the range of -700 to -2000 V, while the substrate bias is on the order of -50 to -300 V [9.6].

In sputter etching or cleaning, no deposition is allowed to occur on the wafer by using a shutter to block sputtered material from the target. During this step, a controlled thickness of surface material is sputtered off the wafer, removing any contaminants or native oxides. A film can then be sputter deposited immediately afterward without breaking the vacuum, inhibiting recontamination or reoxidation. Redeposition of contaminants due to sputtering of chamber walls and other parts of the wafer can occur during sputter etching and may be a concern. Charging damage can also be a problem.

In bias-sputter deposition, or sometimes just "bias sputtering," sputter etching of the wafer and deposition on the wafer (by sputtering the target) are allowed to occur concurrently. Conditions are chosen so that more deposition occurs than etching. But by allowing some sputtering of the wafer surface to occur during the deposition, both the topography and the properties of the deposited film can be altered. The most common application is SiO_2, called "bias-sputtered quartz" when deposited this way. By applying bias to the substrate, sharp edges of the SiO_2 due to the underlying structure are reduced during the sputter deposition. This allows for a more planarized film as well as for better filling between metal lines. The common reason given for these results is that the sputtering of the wafer is a function of the geometry of the surface features, with pointed and sloped features being sputtered, and removed, more easily than horizontal or vertical surfaces. Not only does this help planarize the film, but the overhang that can develop when depositing into a trench or hole is preferentially sputtered away, allowing for better filling of the hole, as illustrated in Figure 9–29. Another factor may be that the resputtered atoms allow for some redeposition on sidewalls. We will demonstrate these ideas with simulations of high-density plasma CVD and ionized PVD depositions which make use of these same phenomena later in this chapter. Another technique for filling spaces based on these same ideas is to do the deposition, usually by PECVD, and sputter etching sequentially, as opposed to simultaneously. This can be done in a cluster tool with separate chambers for each process. The wafer is physically moved back and forth between the two.

Because of the energetic ions bombarding the film surface during bias-sputter deposition, the properties of the films being deposited can also be altered by changing the substrate bias. These properties include resistivity, residual stress, density, etch rate, and dielectric constant. SiO_2 that is deposited by regular sputtering (no substrate bias) is very porous and unsuitable for use in microelectronics, while SiO_2 that is deposited using bias sputtering is much denser and has been used for intermetal dielectric applications.

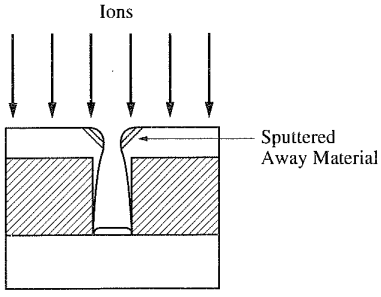


Figure 9–29 Illustration of angle-dependent sputtering which removes nonplanar features in bias-sputter deposition. Here the angled surfaces of the overhang are preferentially sputtered by the directed ions, allowing for better filling of the hole.

However, this method for depositing SiO_2 also has shortcomings. First, because of the combined deposition and etching of the wafer, the net deposition rate is low. Also, the SiO_2 will deposit on the chamber walls and other metal surfaces in the chamber and can flake off and cause contamination and particulate problems. For these reasons bias sputtering is not commonly used today. A method which combines bias sputtering with CVD deposition methods, the High-Density Plasma (HDP) CVD method described earlier, has recently been developed and does not suffer from these problems.

Magnetron Sputter Deposition In both conventional DC and RF sputtering, the efficiency of ionization from energetic collisions between the electrons and gas atoms is rather low. Most electrons lose their energy in non-ionizing collisions or are collected by the anode, and only a small percentage of them take part in ionization processes with the Ar atoms. In RF plasmas electrons gain energy directly from the oscillating RF fields, increasing the ionization efficiency somewhat. But the overall ionization efficiency is still fairly low. As a result, the deposition rates are usually rather low. Few Ar atoms are converted to positive ions and few positive ions strike the target.

In the technique known as magnetron sputtering, magnets are used to increase the percentage of electrons that take part in ionization events, and the ionization efficiency is increased significantly. A magnetic field is applied at right angles to the electric field, usually by placing large rectangular magnets behind the target. This traps the electrons near the target surface, and causes them to move in a spiral motion until they collide with an Ar atom. The ionization and sputtering efficiencies are increased significantly, and deposition rates of up to one μm per minute can be achieved—about 10 to 100 times faster than without the magnetron configuration. Unintentional wafer heating is also significantly reduced, since the dense plasma is confined to near the target and the ion loss to the wafers is less. In addition, a lower Ar pressure can be utilized since the ionization efficiency is so much larger. This can lead to better film quality since less argon will be incorporated in the film. Magnetron sputtering can be done in either DC or RF modes, but the former is more common. At these higher ion currents, cooling of the target can be a problem for RF systems which have insulating targets.

Collimated Sputter Deposition and Ionized Sputter Deposition A small range of arrival angles during deposition can cause nonuniform, or poor, step coverage over a step in topography. However, sometimes a small range of arrival angles is desirable. If material is required to be deposited into of a deep contact or via, as when filling it with Aℓ or depositing Ti/TiN contact and barrier layers, a large arrival angle distribution can cause problems. As illustrated in Figure 9–30(a), a relatively large arrival angle distribution, when little surface diffusion or reemission occurs, can result in little deposition at the bottom of a hole due to shadowing effects. In addition, overhang formation can occur at the top corners of a deep hole, enhancing the shadowing effect. All this leads to poor coverage at the bottom of the hole or incomplete filling.

One way to improve this is by having a narrow range of arrival angles, with most of the depositing atoms arriving at the wafer perpendicular to the surface. Evaporation from a stationary small source would result in a narrow range of arrival angles compared to sputtering due to geometric arguments (Figure 9–26), as well as due to the fewer gas-phase collisions, but sputtering is preferred due to its ability to deposit uniform alloys and compounds and its in situ sputter etch capability. While standard sput-

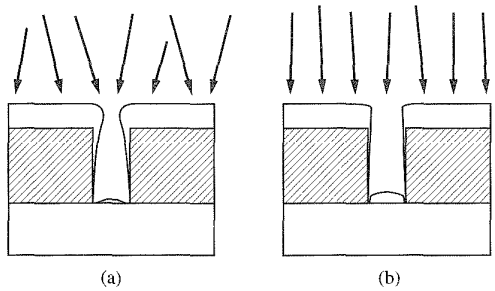


Figure 9–30 Effect of arrival angle distribution of depositing species on filling of trenches or holes in a PVD system. In (a) a relatively wide arrival angle distribution leads to poor bottom filling or coverage, while in (b) a narrower arrival angle distribution leads to better bottom filling. The higher the aspect ratio of the feature, the narrower the arrival angle distribution must be for adequate bottom coverage.

ter deposition equipment may adequately cover the bottoms of holes with low aspect ratios (less than 1 or so), its arrival angle distribution is too large for holes with high aspect ratios. Using a smaller sputtering target would result in a narrower arrival angle distribution at the wafer, with a single angle coming from a point source in the extreme case. But that angle would vary across the wafer. This would result in uneven asymmetric deposition, depositing more on one sidewall of the hole or the other for much of the wafer. (A small evaporation source would have the same problem.) It would be better to somehow filter or focus the incoming atoms in a sputtering system so that a narrow range of angles, all nearly perpendicular to the wafer surface, arrived at the wafer. Better filling of a hole, and more uniform directionality across a wafer, could thus be obtained as illustrated in Figure 9–30(b). Two methods to do this have been developed: collimated sputtering and ionized sputtering.

In collimated sputtering, a plate of circular or hexagonal holes, or cells, is placed between the target and the wafer. The height to width aspect ratio of these cells can range from 1:1 to 3:1. As the sputtered atoms travel through the collimator toward the wafer, only those with a nearly normal incidence trajectory will continue on to strike the wafer. Those at more oblique angles will strike the walls of the collimator, depositing there. The collimator thus acts as a physical filter to low-angle sputtered atoms. A problem with this method is that a significant percentage of the sputtered atoms are filtered out, up to 95 percent, and never make it to the wafers. Thus the deposition rate is significantly reduced and the collimator must be cleaned regularly. As a result, collimated sputtering is rarely used for completely filling contacts or vias where a lot of material needs to be deposited but is more suited for contact and barrier metal deposition where the amounts of material deposited are much less. It does do a fairly good job of covering the bottoms of contacts and vias with contact and barrier metals. Some sidewall coverage, important for via liners, is also obtained from the flux that is slightly off-normal, but the sidewall film is often less dense. In addition, uniformity across a wafer or from

wafer to wafer may be affected by shadowing by the collimator.

Another type of sputtering is called "long throw collimation," or "long throw sputtering." No collimation plate is actually used. Instead, a limited size target, about the size of the wafer in a single-wafer system, is used and the distance between the target and wafer is increased (up to about 30 cm compared to about 5 cm in more standard systems), decreasing the distribution of arrival angles at the wafers. In effect the sputtering chamber acts as one collimator cell, with the high-angle atoms not reaching the wafer but striking the chamber wall instead. However, as the spacing is increased, the deposition rate, the target utilization, and the uniformity all decrease; in addition, the chance of gas-phase collision and scattering increases. Low pressures are usually utilized in these systems to minimize gas-phase collisions. Because of the limited target size and target-to-wafer spacing, asymmetric depositions in trenches and vias can occur at the edges of wafers in these systems [9.13].

In ionized sputtering a narrow arrival angle distribution, perpendicular to the surface, is obtained as in collimated sputtering. However, most of the sputtered material reaches the wafers, giving it a significant advantage over collimated sputtering. In this technique, the sputtered atoms—mostly neutral as they leave the target—are ionized as they travel between the target and the wafers. An inductively coupled RF antenna is placed around the plasma region which ionizes the sputtered atoms through collisions with electrons in the plasma, as illustrated in Figure 9–31. Fifty to eighty-five percent of the atoms are ionized. The ionized atoms, Al^+ or Ti^+, for example, will be accelerated and directed across the wafer sheath, striking the wafers in a nearly perpendicular direction. A DC or RF bias is usually applied to the wafer holder, similar to bias sputtering. This will further accelerate these ions to the wafer surface, resulting in a highly directional flux. In addition, the energy of the incoming ions (both Ar and the deposited element) can be tightly controlled, and the deposition behavior and film properties can

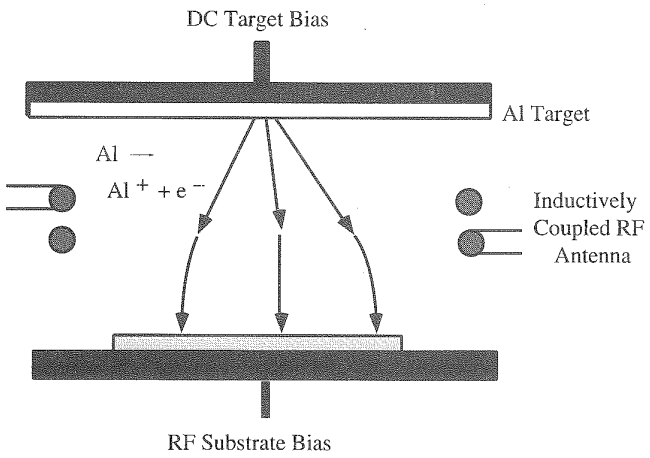


Figure 9–31 Schematic diagram of ionized sputter deposition system (or ionized PVD), showing atomic flux lines. (After [9.15].)

be altered, just as in bias sputtering. The highly directional ionized atoms and angle-dependent sputtering can result in very good coverage at the bottom of deep vias or contacts. However, the sidewall coverage may be poor. By increasing the ion energy, the resputtering and redeposition of the atoms can be increased, resulting in better sidewall coverage [9.14], often still thinner than the bottom, but at least continuous. Bevelling of the top corners due to angle-dependant sputtering may also occur.

Ionized PVD (IPVD) is also referred to as "ionized metal plasma" (IMP) deposition. It is also sometimes called "high density plasma sputtering," analogous to high density plasma CVD, because of the higher density plasma produced which ionizes the metal atoms. With the high density plasma produced, much more efficient deposition occurs compared to bias-sputtering or collimated sputtering, and IPVD is beginning to be more commonly used in chip production. The main problems with ionized PVD have to do with uniformity and manufacturing issues. The ionizing coils are usually located inside the chamber, and not remotely as in HDP CVD and etching systems. (Otherwise metal deposited on the chamber windows would block the signal from a remote RF source.) Having the coils inside the chamber leads to deposition on the coils and other associated problems. Another type of IPVD system utilizes a hollow cathode magnetron source rather than RF coils.

High-Temperature or Hot-Sputter Deposition Another method used to fill spaces during deposition as well as to improve overall coverage is high-temperature, or hot-, sputtering. This will be discussed further in Chapter 11 since it is most often used for Al deposition. The basic idea is to heat the substrate to 450°C or higher during deposition, usually by a wafer backside gas heating system. Surface diffusion is significantly increased so that filling in spaces, smoothing edges, and planarization are accomplished, all driven by surface energy reduction. This can be considered as *in situ* reflow. Usually a thin "cold" deposition is done first with the substrate at room temperature, which serves as a nucleation-wetting layer. CVD deposition of Al may also be used for this. This is then followed by the hot-PVD deposition. High-temperature sputter deposition is still in the development stage. One of the main drawbacks is the relatively high temperature required. Excessive reaction between the hot Al and the liner can occur. The high temperature can also adversely affect other materials and layers already present on the wafer at this stage in the process.

Even more recently developed is high-pressure sputter deposition. In this method, called the aluminum extrusion fill, or the "ForceFill" [9.16] technique, the aluminum is first sputter deposited, which covers and seals the holes. Then the chamber is pressurized with argon to up to 700 atmospheres while the wafer is heated to about 400°C. This forces the aluminum to completely fill any hole and does so quite effectively. Applying the high argon pressure after the deposition minimizes any argon incorporation in the film. The high temperature and specialized equipment are drawbacks to this technique.

9.3 Manufacturing Methods

The different types of equipment used for the deposition of thin films in microelectronics were discussed in the previous section. In this section the techniques used today to deposit specific thin films will be briefly described. Table 9–1 provides a summary.

Table 9–1 Common deposition methods for thin films in integrated circuit fabrication

Thin Film	Equipment	Typical Reactions	Comments
Epitaxial silicon	APCVD, LPCVD	$SiH_4 \rightarrow Si + 2H_2$	1000–1250°C.
		$SiCl_4 + 2H_2 \rightarrow Si + 4HCl$ Also $SiHCl_3$, SiH_2Cl_2	Reduce pressure for lower-temperature deposition.
Polysilicon	LPCVD	Same as epitaxial Si	575–650°C. Grain structure depends on deposition conditions and doping.
Si_3N_4	LPCVD, PECVD	$3SiH_4 + NH_4 \rightarrow Si_3N_4 + 12H_2$	650–800°C for oxidation mask. 200–400°C (PECVD) for passivation.
SiO_2	LPCVD, PECVD, HDPCVD, APCVD	$SiH_4 + O_2 \rightarrow SiO_2 + 2H_2$ $Si(OC_2H_5)_4 (+O_3) \rightarrow SiO_2$ + byproducts	200–800°C. 200–500°C (LTO)—may require high T anneal. 25–400°C (TEOS-ozone, PECVD, HDPCVD).
Al	Magnetron sputter deposition		25–300°C (standard deposition). 440–550°C (hot Al for in situ reflow). CVD difficult for alloys (Al-Cu-Si).
Ti and Ti-W	Magnetron sputter deposition (standard, ionized, or collimated)		CVD difficult. Nitrogen can be added to Ti-W to stuff grain boundaries.
W	LPCVD	$2WF_6 + 3SiH_4 \rightarrow 2W + 3SiF_4 + 6H_2$ $WF_6 + 3H_2 \rightarrow W + 6HF$	250–500°C. Blanket deposition with two-step process using both reactions is common.
$TiSi_2$	Sputter and surface reaction	Ti(sputtered) + Si(exposed) $\rightarrow TiSi_2$	Sputter/reaction gives self-aligned silicide.
	Cosputtering or CVD		
TiN	Reactive sputter deposition	$Ti + N_2$ (in plasma) $\rightarrow$ TiN	Organometallic source possible for MOCVD deposition.
	CVD	$6TiCl_4 + 8NH_3 \rightarrow 6TiN + 24HCl + N_2$	TiN can also be formed in $TiSi_2$ process.
Cu	Electroplating, electroless, sputtering, CVD	$Cu^{2+} + 2e^- \rightarrow Cu$	Electroplating is most common method today.

9.3.1 Epitaxial Silicon Deposition

"Epitaxy" is from the Greek for "arranges upon." Epitaxial silicon is silicon that is deposited on top of single-crystal material (usually the silicon substrate) and "arranges upon" and takes the crystalline form of the substrate. If no underlying single-crystal silicon is present, or if the deposition conditions are not right, then either amorphous or polycrystalline silicon will result. An example of an epitaxial layer is shown in Figure 2–20 in the CMOS process in Chapter 2.

Epitaxial silicon is deposited by CVD by either decomposing silane via the reaction

$$SiH_4 \rightarrow Si + 2H_2 \tag{9.28}$$

or by the reaction of a silicon chloride compound with hydrogen. Silicon tetrachloride, $SiCl_4$, is often used, and the overall reaction is

$$SiCl_4 + 2H_2 \rightarrow Si + 4HCl \tag{9.29}$$

Two other chlorides, $SiHCl_3$ and SiH_2Cl_2, are also sometimes used as the source species, with similar reactions. Eq. (9.29) is the overall, or net, reaction for the tetrachloride source. Actually many intermediate and competing reactions occur, and the process is quite complex. The silicon chloride species are all liquids at room temperature and are delivered to the chamber using hydrogen as a carrier gas. Using silicon chlorides rather than silane as sources of silicon is beneficial in that chlorine removes metal contaminates from the deposited silicon film.

Cold-walled CVD systems have traditionally been used for epitaxial silicon, either in the horizontal tube configuration shown earlier in Figure 9–4(a), or in a "pancake" configuration, with a round vertical wafer holder with source gases coming from above. Systems using quartz halogen radiant heating are also used. In each of these systems the wafers are placed side by side (end-to-end) to ensure uniform depositions.

For conventional epitaxial growth, relatively high deposition temperatures of 1000–1250°C are used, the lower temperature for silane and the higher for $SiCl_4$. These high temperatures are used so that single-crystal layers are readily obtained. High temperatures are required to desorb any native oxide on the surface which would inhibit single-crystal growth. At temperatures over 1000°C at 1 atmosphere, thin silicon dioxide films that may form during the deposition become unstable and vaporize. Also at the higher temperatures the surface reactions for deposition are enhanced, especially the surface migration to kink sites and rearrangement of the atoms, allowing for single-crystal growth. It is important that the atoms on the surface have time to attach at the proper sites before too many more atoms arrive. If the deposition rate is too fast relative to the surface processes, polycrystalline material results. Because of the high temperatures and fast surface reactions, epitaxial growth in these systems usually occurs in the mass transfer limited regime. Therefore thickness uniformity requirements preclude the wafers from being stacked close together. Recently lower temperature systems have been developed, as will be described later in this section.

Conventional Si epitaxial deposition is done either at atmospheric pressure, or at somewhat reduced pressures of 80–200 torr, but still above normal LPCVD pressures

which are in the 0.5–2 torr range. Reducing the pressure gives better control over switching on and off dopant source flows and reduces autodoping so that more abrupt doping profiles may be obtained. This works well for arsenic and phosphorus, but not as well for boron. Decreasing the pressure too much may cause the deposition rate to be quite low, by perhaps limiting the reactant species, especially for the chlorine-based sources. In earlier years when most epitaxial deposition was done using these conventional systems, the epitaxial layers were up to 20 μm thick and higher deposition rates were required.

A clean surface is critical to obtaining epitaxial growth. If contaminants or native oxide films are present, single-crystal growth will not occur. Wafers must be carefully cleaned prior to loading into the system. In addition, after the wafers are loaded into the deposition chamber, they may be subjected to an in situ HCl gas clean. This will etch the top surface of the silicon and remove any contaminates or native oxide.

Doping of epitaxial Si films can be done during the deposition by introducing appropriate gas sources of the dopants into the chamber along with the silicon source gases, as discussed earlier in the section on APCVD. Common dopant source gases are arsine (AsH_3), phosphine (PH_3), and diborane (B_2H_6). Unintentional doping due to diffusion and autodoping can be a problem, as discussed earlier. Reducing the pressure and increasing the temperature seem to reduce the problem for arsenic and phosphorus, while increasing the pressure and reducing the temperatures work better for boron. Using antimony instead of arsenic or phosphorus helps due to antimony's low vapor pressure compared to arsenic and low diffusivity compared to phosphorus.

Selective deposition of epitaxial silicon is sometimes desired. Selective deposition means that the silicon deposits on exposed regions of silicon, but not on other films, such as silicon dioxide. Under non-selective conditions, epitaxial silicon would be deposited over the exposed silicon, and amorphous or polycrystalline silicon would be deposited over the other regions. But by using the right conditions, one can actually achieve selective deposition. Conditions are chosen to reduce the nucleation and growth rate on the less-favored nonsilicon regions and to enhance the surface migration of the silicon atoms. Chlorides in the source gas tend to inhibit nucleation on the oxide by Cl etching of Si and by increasing the surface migration of Si atoms. Therefore using $SiCl_4$, $SiHCl_3$, or SiH_2Cl_2 increases the selectivity over using SiH_4. HCl gas is often added to improve selectivity. Reducing the pressure also helps.

Recently the need for low-temperature epitaxial growth of Si has greatly increased. In advanced process technologies, minimizing the thermal processing that the wafer is exposed to is crucial in order to decrease dopant diffusion, control abrupt interfaces and junctions, and reduce damage to device structures. Reducing the growth temperature from 1000°C or higher to temperatures in the 500–900°C range is of great benefit in these regards. Recently techniques have been developed to provide good quality epitaxial silicon films at low temperatures. The key to this is to provide an ultraclean environment so that oxides and other contaminants that would inhibit single-crystal growth are not present on the wafer surface. This can be achieved by using high-vacuum compatible CVD systems, ultrapure gases, and utilizing low pressures. A good predeposition clean is critical, and the wafers can also be initially heated to high temperatures in vacuum or H_2 in the chamber prior to the de-

position to desorb any native oxide. The low deposition temperatures and pressures usually place the deposition in the surface reaction limited regime. The deposition rates in these systems may be lower than in conventional epitaxial systems, but the layers that are grown today are generally much thinner than in older technologies.

Another technique for low-temperature Si epitaxy is Molecular Beam Epitaxy (MBE). Used for many years to produce III-V epilayers, this method is now being used more frequently for Si and SiGe films. In this technique, the silicon and dopant atoms are evaporated under ultrahigh vacuum conditions onto the Si substrate. Low-temperature epitaxial deposition, in the 500–800°C range, is obtainable, with ultraabrupt interfaces and doping profiles.

9.3.2 Polycrystalline Silicon Deposition

Polycrystalline silicon, unlike epitaxial silicon, can be deposited on arbitrary substrates and does not need exposed silicon underneath. It is used for the gate electrode in CMOS technology and for local interconnects and resistors. Its good thermal stability, good interface to silicon dioxide, good conformality, and ease of deposition and processing have made it a mainstay of silicon microelectronics technology. Polycrystalline silicon, as its name implies, is made up of regions or "grains" of crystalline material, but with each grain oriented differently from the neighboring grain. Separating each grain from the next one is a region called, appropriately, the grain boundary. Polycrystalline structures, grains, and grain boundaries will be discussed in detail in Chapter 11.

Polycrystalline silicon, or polysilicon, can be sputter deposited; however, CVD deposition is the usual method today. This is mainly due to two reasons, one historical and one technological. As we described above, epitaxial silicon must be deposited at relatively high temperatures in order for single-crystal material to be obtained. The low temperatures of sputtering (room temperature up to a maximum of about 400°C) preclude its use in depositing epitaxial film. Once CVD methods were developed for epi, it was a simple matter to use this method for polysilicon, where the conditions are much less restrictive. The other reason that CVD is used for polysilicon rather than sputtering is due to step coverage issues. In actual devices and circuits, polysilicon must often be deposited across steep topography. (See Figure 2–25 for example.) Good step coverage is important, and as we have mentioned many times, CVD usually gives better step coverage than sputtering. The conformality of LPCVD polysilicon is extremely good, with a sticking coefficient of 0.001 or less.

Silane is most often used for polysilicon deposition, following the decomposition reaction in Eq. (9.28). Silane is used rather than the silicon chlorides because better coverage over amorphous materials such as SiO_2 is obtained. This may be related to enhanced surface mobility in the presence of the chlorides or to the chlorine etching of Si over an insulator, leading to better nucleation and growth on the silicon and worse on the oxide. What is better for selective epitaxy is what is worse for depositing polysilicon over oxide layers, and vice versa. In addition, moderately low temperatures are usually used for polysilicon deposition, and silane decomposition occurs at lower temperatures than with the chloride reactions. If a dilutant or carrier gas is used, it is commonly nitrogen or hydrogen. However, using hydrogen increases the reverse reaction of

Eq. (9.28), and can decrease the deposition rate at lower temperatures. This decrease in the surface reaction shifts the deposition rate versus $1/T$ curve to the left compared to when nitrogen or no carrier gas is used, as shown in Figure 9–8 [9.17].

In order to achieve good throughput with uniform films, LPCVD systems with stacked wafers are used for polysilicon. Such a system was shown schematically in Figure 9–4(b). The deposition pressure is generally in the 0.2 to 1 torr range. A dilutant or carrier gas may or may not be used with the silane. By using a low pressure at the relatively low deposition temperatures described below, one ensures that the deposition occurs in the surface reaction limited regime while maintaining adequate deposition rates. The use of low pressure has been seen to greatly improve the uniformity of polysilicon depositions in closely stacked wafer systems at these moderate temperatures. Using lower pressure also results in fewer gas-phase reactions and less gas usage.

The deposition temperature for polysilicon films is moderately low, usually in the 575 to 650°C range. One reason for these low temperatures is to minimize the thermal budget of the wafer, reducing dopant diffusion and material degradation. Also, fewer gas phase reactions occur at the lower temperatures, leading to smoother and better adhering films, and silane depletion is less at lower temperatures. Another reason is that small grains are usually desired for polysilicon films. This is because films with finer grains are easier to mask and etch to give smooth and uniform edges. Deposition of polysilicon at higher temperatures usually results in larger grained films. For temperatures less than 575°C or so, the deposition rate is too low.

The deposition rate depends on the temperature, the silane partial pressure, and the presence of hydrogen or doping gases. At 620°C with a partial pressure of silane (with nitrogen carrier gas) of 0.5 torr, the deposition rate is about 15 nm per minute. According to our simple model for deposition, the deposition rate should be a linear function of the silane partial pressure, as given in Eq. (9.10). But the deposition rate for polysilicon has been found to be a sublinear function of the silane partial pressure, saturating as the silane is increased. This may be due to the desorption of hydrogen from the surface, which could be the limiting step in the deposition mechanism.

The as-deposited grain structure depends on these same factors: the temperature, the silane partial pressure, and the presence of other gases or species [9.17, 9.18]. Usually the polysilicon film is columnar in structure, meaning that most grains span the entire height of the film, with the grains quite tall and narrow in some cases. The average size at the top of the grains is often a little larger than at the bottom. Figure 9–32 shows several TEM cross-section images of deposited polysilicon films with the grain structure evident [9.19]. Polycrystalline films often exhibit a certain crystallographic texture. Crystallographic texture means that while the in-plane orientation of the columnar grains is more or less random, there is a preferred orientation of the grains in the vertical direction or out-of-plane orientation. For example in <110> textured films, which are common for typical polysilicon deposition conditions, most of the grains would have their <110> orientation pointing upward, or close to it. (And a more textured film means that a higher percent of the grains are at, or close to, a single orientation.) These orientations are a function of the deposition conditions, particularly deposition temperature and partial pressure of silane. If the deposition temperature is below 570–600°C, depending on the silane partial pressure, the as-deposited film will not be polycrystalline, but amorphous. Annealing the film at 600–1000°C will convert the struc-

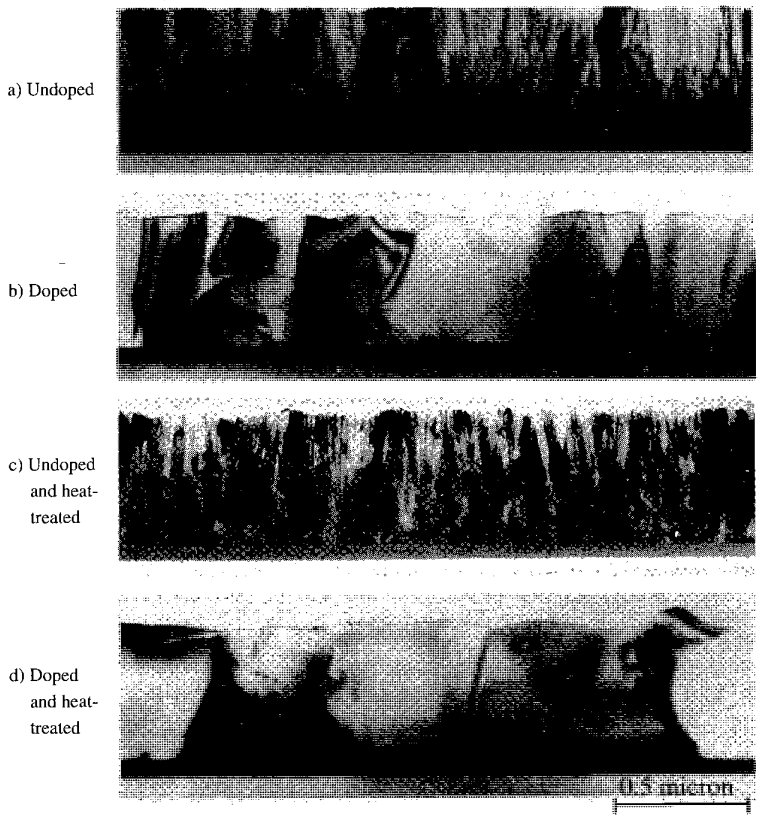


Figure 9–32 TEM cross sections of CVD polycrystalline films deposited at 625°C: (a) as-deposited, undoped film, showing the thin grains in a columnar structure; (b) as-deposited phosphorus-doped film, showing much larger grain size; c) annealed (1000°C), undoped film, showing little grain growth as compared to (a); (d) annealed (1000°C), phosphorus-doped film, showing evidence of grain growth as compared to (b). Reprinted with permission of the Electrochemical Society [9.19].

ture to polycrystalline material.

The average diameter of the columnar grains, the grain size, also depends on the deposition conditions and can range from 0.01 to 0.4 μm, with a typical size of about 0.05–0.1 μm for undoped polysilicon deposited at 625°C. Higher deposition temperatures generally result in larger grain sizes, as do thicker film depositions. High doping with phosphorus or arsenic (greater than about 3×10^{20} cm^{-3} for P and greater than about 1×10^{19} cm^{-3} for As) can greatly increase the grain size of as-deposited films, as shown in Figure 9–32. During any subsequent high-temperature annealing, the average grain size will increase only slightly for as-deposited undoped polycrystalline silicon when annealed at 1000°C (more grain growth occurs at higher temperatures above 1200°C). For

as-deposited amorphous silicon, the crystalline grains will nucleate and grow much more, even exceeding the size of the grains of annealed polycrystalline films. They also end up being somewhat less columnar, with some grains nucleating and growing above and below each other. The final size in either case will depend on deposition and anneal temperatures, anneal time, dopants, impurities, and film thickness. For example, phosphorus and arsenic doping will significantly increase the grain growth, while boron does not. This may be related to the increased vacancy concentrations with n-doping [9.20]. Grain growth and the dependence on film thickness will be discussed in more detail in Chapter 11.

Polysilicon can be doped either n type or p type. This can be done either during the deposition or afterward by implantation or diffusion. Doping during deposition is done by adding dopant source gases, such as those used for epitaxial growth, into the chamber along with the silane. However, the addition of arsine (AsH_3) or phosphine (PH_3) greatly decreases the deposition rates, by competing for sites perhaps, as well as creates thickness variations in the films. For these reasons, n-type doping of polysilicon is usually done after the deposition by implantation or diffusion. Diborane (B_2H_6) on the other hand, increases the deposition rate slightly and therefore can be used to dope the film during the deposition. As and P dopant atoms in polysilicon tend to segregate to the grain boundary regions. Here they are electrically inactive, so that only the dopants that remain in the grain interiors contribute to the doping of the film. B does not appear to segregate to grain boundaries. In addition, the grain boundaries can trap carriers regardless of the dopant, also reducing the net doping. The carrier mobility is also decreased by the grain boundaries. As a result, polysilicon must be doped at higher than normal concentrations in order to achieve desired doping and resistivity levels. Figure 9–33 extends the resistivity data we discussed in Chapter 1 (Figure 1–18) to the polysilicon case and shows the resistivity of annealed films as a function of phosphorus doping concentration. One can see that the resistivity for polycrystalline material is higher than single-crystal silicon for the same total doping level. This difference gets smaller at higher doping concentrations, presumably because the finite number of traps at the grain boundaries get saturated above some critical concentration (about 1×10^{17} cm^{-3} here). One sees that for samples deposited at 580°C (indicated by the LP-580 label, for LPCVD deposition at 580°C), the critical doping concentration is less. This is consistent with a larger grain size for annealed amorphous films, as discussed above. Larger grains means less grain boundary area and thus fewer segregated dopant atoms and fewer carrier traps.

9.3.3 Silicon Nitride Deposition

Silicon nitride, Si_3N_4, films are used primarily for two purposes. One is for a mask against oxidation. In earlier chapters we saw how Local Oxidation of Silicon (LOCOS) occurs using silicon nitride as a mask against oxidation (Figure 2–5). The diffusion of oxygen through the film is very slow, preventing the underlying silicon from being oxidized. The other main use of silicon nitride is as a final passivation layer on chips. Silicon nitride is a very good barrier against water and sodium diffusion (unlike silicon dioxide) and it can be deposited conformally, making it an excellent final passivation

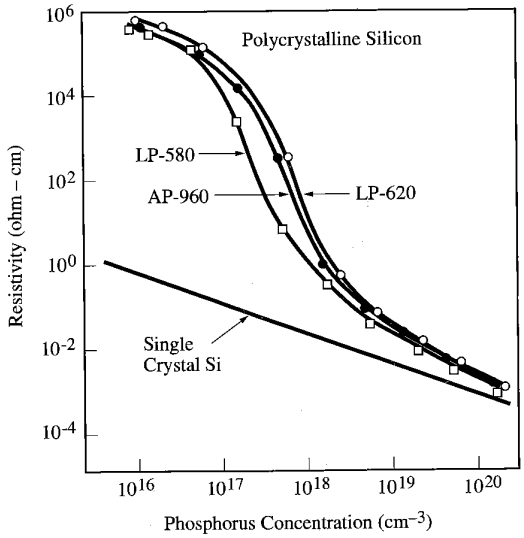


Figure 9–33 Resistivity of annealed phosphorus-implanted silicon films as a function of phosphorus concentration. LP refers to LPCVD deposition, AP refers to APCVD deposition, and the numbers in the labels refer to the deposition temperature. (After [9.21].)

layer for silicon microelectronics. It is generally not used in direct contact with silicon because of its relatively poor interface properties, especially with regard to fixed or interface trapped charges, and its high stress.

The two uses of silicon nitride require different deposition techniques as dictated by the allowable deposition temperatures. The masking nitride film is deposited relatively early in the fabrication process, and the temperature can be high. However, the passivating nitride is deposited after the metallization is in place, and therefore the deposition and processing temperature is limited to 450°C.

Masking silicon nitride films for LOCOS processes are deposited by CVD in the 650–800°C range. Low-pressure CVD is frequently used for good uniformity and high wafer throughput. The source gasses are ammonia and either dichlorosilane or silane, following the reactions

$$3SiH_2Cl_2 + 4NH_4 \rightarrow Si_3N_4 + 6HCl + 6H_2 \quad (9.30)$$

or

$$3SiH_4 + 4NH_4 \rightarrow Si_3N_4 + 12H_2 \quad (9.31)$$

To ensure the proper stoichiometry of the film, an excess of ammonia is used often by as much as 20-to-1 over the chloride or silane. Nitrogen can also be used in place of the ammonia. Significant amounts of hydrogen, up to 8 atomic percent, are incorporated into these films, bonded to both the silicon and nitrogen. This can affect the film properties, such as density and etch rate.

Silicon nitride films used for final passivation films must be deposited at tempera-

tures of 450°C or below, as mentioned above. Because of this, PECVD is the common technique employed here. Silane and ammonia are used, following Eq. (9.31), since silane works better at the lower temperatures than do the chlorides. Typical PECVD nitride deposition temperatures range from 200 to 400°C. The film that is deposited at these low temperatures using PECVD is often not stoichiometric Si_3N_4 and is often referred to as just SiN or Si_XN_Y. In fact, even more hydrogen than in the high-temperature LPCVD films is incorporated, up to 25 atomic percent, and the film is often designated by $Si_XN_YH_Z$. Oxygen can also be present in the film. The deposition rate and the film properties are very dependent on the deposition conditions, including the RF power and frequency, the gas flow, and the chamber pressure. The various film properties, such as composition, stress, dielectric constant, thermal expansion coefficient, and density, can be optimized, or at least traded off between each other, by varying the various deposition conditions.

To reduce the Si/SiO_2 interface trapped charge density in the active devices, an anneal is done in a hydrogen ambient near the end of the fabrication process. Because a nitride film is a barrier to hydrogen diffusion, the hydrogen anneal must be done prior to the deposition of the passivation nitride film.

9.3.4 Silicon Dioxide Deposition

Deposited films of silicon dioxide are utilized throughout silicon microelectronic structures as described in the CMOS process flow in Chapter 2 (see, for example, Figures 2–29 and 2–38). They are used for dielectric layers to electrically separate conducting regions as well as for masks against implantation and diffusion. The requirements for the films will be discussed in more detail in Chapter 11, but obtaining films of good quality in terms of density, purity, and thermal stability, as well as good step coverage and filling properties are of major concern.

Most oxide films today are deposited at relatively low temperatures, from 200°C to 500°C using CVD techniques. Sputtered oxide films have worse coverage and more particulates (mostly due to flaking off of the metal chamber walls) than CVD oxides. Atmospheric systems have been used for many years, but today LPCVD and PECVD systems have become more commonly used.

Silane and oxygen are common source gases for CVD silicon dioxide deposition, following the net reaction

$$SiH_4 + O_2 \rightarrow SiO_2 + 2H_2 \tag{9.32}$$

Oxidants other than O_2 are sometimes used, such as N_2O, NO, and CO_2. When nonplasma, silane-based deposition of SiO_2 is done at temperatures less than 500°C, it is commonly referred to as Low-Temperature Oxide, or "LTO," deposition.

Other sources of silicon are also used. One such source is tetraethooxysilane—$Si(OC_2H_5)_4$—or "TEOS." It can form silicon dioxide by decomposing via the reaction

$$Si(OC_2H_5)_4 \rightarrow SiO_2 + \text{byproducts} \tag{9.33}$$

(O_2 is often added as an extra oxidizer to reduce the carbon content in the film.) This

deposition occurs at higher temperatures, in the 650–800°C range, but because of the low sticking coefficient of TEOS compared to silane, more conformal films are produced. The deposition temperature can be lowered to below 400°C by either utilizing plasma enhanced deposition, or by including ozone, O_3, in the chamber:

$$Si(OC_2H_5)_4 + O_3 \rightarrow SiO_2 + \text{byproducts} \tag{9.34}$$

Because at low pressures (<150 torr) the deposition using TEOS and O_3 is limited by O_3 and gives worse film properties, a so-called Sub-Atmospheric CVD (SACVD) system has been developed for doped premetal films and undoped intermetal films [9.22, 9.23]. The deposition pressure in this system is at higher pressures (200–400 torr) and the deposition temperature is 300–550°C. Oxide films with good gap fill and reflow ability can be obtained with this system.

The major drawbacks to using TEOS are that it is a more difficult source to use (it is an organometallic liquid), the relatively high levels of carbon in the deposited film (from the TEOS itself), and the more porous films that are sometimes formed. But TEOS is still commonly used due to the conformality of the films it produces.

Silicon dioxide films deposited using LPCVD or APCVD at low temperatures (less than 500°C) can be nonstoichiometric and are often porous. They can also exhibit inadequate step coverage or filling. The films that are deposited before aluminum is on the wafer are usually annealed at high temperatures after their deposition. This densifies the films, as well as allowing them to flow, or "reflow," in order to achieve better filling and step coverage, as discussed in Chapter 11. This is done at temperatures of 1000–1100°C for undoped films, and as low as 800°C for doped (with B and/or P) films.

For films that are deposited after aluminum is already on the wafer, other measures must be taken. No high temperature densification anneal can be done, so the films must be of adequate quality and have good coverage "as-deposited." TEOS with ozone and plasma enhanced CVD TEOS are commonly used. While PECVD TEOS gives dense films and fairly good coverage and filling at low temperatures, the films usually contain a lot of hydrogen or nitrogen, up to 10 atomic percent, which can affect their properties. TEOS/O_3 films may be more conformal and have better filling ability, but are also much more porous and can absorb moisture.

Another method of silicon dioxide deposition, bias sputtering, has been developed to improve filling and planarizing. But because of its low deposition rate and a high density of particulates, this method is not commonly used today. A relatively new technique called High-Density Plasma (HDP) CVD, which combines PECVD with bias sputtering, is being used for low-temperature deposition (150°C down to room temperature). Figure 9–34 shows examples of SiO_2 deposited by HDPCVD, displaying excellent filling and good planarizing properties. In this method, both silane-and TEOS-based chemistries can be used, giving good quality films in high aspect ratio structures.

Deposited silicon dioxide films are often doped with phosphorus or boron, or both. Phosphorus is used to reduce stress and enhance Na gettering, while the boron allows more reflow at lower temperatures, as will be discussed in Chapter 11. Doping can be done by using hydrides for the silane case and organometallic compounds for the TEOS case in APCVD, LPCVD and PECVD configurations.

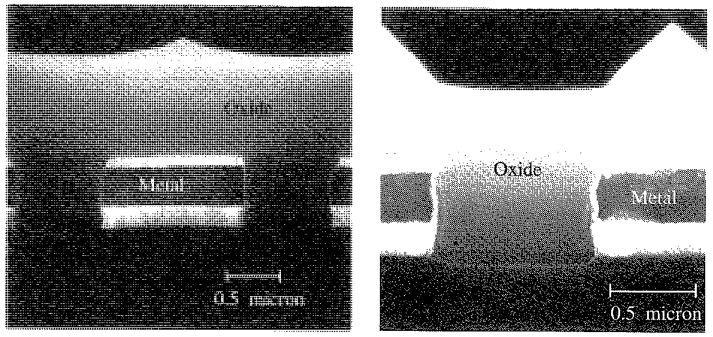


Figure 9–34 SEM images of High-Density Plasma (HDP) CVD deposited SiO_2, showing good planarizing and excellent filling properties. Photos courtesy of VLSI Technology, Inc.

9.3.5 Al Deposition

Aluminum is used as the main interconnect material in silicon microelectronics. As explained in Chapter 11, pure Al is usually not used, but an alloy of Al with 0.5–1 wt. percent Cu (to reduce hillocking, voiding, and electromigration) and sometimes with about 1 wt. percent Si (to prevent spiking and other interactions). The technique almost always used is DC magnetron sputter deposition. CVD techniques are not currently used for Al (but are being investigated) because of the problems with CVD of metal alloys, and Al is almost always used as an alloy. These problems include unwanted reactions between the precursors for the various metals, carbon contamination, rough films, low deposition rate, and inefficient use of the Al precursor which can result in a very costly deposition. Sputter deposition does not have these problems. Magnetron sputtering is used primarily due to its high deposition rate of up to one μm per minute for Al.

The argon pressure and the presence of impurities can affect the deposition and properties of the film. While a higher pressure of argon will increase the number of collisions of the Al atoms in the gas phase, decreasing the mean free path and improving the step coverage, a higher pressure of argon also means that more Ar will be incorporated in the Al film, up to 1 atomic percent. This can have deleterious effects on the film properties, including higher electrical resistivity, more stress, and smaller grain size (the argon acting as nucleation sites for the grains). Therefore lower pressures are usually desired. The presence of other gas species also affects the deposition and film properties. Oxygen, nitrogen, hydrogen, and water vapor may all be present in small amounts in the deposition chamber or on the wafer surface. These can all be incorporated into the film, having the same effects on the film properties: higher electrical resistivity, more stress, and smaller grain size. It is important to minimize the amount of these impurities

in the deposition chamber. This is done by pumping down the chamber to very low pressures before the argon is introduced (to less than 1×10^{-10} torr in some cases), by using ultrapure argon, and by heating the wafers and wafer holders before the deposition occurs to desorb any moisture.

It is common to heat the wafers to 150–300°C during Al sputter deposition, mainly to improve step coverage by increasing the surface mobility. Heating the substrates to even higher temperatures, up to 450–550°C, is called high-temperature or hot sputter deposition, or just "hot Al". The surface diffusion (and possibly bulk viscous flow) increases significantly, allowing for filling of deep contact and via holes.

There has been quite an interest recently in CVD Al. This is because of the technique's inherently better step coverage and excellent filling ability. There has been some limited success at depositing an Al-Cu alloy by using different Al and Cu organometallic precursors [9.24]. However, a more common approach involves depositing a film of pure Al by CVD, then doping it with Cu afterward. Al itself can be deposited by CVD using metal-organic sources, such as tri-isobutyl-Al (TIBA), or dimethylaluminumhydride (DMAH). The presence of Al-H bonds in the latter precursor results in less carbon incorporation in the Al film. To alloy or dope the film with Cu, one method is to sputter a layer of Cu, or Al-Cu alloy, on top of or below the CVD Al layer. Then a 250–400°C anneal is done in which the Cu diffuses into the pure Al layer, producing a uniform layer of Al-Cu. However, controlling this process can be difficult.

9.3.6 Ti and Ti-W Deposition

Titanium is often used as an underlayer for contacts, vias, and interconnects because of its good adhesion to other materials, its ability to reduce native oxides, and its good electrical contacting properties (as discussed in Chapter 11). It is usually deposited by sputtering, either using standard magnetron sputtering, or by using collimated or ionized sputtering for good coverage in contact or via bottoms. CVD of Ti is difficult because the precursors are so thermodynamically unstable. Impurities incorporated into the film from the precursors can also be a problem.

Ti-W alloys have been used as barrier metals in contacts and as underlayers and antireflective layers in interconnects, as also discussed in Chapter 11. However, TiN is more commonly used today for these purposes due it its better thin film and barrier properties and will be discussed separately later. Ti-W films are deposited using magnetron sputtering, usually from a single target. The most common target is 10 wt. percent Ti (30 atomic percent) and 90 wt. percent W (70 atomic percent). Nitrogen is commonly added to the chamber and is incorporated into the polycrystalline film, plugging or "stuffing" the grain boundaries. This reduces atomic diffusion along these high diffusion paths and improves the film's barrier performance. Film properties, such as resistivity, stress, composition, and barrier performance. depend mainly on four deposition conditions: sputtering pressure, power, temperature, and Ar/N_2 ratio. [9.25] The films can be rather brittle, leading to cracking on the wafer and to flaking off of deposition chamber walls onto the wafers.

9.3.7 W Deposition

Tungsten is commonly used as a contact or via metal, called "tungsten plugs," or in some cases is used as a first-level metal interconnect. (See, for example, Figure 2–41 in Chapter 2.) It is usually deposited by CVD, due to its very good filling ability and conformal coverage, in a hot-wall, low-pressure system. Temperatures range from 250–500°C, and total pressures are in the 0.1-to 2-torr range.

Tungsten can be deposited either selectively or as a blanket deposition. In selective deposition, W is only deposited where Si is exposed, such as in contact holes, and not on the surrounding SiO_2. In this process, the exposed Si is used to reduce the W source gas, WF_6, depositing W only in those regions.

Much more commonly used is blanket tungsten deposition. Here tungsten is deposited everywhere, both in the contact or via holes, and over the dielectric. (CMP shown in Figure 2–42 or an etchback process described in Chapter 11 can then be used to remove all the tungsten except in the contact or via holes.) Tungsten can be deposited by either the hydrogen reduction reaction

$$WF_6 + 3H_2 \rightarrow W + 6HF \tag{9.35}$$

or the silane reduction reaction

$$2WF_6 + 3SiH_4 \rightarrow 2W + 3SiF_4 + 6H_2 \tag{9.36}$$

A two-step process utilizing both these reactions can also be used, as will be described in Chapter 11.

Tungsten films deposited by CVD have significant amounts of fluorine incorporated in them as a result of the WF_6 source, often 1 atomic percent or more. This increases their resistivity over bulk values by 1.5 to 2 times. The grain size is usually less than 0.2 μm , yet the films are rather rough. This is due to random crystal faceting. This makes the films harder to etch evenly and can also lead to problems in the dielectric overlayers. Overall, good filling of contacts and vias with decent film properties can be achieved, and the use of tungsten plugs in current microelectronics is widespread.

9.3.8 $TiSi_2$ and WSi_2 Deposition

Titanium silicide, $TiSi_2$, and tungsten silicide, WSi_2, are common silicides used today in Si microelectronics. Other common silicides include $CoSi_2$, $TaSi_2$, $MoSi_2$, NiSi, and PtSi. We will focus on $TiSi_2$ and WSi_2 here, but many of the same deposition methods and ideas can apply to the other silicides as well. As we saw in the CMOS example in Chapter 2 and as we will discuss in much more detail in Chapter 11, silicides are used in CMOS technology extensively to reduce sheet resistance of polysilicon lines and n^+ regions. (See Figure 2–36.) They are often part of contact structures, either because they are already on top of the gates, sources, and drains or because they are deposited in contact holes. Silicides can also be used as local interconnects themselves.

The formation of silicides can be done in two general ways: by direct deposition of the silicide; or by deposition of the metal on top of Si followed by the reaction between the metal and Si to form the silicide. The direct deposition method can be accomplished in several ways: (1) sputtering from a composite target, (2) cosputtering from two targets of the metal and Si, (3) co-evaporation of the metal and Si, and (4) CVD.

The reaction method is more common today than the direct deposition method. (See Figures 2–35 to 2–37.) Here the metal, Ti for example, is deposited by sputtering on the exposed gate and/or source drain regions, all of which are silicon. The wafer is annealed and the silicide forming reaction occurs wherever the silicon and metal are in contact, that is, on the gate and source/drain regions. The unreacted metal is then selectively etched away. This method produces self-aligned silicide structures on the gate and/or source drains. This is often called the "salicide" technique, for "self-aligned silicide," and will be discussed more in Chapter 11. Another reaction method deposits the metal and then amorphous Si, both usually by sputtering, then patterns and selectively etches the amorphous Si. The wafer is then annealed so that the metal and amorphous Si react to form the silicide, and the unreacted metal is selectively etched off. This nonself-aligned method produces local interconnects independent of either gate or source/drain regions. For either type of reaction method, it is imperative to have the equipment (both sputtering chamber and annealing chamber) and the wafer surfaces free of oxygen, water, and native oxides. Their presence can result in films that are high resistivity, have uneven thickness and rough surfaces, and have poor adhesion properties.

WSi_2 has historically been used in a "polycide" structure. Here the silicide is blanket deposited over a blanket polysilicon layer, and both polysilicon and silicide are masked and etched at the same time to produce a low-resistivity gate structure. Both sputtered and CVD deposition of the WSi_2 have been used. Sputtering from a composite (WSi_2) target has been used, but CVD has the advantage of better step coverage, lower stress, and better adhesion. (WF_6 and SiH_4 are used in the CVD reaction, but a high SiH_4:WF_6 ratio is used so that WSi_2 is deposited instead of W.) However the trend today is to use the self-aligned silicides (namely $TiSi_2$ or perhaps $CoSi_2$) instead and form the silicide on the gate and source/drain regions all at once.

9.3.9 TiN Deposition

TiN films are used as barrier layers in contacts, and as underlayers and antireflective layers in interconnects, as discussed in Chapter 11. They have mostly replaced Ti-W in these applications due to their better barrier and film qualities. TiN films are often deposited or formed on top of Ti films, which have better contact and adhesion properties to the films underneath.

TiN is commonly sputtered using the reactive sputtering technique. As discussed earlier, reactive sputtering involves sputtering a metal in the presence of a reactive gas, such as nitrogen or oxygen, to form the metal oxide or nitride. In this case, nitrogen gas is mixed with the argon gas while Ti is being sputtered. The plasma breaks up the N_2 molecules into atomic nitrogen, which then react with the Ti to form the compound both on the target and on the wafer surface. The properties vary significantly with the

deposition conditions, which include the nitrogen flow or partial pressure, the total pressure, the sputtering power, the substrate temperature, and the substrate bias, if one is applied. While excess nitrogen, or the presence of oxygen, increase the film's resistivity, they also enhance the barrier properties by saturating the grain boundaries. The deposited film is typically fine grained, columnar, and with <111> orientation, although <100> orientation can be achieved under certain deposition conditions [9.25]. The microstructure and orientation also depend on the film underneath. For example, the crystallographic orientation or texture often follows that of the Ti film that is commonly underneath TiN films. TiN is a very hard metal, and the films can be highly stressed and brittle, like Ti-W. TiN is often sputter deposited when the TiN is part of the Al interconnect stack. Step coverage and filling are not usually issues here. Generally it is better to deposit the whole stack by one technique without leaving the deposition chamber to avoid surface oxidation and contamination. Ionized PVD of TiN can be used when TiN is utilized as a liner/barrier in vias.

To improve the step coverage and filling ability of TiN films, CVD methods for deposition of these films have gained a lot of interest recently [9.26]. Titanium tetrachloride, $TiCl_4$, can be reacted with ammonia, NH_3, to form the films at 400–700°C via the reaction

$$6TiCl_4 + 8NH_3 \rightarrow 6TiN + 24HCl + N_2 \tag{9.37}$$

In general, the higher the deposition temperature, the better the film properties. As the temperature increases, the resistivity of the films decreases, the film density increases, and the chlorine content in the films decreases. Chlorine incorporation can be significant, 0.5–5%, which can lead to corrosion problems in Al interconnects. CVD processes using Metal-Organic (MOCVD) precursors instead of ammonia have also been developed. These include tetrakis-dimethyamino titanium (TDMAT) and tetrakis-diethyamino titanium (TDEAT). TiN films can be deposited with these sources at lower temperatures and with no Cl contamination. However, the films typically have carbon and oxygen incorporation, lower density, particulates, and not as good step coverage or filling ability. Methods to improve these properties, including postdeposition annealing treatments, or plasma treatments, are being developed. PECVD methods using ammonia are also being developed.

Another method to deposit TiN films is by first sputtering Ti on the wafers and then annealing them in a nitrogen or ammonia ambient above 600°C [9.25]. This was the process used in the CMOS technology described in Chapter 2. The Ti film reacts with the nitrogen to form the TiN film. The reaction is limited by the diffusion of the nitrogen through the forming TiN film, and the formation rate exhibits a diffusion-limited square root of time dependence, as expected. The reaction is very sensitive to the presence of oxygen, which can actually stop the reaction. The anneal is often done in an RTA system in which the heating time and exposure to oxygen are minimized. This nitridation method is often done in conjunction with silicide formation (Figure 2–36), as discussed in more detail in Chapter 11. Here the anneal done to react the deposited Ti with the underlying Si is performed in a nitrogen ambient. While $TiSi_2$ forms on the bottom of the Ti, TiN forms on top.

9.3.10 Cu Deposition

As we will discuss in Chapter 11, manufacturing processes that use copper instead of Al as a global interconnect metal have recently been developed. Cu has a lower electrical resistivity than Al and has better electromigration resistance. However, there are several processing and process integration issues concerning the use of Cu. One problem with Cu is that it is difficult to plasma etch due to the low volatility of the etch byproducts. Therefore a process that does not require etching, such as selective deposition, would be attractive. However, by using the damascene process, which was described in the CMOS technology in Chapter 2, the deposition and patterning of Cu can be accomplished by doing a blanket deposition in an etched-out dielectric layer, followed by Chemical-Mechanical Polishing (CMP). In this manner, neither etching or selective deposition is required. The damascene process will be described further in Chapter 11.

Copper can be deposited by PVD methods, including both sputter deposition and evaporation. However, the step coverage for these is quite poor and filling small vias is difficult. A high-temperature reflow step can be done, but that raises the thermal budget of the process.

CVD processes for Cu deposition are being developed to give better step coverage and filling properties. Like Al CVD processes, metal-organic sources are commonly used. One such source is 2Cu(hfac)L, where (hfac) is hexafluoroacetylacetoneate and L is a neutral ligand weakly bonded to the Cu. One ligand is vinyltrimethylsilane. Both blanket and selective depositions can be done depending on the ligands and deposition conditions. However, it is difficult to simultaneously achieve low resistivity, good step coverage and via filling, and a high deposition rate with these methods [9.27].

An alternative deposition method that has been developed for Cu is electrolytic plating. This can be done with the use of external electrodes and applied current. Applying an external voltage is generally referred to as electroplating. Deposition without an applied field is called electroless deposition. Both processes are used commonly in printed circuit board manufacturing, and work is being done to apply these processes to on-chip interconnect fabrication. While the electroless method has received a lot of attention in recent years, mostly due to its selective deposition capabilities, the electroplating method has become the current method of choice for Cu deposition in integrated circuits.

In electroplating, the wafers are mounted on a cathode and immersed into a plating solution that contains Cu ions. An inert anode, made of platinum, for example, is also immersed into the solution. A voltage is applied between the two electrodes and the current drives the Cu ions toward the wafer, forming metallic Cu on the surface, according to the electrochemical reaction

$$Cu^{2+} + 2e^{-} \rightarrow Cu \tag{9.38}$$

A thin seed layer, commonly Cu as well, is deposited first over the wafer so that electrical contact can be made to the surface of the wafer and so that Cu is electroplated over the entire wafer surface. This can be done by sputtering or CVD. Excellent filling can be achieved with electroplating of Cu. Filling of gaps is enhanced by using "level-

ing" agents or additives in the solution. These chemicals inhibit the deposition. They preferentially adsorb on the top surfaces of the topography rather than on the bottom of trenches or vias due to diffusion limitations. The top surfaces and corners then receive less deposition than the bottoms of trenches or vias, resulting in excellent filling ability. Low resistivity and high deposition rate can also be achieved with electroplating, and this method is becoming the most common method to deposit Cu for interconnects.

The electroless deposition technique involves the formation of thin metallic films from electrolytic solution without the use of external electrodes and without applying an external current. Simultaneous oxidation-reduction reactions between two half-reactions occur to deposit the film from solution. By this method, both blanket and selective Cu films of good quality, low resistivity, and good filling properties can be obtained. In a typical process, the solution contains formaldehyde (HCOH) as a reducing agent, and CuEDTA as the oxidizing agent and source of Cu. The two half-reactions with their redox potentials are

$$2HCOH + 4OH^- \rightarrow 2HCOO^- + 2H_2O + H_2 + 2e^- \quad V = 0.32 \ (pH = 0.12) \tag{9.39}$$

$$[CuEDTA]^{2-} + 2e^- \rightarrow Cu + EDTA^{4-} \quad V = -0.26 \tag{9.40}$$

The net potential is positive so that the combined reaction is thermodynamically favorable:

$$[CuEDTA]^{2-} + 2HCOH + 4OH^- \rightarrow Cu + 2HCOO^- + 2H_2O + H_2 + EDTA^{4-} \tag{9.41}$$

Therefore by placing the wafer in a very basic solution containing Cu-EDTA and formaldyde, copper will be deposited from the solution. Complexing and surfactant agents are also commonly added to the solution.

In order for the copper to be deposited on the wafer surface, thin seeding layers are required in the electroless process, as in the electroplating process. Here the seed layer is needed to catalyze the electroless process. These are often Pd, Au, or Cu layers, deposited by CVD, sputtering, or evaporation. For selective electroless deposition, the seed layers should only be in the locations where the Cu deposition is desired. This can be done by selectively depositing the seed layer material itself, doing very directional PVD deposition with subsequent etchback or CMP, or using lift-off techniques [9.28]. For nonselective Cu deposition, the seed layer is deposited everywhere followed by the Cu electroless deposition. A problem with the electroless method has to do with contamination. The seed layer materials, Pd, Au, or Cu, can all act as deep-level traps in the underlying Si. Of more concern are the alkali ions needed for the highly basic solution. The hydroxide ion, OH^-, is commonly produced from sodium or potassium hydroxide. Sodium and potassium ions can drift in SiO_2 and alter the electrical properties of the active devices as we saw in Chapter 4. Tetra-methyl-ammonium-hydroxide ($N(C_2H_5)_4OH$), or TMAH, is often used instead of sodium or potassium hydroxide. However, the complex solution and reactions involved with the electroless technique compared to the electroplating method and the slower deposition rate, along with the trend toward blanket deposition of Cu in conjunction with the damascene process, have made the electroless process less attractive than the electroplating method for Cu deposition in IC technology.

9.4 Measurement Methods

The properties of the deposited films can be measured either during the deposition (in situ) or after the deposition is complete. The latter case is much more common. Properties such as thickness, composition, impurity concentration, step coverage, and stress, which often depend on the underlying layers and topography, are usually measured on the films after the deposition is complete, outside the deposition chamber. The measurement of these properties is discussed in the measurement methods section of Chapter 11 for films used in back-end technology. These films include silicon oxide, silicon nitride, silicide, and metal films, but the techniques are also used for other dielectric and interconnect materials, such as polysilicon, as well. In this section we will briefly describe some in situ measurement techniques.

The measurement of properties during the deposition is not very common today during IC fabrication and is mostly limited to film thickness and deposition rate measurements or to research applications. But as the technology progresses and more control is needed over film properties, there is much more interest in having automatic feedback during deposition. This requires in situ measurement of the film properties so that the deposition conditions can be adjusted as the deposition proceeds. The two most common methods of in situ measurement techniques are ellipsometery and the quartz-crystal microbalance.

Ellipsometry was discussed in Chapter 6 and can measure the thicknesses, and hence the deposition rates, of thin films on a reflective substrate. This technique is based on the change of polarization that occurs when a polarized light beam is reflected by a film. When data for multiple wavelengths are acquired, the technique is known as spectroscopic ellipsometry. This provides better measurement when multiple layers are present on the wafer. Ellipsometers are well suited for in situ deposition measurements. This is because the ellipsometer can be placed outside a chamber, using a window to allow the measurement. Even if the window becomes coated with the deposition material or byproducts, the measurement can still be done since the technique relies on relative intensities of light, and not on an absolute intensity which can change with time. The technique is only useful for transparent thin films.

The other common in situ measurement technique, QCM (Quartz-Crystal Microbalance), is based on the fact that the oscillating frequency of a crystal depends on its mass. A quartz crystal, sandwiched between two electrodes, is mounted in the deposition chamber in the same area as the wafers. As the film material is deposited on the wafers and the crystal, the mass of the crystal increases. This changes the oscillating frequency of the crystal in a well-understood manner, which is monitored through an external circuit, and the change in frequency is determined. This is then converted to a change in mass, which is then converted to a change in thickness. This technique is so sensitive that it is often used to calibrate other, less common, techniques, such as atomic emission and atomic absorption methods. This method can be applied to the deposition of any thin film.

9.5 Models and Simulation

While one important goal in thin film deposition is to obtain films with the desirable physical and chemical properties, an equally important goal is to achieve proper coverage and filling of the film over the underlying topography. This becomes even more critical as lateral dimensions decrease and aspect ratios get larger.

In this section we will focus on the modeling of thin film deposition as it relates to film coverage and filling over nonplanar topography. In the first part of the section we will utilize SPEEDIE [9.29], a 2D simulator developed at Stanford University. SPEEDIE uses physically based models for deposition and etching. Many other simulation tools have been developed for deposition and etching in integrated circuit fabrication, such as EVOLVE at Arizona State University, SAMPLE at the University of California, Berkeley, SIMBAD at the University of Alberta, and SHADE at IBM. Many of these are also physically based and contain many of the same basic mechanisms and models as in SPEEDIE. We will use SPEEDIE's simulation capability to predict the thickness and shape, in two dimensions, of deposited thin films as a function of various deposition conditions and configurations. These simulations will be used to illustrate many of the concepts discussed earlier in this chapter. Later in this section we will describe simulations using some commercially available simulation tools.

9.5.1 Models for Deposition Simulations

Figure 9–35 shows a generalized deposition picture of both CVD and PVD techniques. PVD systems use sputtering or heating of a source, whereas CVD systems use gas flow introduction of chemical precursors to provide the reactant species for deposition. The surface processes may also be different. Nevertheless, the steps are in general terms the same and can be modeled with the same components. Only the values used for the specific terms of the model will vary depending on the deposition technique and conditions. This is in fact a good way to analyze and compare the different deposition techniques. By starting with the same basic components and equations, and modeling the different techniques by varying the different parameters, one can more clearly see the similarities and differences between them.

In Figure 9–35 we focus on the wafer surface and the gas phase adjacent to the wafer surface. In the figure we see the direct flux of species, both neutrals and ions, arriving from the gas phase. The deposition species, Al in this case, can be either neutral or ionic. The ions, both Al^+ and Ar^+ in the figure, can also sputter atoms which have been deposited on the surface. In addition the ions, by their surface bombardment, can enhance the deposition rate of the neutral species. We also see in the figure species being desorbed, or "emitted." We see atoms being resputtered by ions. Both emitted and resputtered species can then be redeposited elsewhere, as shown at the bottom of the trench, and are called indirect fluxes. We also see another type of indirect flux: atoms moving due to surface diffusion. The indirect fluxes are all surface, or near-surface, processes.

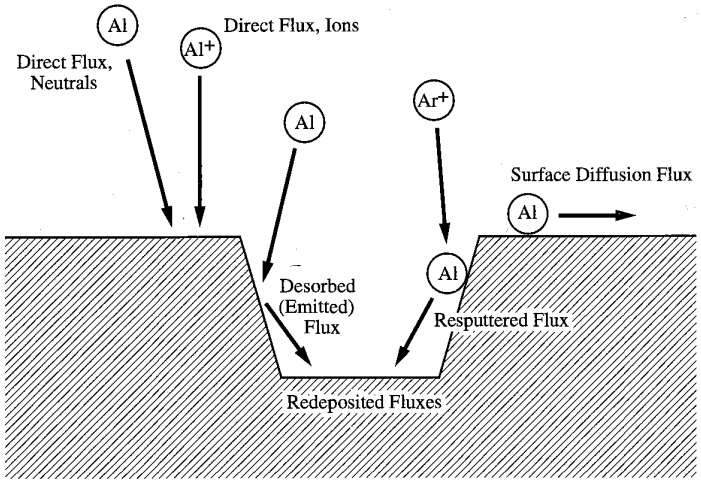


Figure 9–35 Generalized surface region processes in both CVD and PVD deposition systems used in simulation models.

A gas-phase boundary layer is not considered in this formulation since the gas pressure for most deposition systems is low enough so that diffusion through a boundary layer is not important. Because the mean free path in these systems is large compared to the feature sizes of the topography, free molecular flow (i.e., collisionless, nondiffusive flow) is assumed as the transport mechanism near the surface. (Diffusional flow through a boundary layer near the surface is important in APCVD systems, and SPEEDIE has a separate model for that. With many gas-phase collisions occurring within the topographical features in that case, good coverage and filling can result for processes that are reaction limited, such as with TEOS/O_3 SiO_2 [9.5].)

9.5.1.1 Models in Physically Based Simulators Such as SPEEDIE

In physically based topography models such as SPEEDIE all the phenomena depicted in Figure 9–35 should be taken into account. In SPEEDIE the net flux at each position on the surface is calculated based on all the fluxes arriving at and leaving that spot, either from the gas phase—the direct flux—or from other points on the surface—the indirect fluxes. The surface is then moved in the perpendicular direction according to the net flux, normal to the surface, at each point as the film grows. The net flux at each point i, F^i_{net}, can be expressed as the sum of the fluxes that are positive (due to atoms arriving at each point), minus each negative flux (due to atoms leaving each point):

$$F^i_{\text{net}} = F^i_{\text{direct(neutrals)}} + F^i_{\text{direct(ions)}} + F^i_{\text{redep}} + F^i_{\text{diff.in}} - F^i_{\text{emitted}} - F^i_{\text{sputtered}} - F^i_{\text{diff.out}} \tag{9.42}$$

The fluxes are all in units of atoms (or molecules) $\text{cm}^{-2}\,\text{sec}^{-1}$.

The *direct fluxes*, $F_{\text{direct(neutrals)}}$ and $F_{\text{direct(ions)}}$, are the fluxes of the different species at each position (or small area) i on that surface that arrive without any interaction with the wafer surface topography. These arrive directly from the gas phase above the wafer and include both neutral and ionic species. In calculating the direct fluxes, one has to know the distribution of arrival angles of the incoming species relative to the orientation of the surface at that point. Then the distribution can be integrated over all angles to obtain the total direct flux, normal to the surface, at each point.

There are two ways of modeling the direct fluxes. One is to start with the angular distribution of fluxes as the species depart from the target or gas source, or the emitted angle distribution at each source point. Then by taking into account the gas-phase collisions and the geometric configuration of the chamber (distance from target to source, etc.), as well as the surface topography, one can determine the distribution of angles of the precursor arriving at each point on the wafer surface. This is similar in a qualitative way to Figure 9–30 which illustrates deposition from different size sources. But including the gas-phase reactions and the various equipment geometries can get quite complicated. Techniques such as Monte Carlo methods are often used for this and are available in various deposition simulators, but even so, fitting parameters are usually needed.

A simpler way to account for the incoming flux is to just assume some distribution function of fluxes as a function of angle just above the surface of the wafer. This is illustrated in Figure 9–36. If the fluxes were coming in equally at all angles, that is isotropic arrival fluxes, then one would assume a constant flux for all angles in the distribution at each position in that plane. However, as we discussed earlier, we define the arrival distribution in terms of the normal component of the incoming flux relative to a unit area on a surface. In this case the surface is a horizontal plane just above the wafer surface. By simple geometry, the normal component would be proportional to the cosine of the vertical angle. We would therefore express this

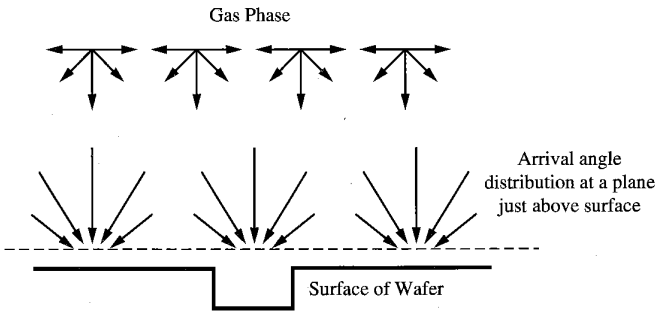


Figure 9–36 Simplified way of considering incoming flux distribution. A distribution of fluxes arriving from the gas phase, with a $\cos^n\theta$ distribution, is assumed in a plane just above the wafer surface. From that plane to the surface features below, no gas-phase collisions occur, just line-of-sight molecular flow.

isotropic arrival as a cosθ distribution, where θ is relative to the vertical direction. To account for nonisotropic distributions as well, this can then be generalized to the form.

$$F_{\text{direct}}(\theta) = F^{\circ} \cos^{n}\theta \tag{9.43}$$

Here F_{direct} is the normal component of the flux coming in at angle θ, F° is the flux at $\theta = 0°$, and θ is the angle from the vertical direction, as was indicated in Figure 9–25. By adjusting n from 1 to large numbers, one can obtain isotropic to very anisotropic distributions of arrival angles, as shown in Figure 9–37. With $n = 1$, the distribution is isotropic, with fluxes coming from above at all directions equally. Higher pressure, more collisions in the gas phase, and shorter mean free path would lead to a more isotropic arrival distribution. A larger source and shorter source-to-wafer distance also give more isotropic arrival. For $n > 1$, a tighter distribution of arrival angles would occur, as shown in the graph in Figure 9–37. This means that the fluxes arriving are more directed in the vertical direction. Lower pressure, fewer collisions in the gas phase, and a longer mean free path would lead to a more anisotropic arrival distribution, as would a smaller source and a longer source-to-wafer distance. For ionic species in biased systems (bias sputtering, ionized sputter deposition, HDPCVD), the n values would be quite high since the ions are accelerated and directed toward the surface by the imposed electric field. In practice, the value of n for the various species is first approximated according to the type of species (ionic or neutral, for example) and type of deposition system, as we will discuss later. The exact value of n for the various species is then often determined by fitting simulations to experimental profiles in calibration structures.

One must also specify the maximum, unobstructed flux of each species to normalize the flux distribution. This is the integrated flux of Eq. (9.43) over all directions above the surface and corresponds to deposition on an unobstructed flat surface.

Knowing the magnitude and distribution of fluxes coming from the gas phase, one must then take into account the local surface topography. This includes the relative orientation of the surface at each position i on the wafer surface, as well as any shadowing

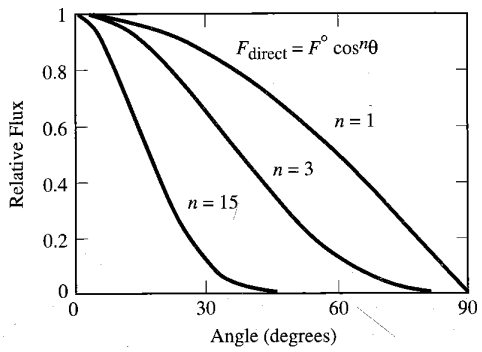


Figure 9–37 Distribution of arrival angles as function of angle to vertical direction, θ, for three values of n. The relative flux is $(F_{\text{direct}}/F^{\circ})$, where F_{direct} is the normal component of flux arriving at point i at angle θ. The $n = 1$ curve corresponds to isotropic arrival, with higher values of n representing more vertically directed arrival.

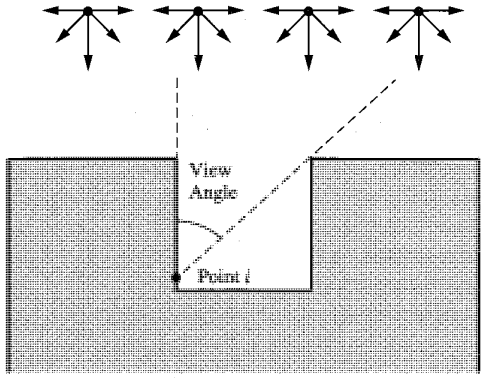


Figure 9–38 Illustration of viewing angle concept. A point, *i*, inside a trench would only see fluxes arriving within a limited range of angles due to geometrical considerations and shadowing.

effects from the topographical features. For example, the position may be on the sidewall of a step or trench and receive little flux from directly above with more coming in at an angle. Or the position may be at the bottom or side of a trench and see a limited amount of direct flux coming in since the fluxes outside the line of sight will be blocked or shadowed. Figure 9–38 illustrates this. One must take into account the line of sight, the "view angle," from the reference plane in the gas phase just above the surface to each position on the surface. (In fact, each small area on the reference plane can be assumed to be an emission source to the topography below, equivalent to a downward pointing Knudsen cell, for these calculations.) Assuming collisionless, line-of-sight transport in this region, all this geometry and shadowing can be taken into account and the total direct flux can be calculated at each position on the wafer surface for each species.

The *indirect fluxes* include the processes that occur on the surface, or near the surface of the wafer during deposition. Surface diffusion, emission, sputtering, and redeposition are all indirect fluxes.

The *surface diffusion fluxes*, $F_{\text{diff.in}}$ and $F_{\text{diff.out}}$, are driven by differences in the curvature in the surface. Here we are not concerned with the short-range diffusion of atoms to attachment sites as part of the surface reaction process, but rather the longer-range diffusion in order to reduce surface energy. The net diffusion flux ($F_{\text{diff.net}} = F_{\text{diff.in}} - F_{\text{diff.out}}$) at each point is calculated by

$$F_{\text{diff.net}} = \frac{D_s}{kT}\gamma_s \Omega \upsilon \frac{\partial^2 K}{\partial s^2} \tag{9.44}$$

D_S is the surface diffusivity of the material, γ_s is the per area surface energy, Ω is the atomic volume of the atom or molecule, υ is the number of atoms per unit area, or the surface density, K is the curvature, and s is the unit length along the surface the atom or molecule travels. In Chapter 11 in the section on reflow modeling we derive the expression for the surface flux and the normal velocity of a surface due to surface diffusion. Here we are interested in the flux relative to the surface normal direction, which is the velocity in that direction times the density. To get this, we multiply the expression

for the normal velocity due to surface diffusion of Chapter 11, Eq. (11.37), by the density, which is the same as dividing by Ω, and we end up with Eq. (9.44).

Because of curvature-induced surface diffusion during deposition, filling in of corners is enhanced and smoother and conformal depositions can result, as we will see in the simulations. Because atoms can diffuse along the surfaces to areas in less direct view of the source, better step coverage can be obtained. Surface diffusion also helps with filling vias and trenches and can lead to more planarized depositions as well. Fast surface diffusion is often used to explain, and model, conformal coverage in particular deposition cases. However, as we will see, emission and redeposition can model the same effects and may be more physically correct in some cases.

The *flux due to emission* of species from the surface, F_{emitted}, occurs because not all deposition species or precursors stay, or stick, where they land. A finite number, usually assumed to be a constant fraction of the incident species, will be emitted or desorbed from the surface without properly reacting or depositing there. The emitted flux at each point i is thus

$$F^i_{\text{emitted}} = (1 - S_c)F^i_{\text{incident}} \tag{9.45}$$

where F^i_{incident} includes both direct and indirect deposition fluxes incident at i and S_c is the sticking coefficient.

The sticking coefficient here is defined as the ratio of the flux of species that properly react and stay on the surface to the flux of incident deposition species:

$$S_c = \frac{F_{\text{reacted}}}{F_{\text{incident}}} \tag{9.46}$$

This is really an effective sticking coefficient or probability, since factors such as site availability, reaction competition and completion, and surface recombination are implicitly included in it. It is also sometimes referred to as the reactive sticking coefficient, and is often empirically determined from deposition profiles. The sticking coefficient depends on the type of deposition system, the chemistry, and material being deposited. It can also depend on substrate temperature. The temperature dependence is often difficult to predict because of all the processes that may be involved. A lower value is seen at higher temperatures in some cases, such as in the case of LPCVD SiO_2 [9.30]. In other cases, such as with CVD W deposition, a lower value is sometimes seen at lower deposition temperatures. In more advanced models S_c may not be a constant, as is assumed here.

The fraction that does not stay at a particular position and is emitted is therefore equal to $(1 - S_c)$. The F_{emitted} flux is thus subtracted in the total flux equation [Eq. (9.42)]. It is often assumed that the emitted species are emitted in a range of directions away from the surface in a perfectly diffuse angular distribution, with a cosine emission distribution, independent of the incident angle [9.31, 9.32]. Thus the emission process is modeled as a small planar surface source at each position with $n = 1$ in the $\cos^n\theta$-emitted angle distribution, and the distribution is centered around the surface normal. This implicitly assumes that there is appreciable residence time on the surface so that the

species lose their memory of their incident angle. It is also generally assumed that only the neutral species can be emitted from the surface. For ionic species the sticking coefficient is thus taken to be one, except when ion reflection effects are considered.

Those emitted species can then be deposited on other surfaces of the structure, leading to the *redeposited (emitted) flux*, $F_{\text{redep(emitted)}}$, term. So for each point i, there is a flux coming from each of the other points, termed k. It is again assumed that the number of gas-phase collisions near the surface area is negligible, so that the emitted species travel in straight lines to other points on the surface. Knowing the emitted flux distribution from each point k [Eq. (9.45) for k and a cosine distribution] and the geometrical relationship to point i, the flux emitted from each point k to point i can be determined. This is similar to what we did earlier in Section 9.2.2.1 when we calculated the flux and deposition rate onto a surface from a planar evaporation source. Of course, only those points that are in line of sight to point i are included. The redeposited flux at point i due to emission from point k is thus:

$$F^{ik}_{\text{redep(emitted)}} = g^{ik} * F^{k}_{\text{emitted}} = g^{ik} * (1 - S_c) * F^{k}_{\text{incident}} \tag{9.47}$$

Here g^{ik} is the geometrical factor that accounts for the distance and orientation between points i and k, similar to Eq. (9.23) where we derived the expression for deposition on a surface from a planar source. Then all the emitted fluxes from each of the points k to i are summed up and contribute to $F_{\text{redep(emitted)}}$.

We mentioned earlier that fast surface diffusion of deposited species is often used to explain good step coverage. However, the emission and redeposition process can result in the same effect. If the sticking coefficient is less than one, species landing on surfaces can be emitted and deposited elsewhere, including vertical surfaces (recall Figure 9–21). The diffuse emission behavior means that surfaces in all orientations can get material deposited on them. The lower the sticking coefficient, the more conformal step coverage becomes. Whether this redistribution of material and better coverage occur by the emission/redeposition process or the surface diffusion process is sometimes hard to distinguish although this has been done in some studies [9.32, 9.33, 9.34]. Using a clever overhang test structure, these experiments concluded that the emission-redeposition process, and not surface diffusion, was dominant for LPCVD and PECVD deposition of SiO_2. Thus a lower sticking coefficient, and not faster surface diffusion, is the appropriate model, at least in the specific cases studied. For other materials or techniques, surface diffusion may be important. For example, for sputter deposited Al, the sticking coefficient is found to be one [9.33], which implies no emission/redeposition, but surface diffusion at high temperatures is significant. In our simulations later we will show how both the sticking coefficient and surface diffusivity affect the degree of coverage.

The *sputtered flux*, $F_{\text{sputtered}}$, term is from the atoms that are sputtered off the wafer surface. This is a momentum-transfer process in which incoming particles knock and displace surface atoms from the surface of the wafer. This is due mainly to accelerated ions, both argon and ionized deposition atoms, coming directly from the gas phase. These ions can achieve energies of several hundred eV. The flux term at each point i due to sputtering (a negative term since atoms are being taken away from the surface) is therefore

$$F^{i}_{\text{sputtered}} = Y * (F^{i}_{\text{argon}} + F^{i}_{\text{direct(ions)}}) = Y * F^{i}_{\text{ions}} \tag{9.48}$$

Y is the sputtering yield and is defined as the ratio of atoms that are sputtered or knocked off the surface to the number of incident ions or atoms. Y is a function of the target material (the wafer surface material, in this case) and the mass of the incoming ion. However, we assume here for simplicity that it is the same for argon and other ions. It is also a function of ion energy and the incident angle, θ, at point i, with a normal incidence being 0°. Ions coming in at low angles are reflected from the surface and do not transfer significant energy to the surface atoms, while ions coming perpendicular to the surface knock surface atoms further into the film and not off. The maximum sputtering yield occurs usually between 40 and 70°.

A typical sputtering yield distribution can be seen in Figure 9–39. The specific characteristics of the curve, such as the maximum yield and angle for which it occurs, could vary for different materials and equipment conditions. Also shown is the normalized etch rate due to sputtering, which equals the product of the sputter yield and $\cos\theta$. The latter term is due to the fact that it is the component of the flux normal to the surface that determines the etch or deposition rate on a surface, as we have described. The sputter etch rate also has a maximum at some angle between 40 and 70°. This angular dependence on sputtering yield and sputter etch rate is often utilized to achieve more planar surfaces during deposition. The sloped surfaces (with incident angles less than 90°) sputter off faster than horizontal or vertical surfaces (with incident angles near 0 or 90°), leading to more planar surfaces.

The emitted angle distribution of the sputtered atoms has been modeled with different types of distributions [9.35]. The sputtered atoms often exhibit an ideal, diffuse angular distribution, being sputtered away at many angles centered around the surface normal, like desorbed or emitted species discussed above. This is illustrated in Figure 9–40(a). The distribution may in fact be more sideways, described by an "under-cosine" distribution (where n is less than 1 in a $\cos^n\theta$ distribution), shown in Figure 9–40(b). In some cases though they can exhibit a specular distribution, being sputtered at just a few angles near an angle opposite that of the incident angle. In this case the angular distribution of the sputtered atoms is described by a $\cos^n\theta$ distribution, with a very large

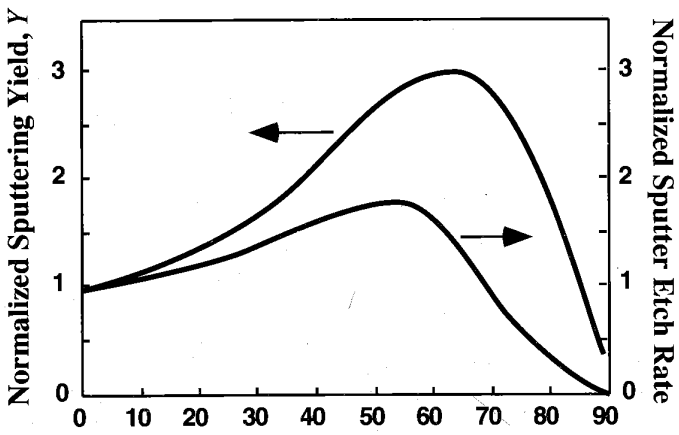
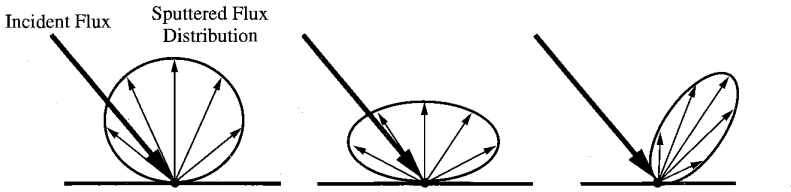


Figure 9–39 Typical example of sputtering yield and sputter etch rate as functions of incident angle, with respect to surface normal.



(a) Cosine ($n = 1$) emission or reflection ("perfectly diffuse")

(b) Under cosine ($n < 1$) reflection

(c) Specular reflection, with $n >> 1$ in cosine distribution, centered opposite the incident angle

Figure 9–40 Different emitted angle distributions for sputtered atoms.

value of n, and the distribution of sputtered atoms is centered opposite the incident angle. This is illustrated in Figure 9–40(c). Generally distributions like Figure 9–40(a) or (c) are used in simulations. The first type is common when the incident ion energy is high and there are collisions and interactions deeper in the sputtered layer, while the latter type is more common when the incident ion energy is lower and single surface collisions dominate. Also, amorphous materials like SiO_2 tend to sputter off with the first type of distribution.

As was the case with desorbed or emitted species, sputtered atoms can be redeposited elsewhere on the surface. The method to calculate this *redeposited (sputtered) flux* and the equations used are analogous to the redeposition due to emission. At each point i the redeposited flux coming from each other point k is determined:

$$F^{ik}_{\text{redep(sput)}} = g^{ik} * F^{k}_{\text{sputtered}} = g^{ik} * Y * F^{k}_{\text{ions}} \tag{9.49}$$

Again, g^{ik} is a geometrical factor that accounts for the distance and orientation between points i and k, as well as for the angular distribution of sputtered atoms from k. Only those points, k, that are in line of sight of i would contribute.

In addition to sputtering off atoms from the surface, ions striking the surface can also enhance the deposition rate. Ion bombardment can transfer energy to the species on the surface and enhance the different surface processes and reactions. One way to model this is by adding an ion-induced deposition rate enhancement term to the total flux expression (rather than trying to modify the other terms). This term is of the form of

$$F^{i}_{\text{ion-induced}} = K_i * F^{i}_{\text{ions}} \tag{9.50}$$

where K_i is the ion-induced deposition enhancement factor and is usually a fitting factor. Remember that F_{ions} includes all ions, both nonprecursor ions, such as Ar^+, and any ionized precursor species (Al^+ or Ti^+, for example). Non-ionized precursor species are still needed for this ion-enhanced process, but they are not a limiting factor in this process so only the ion flux is included in this term. The competition between ion sputtering and ion-enhanced deposition can be very important in obtaining the final topography.

9.5.1.2 Models for Different Types of Deposition Systems

To simulate deposition, one should include all the terms in Eq. (9.42) and calculate the total net flux at each point i on the surface. The deposition rate, or the movement of the surface, at each point is then calculated by dividing the total flux by the film density. For ease of use and computation, one could simplify the net flux equation for the various types of deposition methods (LPCVD, standard PVD, ionized PVD, etc.), including only the important terms for each technique. This is in fact what is done in SPEEDIE. Relative rate constants (K's) are often included in front of some of the flux terms which makes it easier to adjust the parameters when multiple processes are occurring. By examining the expressions that SPEEDIE uses for each type of deposition, one can gain a better understanding of the different deposition techniques.

First we consider *LPCVD deposition*. Figure 9–41 shows an LPCVD system and the relevant fluxes needed to model this type of deposition. In this case we have the direct flux of neutral species, as well as the desorbed or emitted flux and redeposition of these neutral species. There is no flux of ionic species in this system, and no sputtering or redeposition of sputtered species. In addition, long-range surface diffusion is usually not important, as discussed earlier.

Eq. (9.42) becomes

$$F^i_{\text{net}} = F^i_{\text{direct(neutrals)}} + F^i_{\text{redep(emitted)}} - F^i_{\text{emitted}} \tag{9.51}$$

Substituting in Eqs. (9.45) and (9.47) for the emitted and redeposited fluxes leads to

$$\begin{aligned} F^i_{\text{net}} &= F^i_{\text{direct(neutrals)}} + F^i_{\text{redep(emitted)}} - (1 - S_c)(F^i_{\text{direct(neutrals)}} + F^i_{\text{redep(emitted)}}) \\ &= S_c * (F^i_{\text{direct(neutrals)}} + F^i_{\text{redep(emitted)}}) \\ &= S_c * (F^i_{\text{direct(neutrals)}} + g^{ik} * (1 - S_c) * F^k) \end{aligned} \tag{9.52}$$

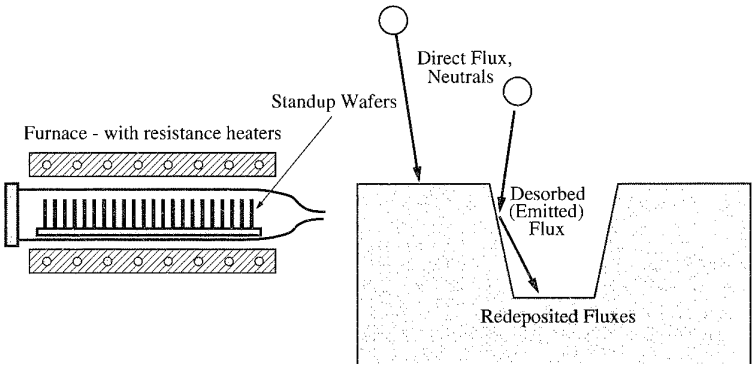


Figure 9–41 LPCVD system and relevant fluxes needed to model this type of deposition.

This must be solved for all points i simultaneously and consistently—emitted flux from one point affects the redeposition flux at another, which in turn affects the flux at the first point, and so on. In the deposition rate equation, the direct neutral and redeposited flux terms ($F^i_{\text{direct(neutrals)}} + F^i_{\text{redep(emitted)}}$) are combined into one term, and called the local deposition flux, F_d. The net deposition rate at each point is just the net flux, F^i_{net}, divided by the film density, N. So for LPCVD, the deposition rate equation is given as

$$\text{rate} = \frac{S_c F_d}{N} \tag{9.53}$$

The other important parameter would be the value of n in the $\cos^n\theta$ distribution for the incoming (direct) species.

Example Calculate the deposition rate using the expression for LPCVD deposition for SiO_2 on a flat surface with no topography. S_c is equal to 0.3, the maximum, unobstructed flux is equal to 3×10^{15} molecules $cm^{-2} sec^{-1}$, and the density of the deposited oxide film is 2.27 gm cm^{-3}.

Answer We are interested in this case in the deposition rate on a flat surface. Therefore there is no shadowing, no redeposited flux, and each surface position is in a horizontal orientation. Thus F_d is just equal everywhere to the maximum, unobstructed flux. The deposition rate for LPCVD in SPEEDIE is given by Eq. (9.53). The density of SiO_2 in terms of molecules cm^{-3} is

$$\begin{aligned}\text{density}(SiO_2) = N &= 2.27 \text{ gm cm}^{-3} * \frac{\text{Avogadro's number}}{\text{\# of gms per mole for } SiO_2} \\ &= 2.27 \text{ gm cm}^{-3} * \frac{6.02 \times 10^{23} \text{ molecules/mole}}{60 \text{ gms/mole}} \\ &= 2.3 \times 10^{22} \text{ molecules cm}^{-3}\end{aligned}$$

Therefore the deposition rate on the flat, unobstructed surface is

$$\text{rate} = \frac{S_c F_d}{N} = \frac{0.3 * 3 \times 10^{15} \text{ molecules cm}^{-2} \text{ sec}^{-1}}{2.3 \times 10^{22} \text{ molecules cm}^{-3}}$$

$$= 3.9 \times 10^{-8} \text{ cm sec}^{-1} = 0.023 \text{ } \mu\text{m min}^{-1}$$

For deposition on nonflat or obstructed surface locations, on sidewalls of trenches, for example, the local deposition rate can be less than this. The same expression is used but the local flux, F_d, can be less at these locations. Figure 9–46 to be discussed later illustrates LPCVD simulation on more complex topography.

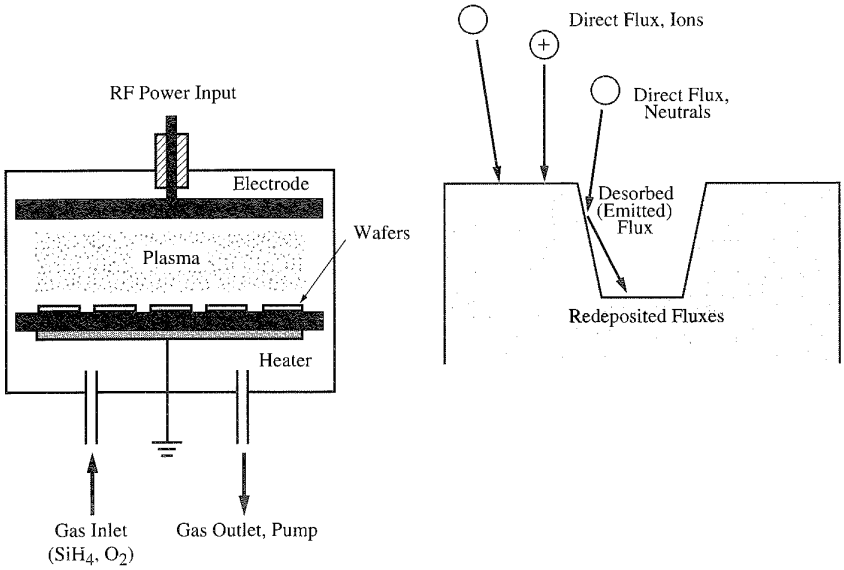


Figure 9–42 PECVD system and relevant fluxes needed to model this type of deposition.

For *PECVD deposition* there is an ion flux onto the wafer which can accelerate the deposition process. Figure 9–42 shows a PECVD system and the relevant fluxes needed to model this type of deposition. As in LPCVD there is a direct flux of neutral species and the emission and redeposition of those species. Also shown is the direct flux of ions, accelerated across the plasma sheath toward the wafer, which can enhance the deposition. The ion energy in this type of system is assumed low enough that sputtering is not significant and direct deposition of ions is also not appreciable. Again surface diffusion is not considered important.

Thus the ion-induced enhancement term, Eq. (9.50), is included in the rate equation, which becomes

$$\text{rate} = \frac{(S_c K_d F_d) + (K_i F_i)}{N} \tag{9.54}$$

K_d and K_i are relative rate constants for the non-ion-enhanced component and the ion-enhanced component, respectively. F_i is the net local ion flux, or F_{ion} as given above. N is the film density. Surface diffusion is not explicitly modeled here; however a low value of S_c can have a similar effect, as discussed earlier, and has been used in SPEEDIE in modeling PECVD oxide deposition [9.4, 9.5, 9.34, 9.36]. Both the non-ion-enhanced and the ion-enhanced terms are important in PECVD depositions [9.5, 9.36].

For *standard PVD*, the same fluxes and terms are used as for LPCVD. Figure 9–43 shows a PVD system, a standard DC sputtering system in this case, and the relevant fluxes needed to model PVD deposition. The rate equation is thus

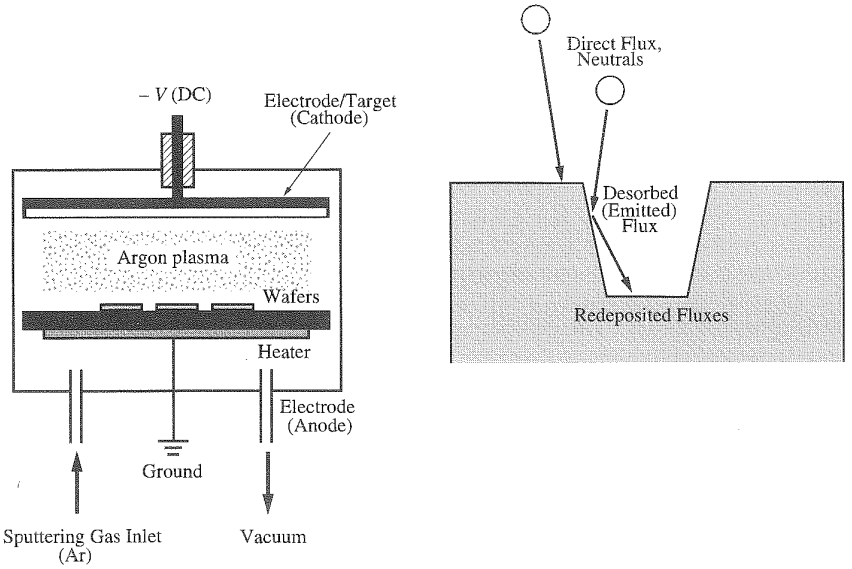


Figure 9–43 PVD (sputter deposition) system and relevant fluxes needed to model this type of deposition. The specific system illustrated here is a DC sputtering system.

$$\text{rate} = \frac{S_c F_d}{N} \tag{9.55}$$

However, different values for the sticking coefficient and the arrival angle distribution would be used for PVD versus CVD methods, as discussed below. There is no significant ion bombardment onto the wafer surface for standard sputter deposition systems, either DC or RF based, as described earlier. This model would also be applicable to collimated and long throw sputter deposition systems, but a narrower arrival angle distribution would be used than for standard sputtering systems. In addition, evaporation can be modeled by this model.

For *high-temperature PVD*, the relevant fluxes are the same as those illustrated in Figure 9–43, except that we now add the surface diffusion flux term, Eq. (9.44), to the rate equation for standard PVD, giving

$$\text{rate} = \frac{S_c F_d + \dfrac{D_s}{kT}\gamma_s \Omega \upsilon \dfrac{\partial^2 K}{\partial s^2}}{N} \tag{9.56}$$

For *ionized PVD*, there are several fluxes in addition to those in the standard PVD model. Figure 9–44 shows an ionized sputtering system and the relevant fluxes. In addition to the direct flux of neutral species, there is a significant flux of ionized species—both ionized Ar atoms and ionized precursor species (Al^+ or Ti^+, for example). These

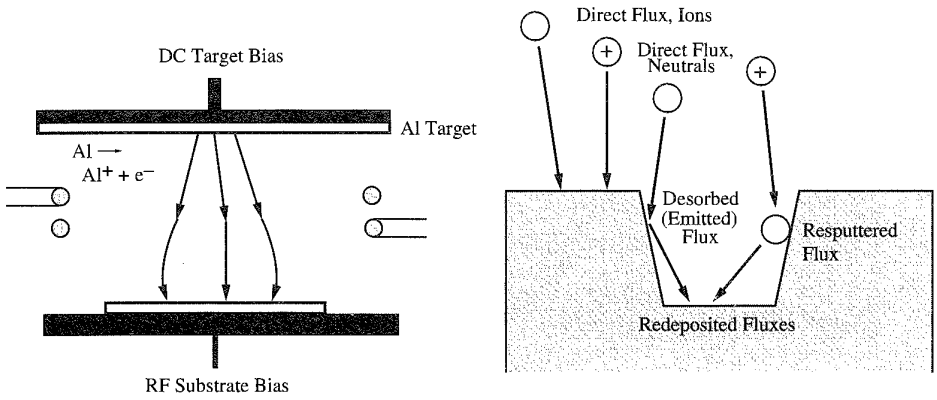


Figure 9–44 Ionized PVD system and relevant fluxes needed to model this type of deposition.

ionized species can be deposited as well as sputter surface atoms on the wafer. Therefore, for ionized PVD, several terms are added to the rate equation for PVD deposition: the term due to direct flux of precursor ions, $F^i_{\text{direct(ions)}}$; the ion-induced enhancement term [Eq. (9.50)]; and the sputtering and redeposition terms [Eqs. (9.48) and (9.49)]. The direct deposition of precursor ions and the ion-induced enhancement terms are combined into one term, equal simply to F_i, the total ion flux. This is done in SPEEDIE partly because SPEEDIE does not currently distinguish between Ar ions and ionized precursor species.

This leads to the expression given for the rate to be

$$\text{rate} = \frac{(S_c F_d) + F_i - (K_{\text{sp}} Y F_i) + (K_{\text{rd}} F_{\text{rd}})}{N} \tag{9.57}$$

Here F_{rd} is the flux given by Eq. (9.49) for redeposition of sputtered material. K_{SP} and K_{RD} are relative rate constants, used mostly as fitting parameters. A term for reflected Al ions may also be important for coverage issues and may be added in future models. (A model for bias sputtering would be similar to this, but the ions would be mostly Ar and not the deposited species.)

Similarly, for *high-density plasma (HDP) CVD* the sputtering and redeposition terms are added to the PECVD case, without any non-ion direct term. Figure 9–45 shows a HDP CVD system and the relevant fluxes used to model this deposition. This gives the following expression for the rate:

$$\text{rate} = \frac{(K_i F_i) - (K_{\text{sp}} Y F_i) + (K_{\text{rd}} F_{\text{rd}})}{N} \tag{9.58}$$

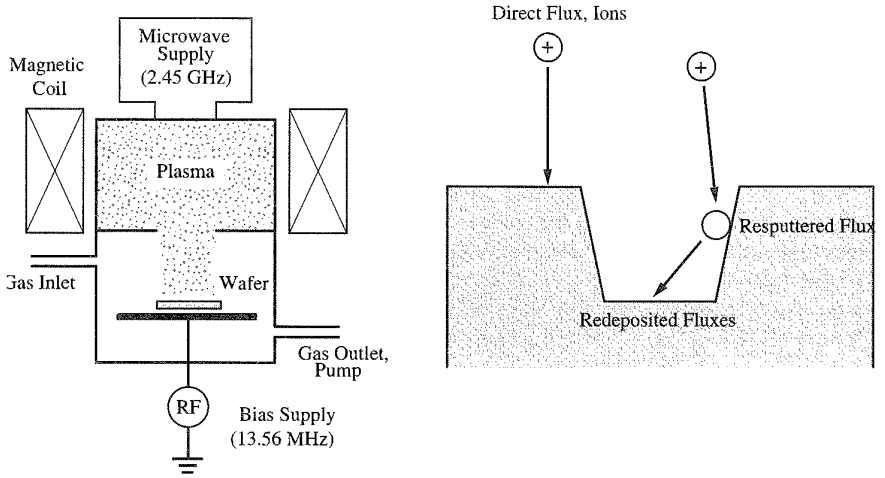


Figure 9–45 HDPCVD system (with a microwave ECR source) and relevant fluxes needed to model this type of deposition.

similar to that for ionized PVD. The ion flux would be quite high in these systems, about 1×10^{17} cm^{-2}. The non-ion (neutral) direct term is not included here because its flux is often too low at these low pressures to be important on its own, that is, when not enhanced by the ions. (However, more recent SPEEDIE models include other terms such as the neutral flux as secondary effects.) There could also be direct deposition of ionized precursor species; this is included in the ion-enhanced term for simplicity.

9.5.1.3 Comparing CVD and PVD and Typical Parameter Values

Before presenting some simulation results, we first want to make some general comments regarding some of the values used for CVD and PVD models. This will be done in terms of the standard models for low-pressure systems, that is, most CVD and PVD systems used today. The discussion will center around the parameter values summarized in Table 9–2.

In CVD deposition systems, the arrival angle distribution for neutral species is basically isotropic, while for PVD systems the arriving flux of neutrals is often more anisotropic and directed. This is due to the different types of sources and the different pressures.

In CVD systems the source gas fills the whole chamber and arrives at the wafers from many directions. This is partly due to injection systems and cages that may be used, as well as gas-phase reactions that may occur. More importantly, the precursor species in the gas phase in CVD systems undergo significantly more collisions than in PVD depositions before the species strike the wafers. This is due to the much higher gas pressure in CVD systems, even for low-pressure CVD systems, compared to PVD systems

Table 9–2 Typical values of modeling parameters for various PVD and CVD depositions

	n (exponent in cosine arrival angle distribution)	S_C (sticking coefficient)
Sputter deposition		
standard	~1.0–4	~1
ionized or collimated	8–80	~1
Evaporation	3–80	~1
LPCVD silicon dioxide		
silane	1	0.2–0.4
TEOS	1	0.05–0.1
LPCVD tungsten	1	0.01 or less
LPCVD polysilicon	1	0.001 or less

(HDPCVD systems with its very low gas pressure being the notable exception). More collisions result in a shorter mean free path and a wider arrival angle distribution. As a result, the arrival angle distribution can be assumed to be isotropic in most CVD systems. This means a value of 1 in the $\cos^n\theta$ arrival angle distribution given by Eq. (9.43) for the direct flux of deposition species, as shown in Table 9–2. (Higher values for n would be used for the ionic species in plasma CVD systems, where the ions are accelerated and directed toward the wafer surface.)

For PVD depositions, the depositing material comes off a remotely located and finite area source or target, and due to geometrical considerations often arrives at the wafers in a more directed manner. In addition, fewer collisions occur during the transport, as mentioned above, because of the much lower pressured used—usually in the 1- to 10-mtorr range for sputtering and even lower for evaporation, as compared to 0.25 to 2.0 torr for LPCVD. As a result of all this, the arrival angle distribution for PVD depositions can be less isotropic and the atoms arriving at the wafer surface can therefore be more directed. For standard sputter PVD the value of n in the $\cos^n\theta$ arrival distribution will often be greater than 1, representing a more directed distribution arriving from above. The exact value will depend on the system geometry (e.g., target size, target-to-wafer distance, rotation of substrate), the operating conditions (e.g., pressure) and the depositing material. Typical values range from very close to 1 (virtually isotropic) to about 4 (fairly directed), as shown in Table 9–2, with higher pressures, larger target sizes, and smaller target-to-wafer distances giving the lower n values.

For ionized species (most of the depositing species in ionized PVD systems and some of them in a plasma CVD system) as well as neutral species in collimated PVD systems, the species are very directed. The n value can be from about 8 to 80 or even higher, depending on the system and the plasma conditions.

The sticking coefficient is also generally different for CVD and PVD systems. For CVD systems, the sticking coefficients used in modeling are almost always less than one, as shown in Table 9–2. Typical values would be 0.05 to 0.1 for TEOS-based CVD of SiO_2, and somewhat higher—up to about 0.4—for silane-based CVD. Similar values would apply to the neutral component in PECVD systems. The value for CVD W (via

hydrogen reduction) and polysilicon is even lower, 0.01 or less. Surface reactions are important for CVD depositions. The species can desorb before the reactions required for deposition can take place. Another precursor required for the reaction may not be present at that spot at that time. In addition, competing reactions and site occupation by other species and byproducts can occur. These all contribute to the effective sticking coefficient for CVD depositions being low. The species not reacted and deposited are desorbed, or emitted, and are usually assumed to have a diffuse emitted angle distribution, as opposed to a specular emission.

For conventional PVD, where physical processes dominate, the atoms generally attach and then stay attached, without desorbing. No surface reactions occur other than simple condensation. The sticking coefficient is therefore usually assumed to equal 1, or close to it, for PVD depositions. (One exception to this may be PVD of Ta, which seems to have a somewhat lower sticking coefficient.) For the ionized species in either PVD or CVD systems, the sticking coefficient is often assumed to equal one.

In PVD and CVD systems where large fluxes of ions are present, such as in ionized PVD and high-density plasma CVD, sputtering of deposited material can occur. This, like desorption, leads to redeposition of species elsewhere. The angular distribution for sputtered atoms is usually assumed to be perfectly diffuse, with n equal to 1, but is sometimes assumed to be specular. For example, sputtered Al is often modeled with specular redeposition with n equal to 2 or more for a cosine distribution that is directed away from the incident angle. HDPCVD deposition is dominated by the highly directional ion flux and actually acts more like PVD deposition rather than other CVD systems in regards to the sticking coefficient (which is effectively equal to one since no neutral species are considered) and arrival angle distribution. It is its angle-dependent sputtering and redeposition action as well as its very directed ion deposition, at the very low pressures, that give HDPCVD its good filling and planarization abilities.

We did not discuss models for evaporation deposition. However, by using a very directed angular arrival distribution (n equal to 15 or higher in the $\cos^n\theta$ distribution) for a stationary substrate or a less directed distribution for a rotated substrate, and using a sticking coefficient equal to one, one could simulate the evaporation deposition processes using the equations given above for conventional PVD.

Based on these generalizations we can now understand why CVD depositions usually give more conformal coverage than PVD depositions in these low pressure systems. An isotropic arrival distribution may lead to better step coverage. More importantly, a lower sticking coefficient, allowing more desorption and redeposition occurrences, leads to better step coverage, and a lower sticking coefficient is usually more important than the arrival angle in obtaining good coverage. Whatever the arrival angle distribution, a low sticking coefficient can lead to good coverage (even in a deep hole) since the deposition species, wherever they land, will bounce around and deposit on all sides of a structure. Thus CVD generally gives much more conformal coverage than PVD.

Having said that CVD techniques generally deposit films with better step coverage, it should be reiterated that a low sticking coeffecient and good step coverage are not always required. This is the case in applications like covering the bottoms of deep, narrow contact or via holes with materials like Ti or filling narrow spaces with SiO_2 or metal. While a low sticking coefficient can greatly help in filling or lining holes, such as with

CVD W or TiN, there is another approach. The idea behind collimated and ionized PVD techniques, and even in HDPCVD with its strong ion component, is to deposit material from the gas phase in a very vertical direction. In this case a narrow arrival angle distribution is desired, not a wide distribution associated with good step coverage, and a low sticking coefficient is not required. Step coverage per se may not be great, but the bottoms of vias will be covered. In addition, good filling of holes or gaps can be achieved by the use of directed ions and sputtering action—filling up the holes from the bottom and sputtering away any overhang that may develop. In addition, the resputtering may help with sidewall covering. Directed ions and sputtering are the keys to ionized PVD and HDPCVD. Of course, CVD techniques with very low sticking coefficients are also good at covering the bottoms of deep vias and filling holes and gaps. But viable CVD techniques have not been developed for some materials, and these other techniques have been shown to do a good job at achieving bottom coverage and gap filling.

9.5.2 Simulations of Deposition Using a Physically Based Simulator, SPEEDIE

We now describe a series of deposition simulations using the SPEEDIE simulator and the equations and ideas presented above. These simulations will be used to demonstrate and further explain some of the concepts covered earlier in this chapter. The main focus will be on issues of film coverage and filling the underlying topography.

The first example is a simple simulation of LPCVD deposition of SiO_2 following Eq. (9.53). We will use parameter values in some cases that may not be realistic for LPCVD deposition but are more appropriate for PVD deposition. We do this to illustrate some of the important concepts of the models.

The initial topography is a wide trench—3 μm wide and 1 μm deep—in underlying Si (Figure 9–46). 0.5-μm-thick SiO_2 layer is deposited over this topography. A $\cos^n\theta$ distribution for the arrival angle distribution [Eq. (9.43)] is used, with $n = 1$ in the first simulation. This corresponds to a uniform or isotropic distribution of arrival angles, that is, molecules coming from all directions equally. This would be the case when many collisions occur in the gas phase so that the mean free path is very short and the direction of molecules becomes randomized. As we mentioned above, an n value of 1 is used for most CVD depositions. The other important parameter is the sticking coefficient, S_c, and a value of 1 is used here. This means that there is no reemission or redeposition of depositing species once they strike the surface. This is unrealistic for most CVD processes, as discussed earlier, but we will use it here for illustrative purposes. Later we will use lower values for S_c.

Figure 9–46(a) shows the SPEEDIE simulation of this simple LPCVD simulation. The different contour lines of the deposited SiO_2 represent different time intervals, which makes the deposition process easier to follow. It is apparent that the thickness is reduced on the sidewalls and corners by about 50% or more, even for an isotropic arrival angle distribution. This is because the species are arriving from one side only to each wall (a 90° maximum view angle on the side versus 180° on top). Also the sidewall deposition is sloped inward, which could cause problems for the deposition of subsequent layers.

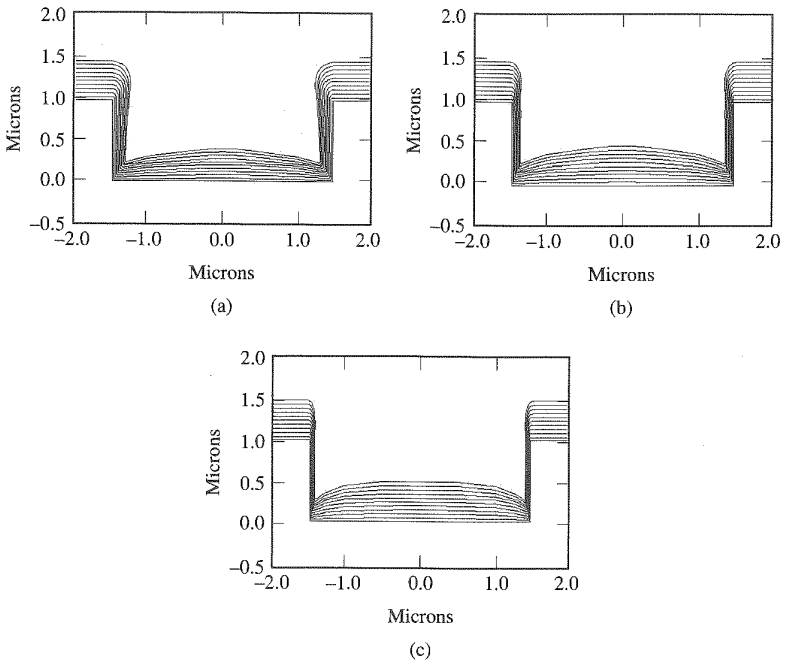


Figure 9–46 SPEEDIE simulations for LPCVD deposition of SiO_2 with $S_c = 1$ (which is more typical of PVD than LPCVD) and varying values of n, the arrival angle distribution factor: (a) $n = 1$; (b) $n = 3$; (c) $n = 10$. Worse step coverage results as n increases (the arrival angle distribution narrows).

The next simulation, shown in Figure 9–46(b) is the same as the previous one, except that the angular distribution of arrival atoms is changed. The value of n used in the $\cos^n\theta$ distribution is 3 here, compared to 1 in the previous simulation. This means that the arrival angle distribution is narrower, with more material coming in closer to the vertical direction. This might be the case when the gas pressure is lower and fewer gas-phase collisions occur, resulting in a longer mean free path. There is even less material deposited on the side walls and corners, as expected. Figure 9–46(c) shows a similar simulation, except with $n = 10$. (Again, this would be unrealistic for CVD processes, but is still illustrative since the same basic model is used for PVD where n can vary greatly—from close to 1 to very large values.) Here the sidewalls receive much less deposition. If the material deposited were a metal conductor, one can see that a narrow arrival distribution could cause problems with conduction of the interconnect line when it travels over steep topography. From these first three simulations we demonstrate the importance of the arrival angle distribution.

In the next example a narrower trench is utilized in the underlying layer (Figure 9–47). This corresponds to a higher aspect ratio (about 1.25 in this example compared to about 0.33 in the previous one). The same model (LPCVD) and parameters are used as

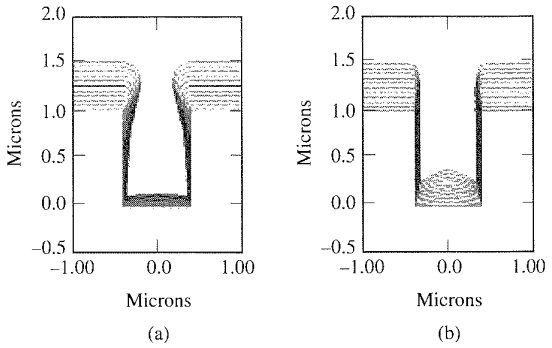


Figure 9–47 SPEEDIE simulations for LPCVD deposition of SiO_2 with $S_c = 1$ (again, more typical of PVD than CVD) and varying values of n, the arrival angle distribution factor, but for a narrower trench: (a) $n = 1$; (b) $n = 10$. Less deposition on sidewalls again results as n increases (the arrival angle distribution narrows), but better bottom filling is achieved due to more vertically directed deposition and less overhanging and shadowing at top corners.

in the previous simulations. Figure 9–47(a) shows the simulation results for the case of $n = 1$, a wide arrival distribution. Here the sidewalls and bottom of the trench have much less SiO_2 deposited on them as compared to Figure 9–46(a), which also had $n = 1$. Clearly geometry and shadowing effects came into play here. Those molecules arriving at large angles to the vertical cannot reach the inside of the trench. Instead they are blocked by the top corners of the trench. Another way to look at it is in terms of the view angle: For locations near the top of the trench the view angle is much larger than for points at the bottom, and more deposition occurs at the top corner and less at the bottom. This becomes even worse as the corners build up due to further deposition, leading to large overhangs. Eventually the top part will seal off, leaving a void in the trench. This angular-dependent blocking effect is called shadowing, for obvious reasons.

In Figure 9–47(b) the value of n is changed to 10. With the more directed incoming molecules, there is less deposition on the top part of the sidewalls. However, there is much more deposition on the bottom of the trench compared to when $n = 1$ [Figure 9–47(a)]. This is because there are more molecules coming vertically into the trench which are not blocked by the top corners or overhangs. Eventually the whole trench could be filled without the development of overhangs or voids.

We see from these examples where the sticking coefficient is unity, and therefore no reemission and redeposition occur, that geometry and line-of-sight considerations are very important. This would be the case for PVD depositions, where the sticking coefficients are usually assumed to be close to one.

We now change the value of the sticking coefficient, S_c, to see its effect. In the previous simulations it was set to 1. Next we set it to 0.1. This means that only one-tenth of the deposition species stay at each arrival point. The rest are reemitted and can be deposited elsewhere. In fact, each redeposited species also has a sticking coefficient

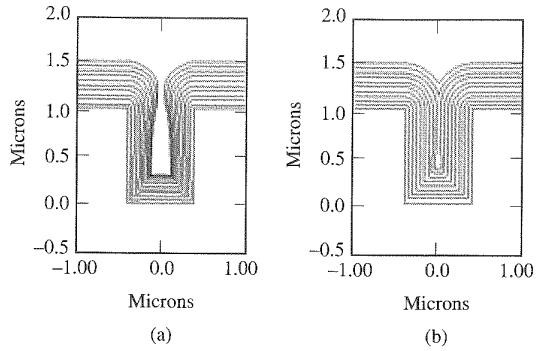


Figure 9–48 SPEEDIE simulations for LPCVD deposition of SiO_2 in a narrow trench with the same isotropic arrival angle distribution (n = 1) but with different values of the sticking coefficient: (a) $S_c = 0.1$, and (b) $S_c = 0.01$. Lowering the sticking coefficient from 1 to 0.1 results in much more conformal coverage. (a) has model parameters typical of CVD SiO_2 depositions, while (b) is typical of CVD W depositions.

of 0.1, so the effect is compounded. The net result is that the species bounce around a lot within the topographical features, giving much improved coverage.

Figure 9–48(a) shows the results for $S_c = 0.1$ and $n = 1$. In this case, these would be realistic parameters for LPCVD SiO_2. Comparing this with Figure 9–47(a), we see that much more uniform deposition occurs, with much more material deposited on both the sidewalls and the bottom. In fact better overall coverage occurs here than in any of the previous simulations (where S_c was set to one in all of them and n was varied), indicating that the sticking coefficient has a larger impact on the profile than the arrival angle distribution. Lowering the sticking coefficient improves both the step or sidewall coverage and the filling ability, and a low sticking coefficient can compensate for orientation and shadowing effects to a very large degree. Figure 9–48(b) shows the simulation results for an even lower value of $S_c = 0.01$, such as would be typical of W or polysilicon depositions, again with $n = 1$. Even more uniform coverage and better filling result.

Figure 9–49 shows the results of some SPEEDIE simulations in which the angle of the sidewall is varied. For all three simulations shown, a sticking coefficient of 0.2 and an n value of 1 are used. One can see that better filling is obtained as the slope is changed from completely vertical (90°) to slightly sloped. Contact and via holes are often sloped slightly to enhance filling. Sloping a step also helps sidewall step coverage.

While the above series of simulations utilized the LPCVD model in SPEEDIE, the same results and trends would occur with the standard PVD model since the same equations are used. Only the values used for the sticking coefficient and the arrival angle distribution would be different. As discussed earlier, CVD processes have sticking coefficients in the 0.001 to 0.3 range and an n value of 1. PVD processes generally have sticking coefficients close to 1, while n is often greater than 1. The above simulations help illustrate the difference between CVD and PVD by varying those parameter values.

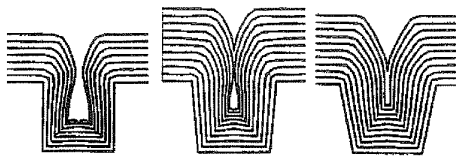


Figure 9–49 Results of SPEEDIE LPCVD simulations in which the sidewall angle is changed. Decreasing the angle from 90 to 80°C greatly improves the trench filling. (After [9.37])

We next compare the methods of LPCVD and HDPCVD. In the latter case there is a strong ion component that not only leads to good filling of holes (since the effective n value of the species would be high) but also causes angle-dependent sputtering of the surface. This causes sloped surfaces to etch away faster than flat surfaces, which can lead to more planar surfaces when depositing material over nonplanar underlying topography. We shall now demonstrate this with SPEEDIE simulations. We first show a simulation using the LPCVD model. The initial topography is a small, flat rectangle on a flat surface, such as would be the case for a polysilicon or metal interconnect line on a flat underlayer (Figure 9–50). We now deposit LPCVD SiO_2 on top. A cosine arrival distribution with $n = 1$ is used, and a sticking coefficient of 0.1 is specified, typical for LPCVD SiO_2 using TEOS. The simulated results are shown in Figure 9–50(a). One sees the very conformal coverage due to the isotropic arrival and, more importantly, the low sticking coefficient [compare this to Figure 9–46(a), which has isotropic arrival but $S_c = 1$]. Conformal coverage is good for obtaining a uniform thickness everywhere, but bad for planarizing the structure.

Next we simulate the same situation, but with the HDP model and typical parameters for that method. A very large and very directed ion flux with a high n value in the arrival angle distribution is used, typical for HDP systems. Sputtering and redeposition of material occur, with a maximum sputtering yield at 62° relative to the surface. This causes sloped surfaces, such as at corners, to be sputtered faster than vertical or horizontal surfaces. The simulation results are shown in Figure 9–50(b). One sees the characteristic pyramid shape that starts to form due to these effects, which then gets reduced to a small bump on the surface. The final topography is more planar than the final topography in the standard LPCVD case. For wider features in the topography, wider and taller pyramidal shapes would form and a thicker HDP oxide layer would have to be deposited for planar topography to be achieved. For this reason, other planarizing techniques, such as CMP, are often used in conjunction with HDP depositions to completely planarize the topography over all the features.

The ability of HDP deposition to fill spaces and to avoid the formation of voids, usually the main reasons for using HDPCVD, are demonstrated in the next two simulations. First in Figure 9–50(c) we show LPCVD deposition in a trench, similar to that described earlier in Figure 9–48(a). The sticking coefficient is still 0.1 and the angular distribution factor, n, is still 1, but the maximum deposition flux is increased a little more to try to fill the space. One sees that a void is formed in the narrow trench, even with a fairly small sticking coefficient. While low sticking coefficients of CVD processes help with filling, the shadowing effects are still important and voids can still form for high-aspect

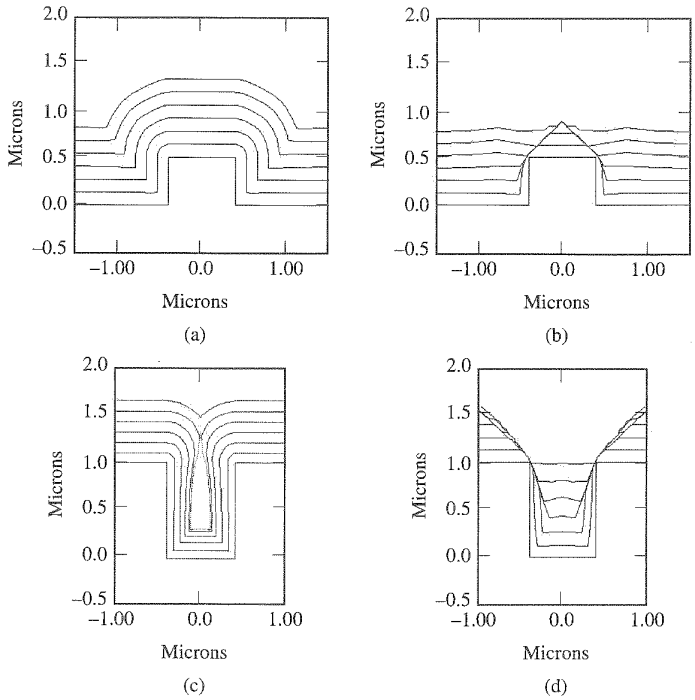


Figure 9–50 SPEEDIE simulations comparing LPCVD and HDPCVD depositions. (a) LPCVD deposition of SiO_2 over rectangular line. (b) HDPCVD deposition, with directed ionic flux and angle-dependent sputtering, over rectangular line showing much more planar topography. (c) LPCVD deposition in trench, showing void formation. (d) HDPCVD deposition in trench, showing much better filling.

ratio structures. The use of standard PECVD oxide deposition can help with the filling of gaps and spaces. With its directional ion component [Eq. (9.54)] which can be approximately equal in magnitude to the purely chemical component for oxide deposition [9.34], a little better filling of the gap occurs as compared to LPCVD deposition. But a void still results for this aspect ratio gap (though slightly higher up in the gap) and gets worse as the aspect ratio increases (see Figure 9–3). Using TEOS/O_3 films or the dep/etch/dep method can also improve the filling.

In Figure 9–50(d) we show how HDP deposition completely fills the gap, doing the same simulation with the trench structure but with the HDP model. No voids are formed. This is due to a combination of the highly directed ion flux component of deposition and the angle-dependent sputtering action. The deposition is very directed, and any overhang that does develop on the top corners is sputtered away, preventing any shadowing due to an overhang. Redeposition of sputtered material also helps with cov-

erage and preventing voids. This is obviously a very attractive technique for filling narrow spaces with dielectric material. Simulations show that void-free depositions can result even for aspect ratios of 3:1 and higher [9.4]. Actual SEM images of HDP deposition of SiO_2 were shown earlier in Figure 9–34 illustrating the filling and planarizing abilities of HDP deposition.

We next describe some PVD simulations. The standard PVD deposition SPEEDIE model uses Eq. (9.55). This is identical to that used for LPCVD, but usually different values for the sticking coefficient and the arrival angle distribution are used, as discussed earlier. In the first simulation we deposit Al onto a very narrow SiO_2 via structure with an aspect ratio of about 3, indicative of more advanced technology. We use parameter values more typical of PVD (sputter deposition) here. For S_c, we use 1. For the arrival angle distribution, we assume a $\cos^n\theta$ distribution with $n = 4$ (at the high end for standard sputter depositions), giving a more directed incoming flux than LPCVD systems. In this case, we specify that the hole is a via (a cylinder) rather than a trench, as was the case in the LPCVD simulations shown earlier. SPEEDIE calculates slightly different normalized direct fluxes because of this, and a via is generally harder to fill than a trench. The simulation results are shown in Figure 9–51(a). Very little Al is deposited inside the narrow via. Even though the flux is somewhat more directed than in LPCVD systems, the shadowing effect and the absence of any reemission (or resputtering) and redepositing of material allow very little material to be deposited in and on the bottom of such high-aspect ratio vias.

We next simulate ionized PVD, or IPVD, of Al. We assume that 80% of the Al atoms being deposited are ionized. For the un-ionized atoms, we assume a $\cos^n\theta$ distribution with $n = 4$. But for the ionized Al, we use a $\cos^n\theta$ distribution with n equal to 80. Eq. (9.57) is used for the local deposition rate, similar to the HDPCVD case [Eq. (9.58)]. We use typical parameter values for PVD, such as a sticking coefficient of 1. The same sputtering yield dependence is used as was used in the HDPCVD simulation, that is, with a

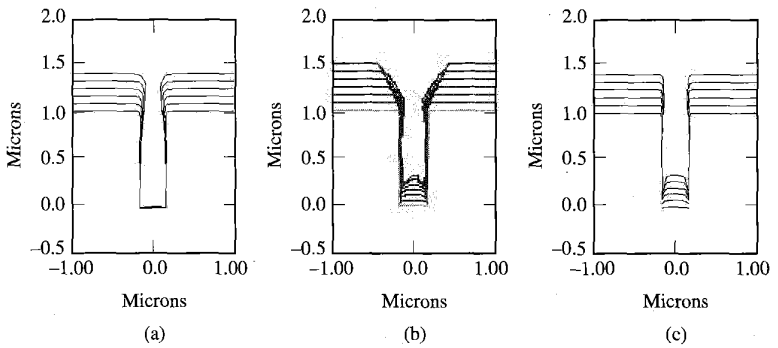


Figure 9–51 SPEEDIE simulations for sputtering (PVD) and ionized sputtering (IPVD) deposition of Al. (a) conventional PVD deposition with $S_c = 1$ and $n = 4$. (b) IPVD deposition with 80% of Al atoms being ionized (with near vertical arrival angles) and angle-dependent sputtering yield, showing much better filling of narrow via. (c) IPVD with no resputtering.

maximum sputtering yield at 62°. The simulation is shown in Figure 9–51(b). Two differences from the standard PVD simulation are evident. First, the bottom of the via gets filled much more than the standard PVD case. This is due to the very narrow (and vertical) ion angle distribution and is the reason why ionized PVD is an attractive alternative to standard PVD for some applications, such as for Ti or TiN deposition in today's deep contacts and vias. The second difference is the sloped sides at the top corners. This is due to the resputtering action, which not only bevels the corners but also redeposits material elsewhere. Resputtering from the bottom also occurs.

To see the effect of the resputtering and redeposition, another simulation is done. Everything is the same as in the first IPVD case except that the sputtering is turned off. The results are shown in Figure 9–51(c). With the very directed Al ion flux, a good amount of Al gets deposited down into the hole. But without the sputtering, the top corners are no longer sloped. The shadowing from the top corners and the lack of redeposition result in virtually no coverage on the sides of the via. Coverage on the side of a via or trench is often important, for example when depositing a TiN liner in a via, and the sputtering and redeposition action in IPVD is seen to help with this, especially at higher ion energies [9.14, 9.35].

In the final SPEEDIE simulations, High-Temperature PVD (or HTPVD) of Al in a trench is simulated. Eq. (9.56) is used in the HTPVD model. This includes the flux term that is due to surface diffusion driven by surface energy reduction, causing reflow during the deposition. For comparison we first show a standard PVD deposition into a trench, equivalent to setting the surface diffusivity in the HTPVD model to zero, in Figure 9–52(a). Next we do a HTPVD deposition with the deposition temperature specified to be 400°C, which determines the surface diffusion coefficient. The surface diffusivity and energy are taken from [9.38]. The resulting SPEEDIE simulated profile is shown in Figure 9–52(b).

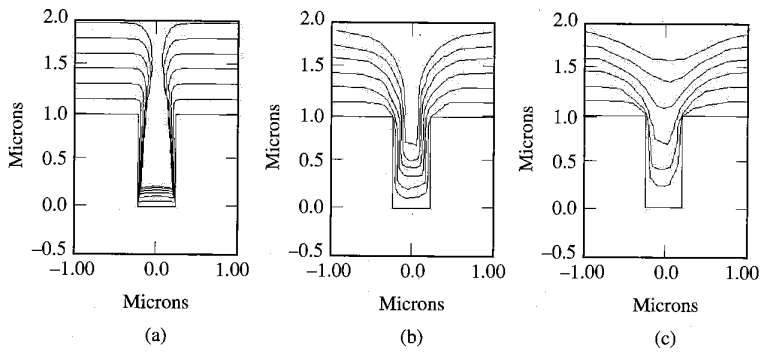


Figure 9–52 SPEEDIE simulations for High-Temperature PVD (HTPVD). (a) Standard PVD deposition with $S_c = 1$, $n = 4$, for comparison, showing poor filling. (b) HTPVD with surface diffusion at a deposition temperature of 400°C. Better filling is seen with the HTPVD. (c) HTPVD again but at a higher deposition temperature of 550°C, showing even more reflow during deposition than in (b), and even better filling and a smoother topography.

One can see that with the surface diffusion, much better filling of the hole and better step or sidewall coverage occur. Next we raise the deposition temperature even more to 550°C, resulting in a higher surface diffusivity, and the results are shown in Figure 9–52(c). The hole is completely filled, and much better planarization of the metal layer results. One can see from these simulations the advantages of high-temperature PVD deposition. More simulations involving surface diffusion driven reflow will be presented in Chapter 11. In that case, reflow of the metal after deposition will be simulated.

9.5.3 Other Deposition Simulations

We now describe some simulations done with a commercially available simulation tool. Here we use Silvaco's ATHENA simulator (using the ELITE topography tool) to simulate deposition of CVD SiO_2 and sputtered Al. While most of Silvaco's front-end simulation models for diffusion, oxidation, and so forth, are directly based on SUPREM IV models, the deposition and etch models used here and in the next chapter from ATHENA (many based on models originally in the SAMPLE simulator [9.39]) are more empirical in nature than SPEEDIE's models, and more user friendly. For example, instead of the user specifying the various fluxes, the user specifies the deposition rate and various parameters to obtain the shape and type of deposition profile.

First the ATHENA CVD deposition model is illustrated. Here the deposition rate, time, and step coverage are the input parameters. In the first simulation where we deposit oxide into a trench, the step coverage parameter is designated to be 1. This means that the same thickness is deposited everywhere. The resulting profile is shown in Figure 9–53(a). Next the step coverage is set to 0.2 instead of 1. The deposition rate, R, for the CVD model is given by:

$$R = dep.rate((1 - step.cov)\cos\theta + step.cov) \tag{9.59}$$

where θ is the angle between the surface segment and the horizontal. For a *step.cov* term less than 1, this results in the full deposition rate on horizontal surfaces (where $\theta = 0$),

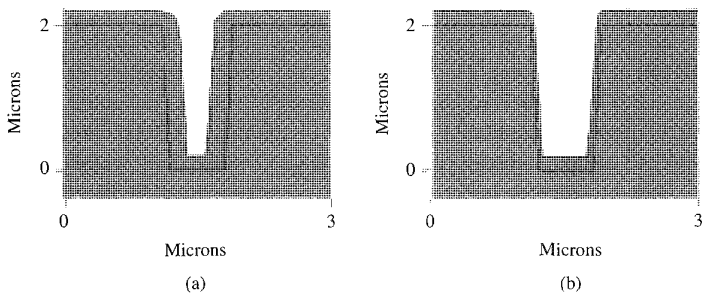


Figure 9–53 Silvaco's ATHENA [9.40] topography simulation of SiO_2 CVD deposition in a via for different values of the step coverage parameter, step.cov, selected by the user: (a) step coverage parameter = 1; (b) step coverage parameter = 0.2.

but less deposition on nonhorizontal surfaces. The result for the step coverage parameter equaling 0.2 is shown in Figure 9–53(b), exhibiting nonconformal coverage. Similar trends in step coverage using SPEEDIE were obtained by varying the value of the sticking coefficient in Figures 9–47(b) and 9–48(b).

The next example simulates the deposition of sputtered or evaporated Al using one of the Silvaco PVD models, the hemispheric deposition model. Here the incoming material flux arrives within a range of directions, coming from an extended, but limited, source such as in a planar sputtering system. The two angles designated by the user are the limits within which the flux is uniformly distributed. In this first case, the flux arrives at angles from 10 to -10 degrees from the normal to the substrate. The growth rate at each point is calculated after [9.41] by

$$R = C(\cos\omega_1 - \cos\omega_2)i + C(\sin\omega_1 - \sin\omega_2)j \tag{9.60}$$

where C is the deposition rate of an unshadowed surface normal to the flux direction, ω_1 and ω_2 are the upper and lower bounds of angles, and i and j are the unit vectors in the x and y directions, respectively. Figure 9–54(a) shows the simulation for such a deposition. One can see the directed deposition with most of the film being deposited on the horizontal surfaces. This is different from the ATHENA CVD model with a low step coverage value shown in the previous example in that now shadowing, overhangs, and cusping are observed on the corners. Figure 9–54(b) shows the result when the upper and lower angles are 45 and -45 degrees. More uniform deposition is observed, with more film deposited on the sidewalls, and less on the bottom of the trench. This model is similar to the SPEEDIE PVD model in which the distribution of arrival angles is controlled by the n term in the $\cos^n\theta$ distribution. However, unlike the SPEEDIE model, the sticking coefficient cannot be specified. The upper and lower bounds of angles used will depend on machine specifications, such as source and wafer geometry and characteristics, and operating pressure.

With this model one could also simulate asymmetric depositions such as might occur at the edge of a wafer in an evaporation system with a small source or in a sputtering

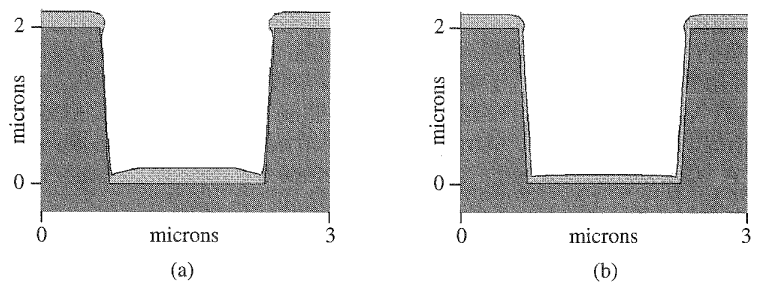


Figure 9–54 Silvaco's ATHENA [9.40] topography simulation of Al PVD deposition in a wide trench using the hemisphere model for different ranges of arrival angles: (a) the arrival angles are from -10 to $+10°$; (b) the arrival angles are from -45 to $+45°$. The wider angle range results in better step coverage, similar to a low n value in SPEEDIE's arrival angle distribution function.

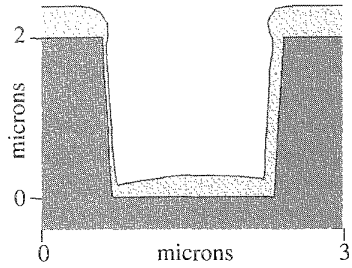


Figure 9–55 Silvaco's ATHENA [9.40] topography simulation of Al PVD deposition in a wide trench using the hemisphere model for asymmetric arrival angles: −25 to +45°. Asymmetric deposition is seen, as might be the case for the right side edge of a wafer in a PVD system with a small source or target.

system with a relatively small target, such as in a long-throw collimation system. In that case the absolute value of the two angles in the hemispheric deposition model could be different. Figure 9–55 shows an ATHENA simulation in which the upper angle is 45° and the lower angle is −25°. Asymmetric deposition is clearly seen in the profile.

The advantages of these models are their ease of use, as well as the ease of calibrating them. For a certain piece of deposition equipment, a given material to be deposited, and specific deposition conditions, one matches the profiles of some test structures to the simulations by varying the model parameters. One can then build up a library of deposition equipment, materials, and normal deposition conditions to use in simulations. Model parameter libraries of various commonly used pieces of deposition and etch equipment with model parameters for various films and normal operating conditions are often included with these simulators [9.40]. One can easily use these models to optimize the deposition process without worrying about all the details of the process.

The advantage of the SPEEDIE-type models is that they are more physically correct and the parameters are more closely linked to actual processes and mechanisms. One inputs the actual flux information in terms of flux rate, type, angular distribution. In addition one specifies physical properties such as sticking coefficients, surface diffusivities, sputtering and reemission coefficients, and the like. The models then calculate the combined effects of the various processes and phenomena in a 2D/3D structure. Obviously, these types of models are more apt to be valid for conditions outside those for which they are calibrated and tested. It should be noted that Silvaco and other commercial simulation tool vendors such as Avant! [9.42] also now offer other deposition and etch models which are less empirical and more physically based, such as those using Monte Carlo calculations. Avant!, for example, in their commercially available TAURUS-TOPOGRAPHY deposition and etch simulator, offers both the simpler, more empirical models as well as models and algorithms similar to those in SPEEDIE involving actual direct and indirect flux calculations. Silvaco has developed such models as well.

Models have also been developed for deposition processes that take into account the specific chemical reactions in CVD systems and the mechanisms for how ion bombardment enhances the deposition rate in plasma systems (see [9.43] to see this how all this has recently been done for HDPCVD of SiO_2, which considers 167 gas-phase reactions and 96 surface reactions!). These, or parts of these models, are sometimes incorporated into physically based simulators to make them even more complete.

9.6 Limits and Future Trends in Technologies and Models

A major problem that process developers face as the technology advances is the ever-increasing aspect ratios of the features. Depositing into smaller, deeper holes and spaces will only get more difficult. The NTRS roadmap [9.44] projects that the aspect ratios of metal lines, the spaces between the lines, and the via and contacts holes will all continue to increase over the next generations. (This is primarily due to the metal thicknesses not shrinking as fast as the lateral dimensions.) This means that the spaces and holes will get smaller and relatively deeper. We saw in this chapter how good coverage and filling are more difficult as the aspect ratio increases. High-density plasma deposition systems have proven successful in filling spaces between metal lines with oxide, and continued use and development of these types of systems are likely. New processes such as dual damascene described in Chapter 11, could place less importance on filling spaces with dielectrics in the future. Most metal films are still being deposited by sputtering techniques. The trend has been to try to use CVD because better filling and conformal step coverage are often achieved by CVD. However there are still problems with CVD deposition of metals which must be solved. These include contamination (impurities, such as carbon, from the precursors mostly) and difficulty with alloys (Al with Cu, for example). A PECVD process for Ti using $TiCl_4$ for filling small holes has recently been announced by one of the equipment manufacturers. New sputter techniques such as ionized, collimated, and hot-sputter depositions have been developed for metal deposition in high aspect ratio features. Future processes may use electroplating and spin-on methods for metals and dielectrics respectively. A potentially new type of deposition involves layer by layer adsorption and reaction of reactant species on the surface, utilizing sequentially pulsed precursors and self-limiting reactions for example. Commonly called Atomic Layer Deposition, or ALD, this can result in extremely conformal coverage of very thin layers.

A trend in deposition equipment is the use of cluster tools. Here various deposition (and etch) steps are performed sequentially with the wafers passed automatically from one process chamber to another, without breaking vacuum. As a result multiple deposition and etch steps can be done without exposing the wafers to the atmosphere between steps. This minimizes exposure of the wafers to ambient particulate contamination and surface oxidation. It also minimizes wafer handling as well as increases throughput. Both PVD and CVD chambers can be connected sequentially, as well as plasma etch chambers. Most of these tools are single-wafer, as opposed to batch, systems. More control over each wafer is possible, but the deposition rates must be high since each wafer must be processed individually. Depositing films on one wafer at a time does give a little more flexibility in choosing deposition conditions. For example, one can operate in the mass transfer regime in CVD deposition since gas-phase mass transfer is easier to control in a single wafer system.

In regards to modeling and simulation, we have been mostly concerned in this chapter with topology. That is, we discussed models and simulations that predict the shape of the deposited films over nonplanar underlying topography—over steps or in vias for example. These have been useful in explaining some of the basic concepts of deposition and the different types of equipment. However, there are other important properties of

the deposited films, such as grain structure, composition, density, mechanical properties, and stress. It is important to model these physical and compositional properties as a function of deposition technique and conditions, as these will affect the film behavior and performance during subsequent processing and operation.

There have been deposition models and simulators developed for physical and compositional properties. An example is the SIMBAD/GROFILMS [9.45, 9.46] group that simulate as-deposited film structure. In SIMBAD, two-dimensional discs rather than atoms or molecules are used in simulating deposition, with each disc representing about 1000 atoms. In these Monte Carlo simulations, regions of dense, closely packed discs form and evolve into a columnar grain structure—the grains separated by the less dense grain boundaries. The exact structure depends on the amount of surface diffusion and flux shadowing, consistent with Thorton's zone model for microstructure evolution [9.47]. (For a good discussion of zone models for microstructure, see [9.48].) GROFILMS extends this modeling to include other processes that can occur during deposition such as wetting, reflow, grain boundary grooving, and crystal facetting. Many of these processes will be discussed in Chapter 11. Deposition simulators that include modeling the physical structure and composition of thin films will be increasingly important in the future.

Making the models more physically correct by including the chemistry and physics of specific deposition systems is important. The use of atomic-level models and simulators could also greatly help in developing more physically correct models. Simulations using Monte Carlo or Molecular Dynamics (MD) calculations, based on atomic-level, first principles models, have already revealed new insights into the deposition processes, as they have in oxidation, diffusion, and implantation modeling. The goal is to make these simulations more efficient and scalable to realistic dimensions and number of atoms. In addition they need to be linked to higher-level, continuum-type simulation tools to make their results more easily accessible.

9.7 Summary of Key Ideas

In this chapter we have examined how thin films are deposited as part of the fabrication of integrated circuits. Important issues in thin film deposition include the physical and chemical properties of the film, and coverage of nonplanar underlying topography, including step coverage and the filling of holes or spaces. Coverage and filling issues are especially important with the ever-decreasing lateral feature sizes, as well as with the increasing use of multilayer film structures.

The two main deposition techniques are Chemical Vapor Deposition (CVD) and Physical Vapor Deposition (PVD). CVD depositions are mostly done at elevated temperatures and are chemical reaction based. In the simple model for CVD presented, the deposition process is seen to be limited by the surface reactions or by mass transfer through a gas-phase boundary layer depending on the temperature, pressure, and system chemistry. At low pressures, mass transfer through the boundary layer is not a limiting step, and the surface reactions become rate limiting, allowing for fewer constraints on reactor geometry. However, shadowing by surface topographical features must still be considered to model coverage and fill profiles. In PVD, the processes are physically,

rather than chemically, based. Arrival angle distribution of the source material at the wafer surface is important. There are fewer gas-phase collisions between the source and the wafer and virtually no reactions on the surface, so the species can arrive more vertical to the wafer from the source and stick where they hit without desorbing and redepositing. Shadowing by topographical features can be very important in PVD depositions. Many different types of both CVD and PVD systems have evolved and are still being used for different applications and different types of films. Other deposition techniques include spin-on techniques for dielectrics and electrolytic deposition of metals.

Most simulation of thin film deposition is based on topography modeling, that is, predicting the shape of the deposited film over nonplanar topography. SPEEDIE is one such topography simulator for both CVD and PVD. Both direct and indirect fluxes of neutral and ionic species are used to model the deposition process in low pressure systems. The direct fluxes are modeled by the arrival angle distribution. Indirect fluxes are due to surface or near-surface processes, such as surface diffusion or reemission and resputtering of surface species with subsequent deposition elsewhere. The sticking coefficient is used to model the reemission process. CVD processes, used mostly for dielectrics, in general have a low sticking coefficient and isotropic arrival angle distribution, leading to good step coverage. A low sticking coefficient also leads to better filling of holes. PVD processes, used more for metals, have a high sticking coefficient. A narrower arrival angle distribution, more characteristic of PVD processes leads to worse coverage over a step but can result in better filling of the bottoms of holes. Concurrent sputtering and redeposition of material along with the directed deposition of ionized species can lead to good gap or hole filling of relatively high aspect ratio features in either PVD or CVD systems, such as in ionized sputtering or HDPCVD. Simulations show the importance of different components and different parameters, especially in terms of arrival angle, sticking coefficient, angle-dependent sputtering, and coverage and filling issues.

9.8 References

[9.1]. F. C. Eversteyn, "Chemical-Reaction Engineering in the Semiconductor Industry," *Philips Res. Rep.*, vol. 29, p. 45, 1974.

[9.2]. A.S. Grove, *Physics and Technology for Semiconductor Devices*, John Wiley & Sons, 1967, p. 12.

[9.3]. T. J. Rogers, "Advanced Integrated Circuit Technology for Micropower Integrated Circuits," Ph.D. thesis, Stanford University, 1975.

[9.4]. J. Li, "Topography Simulation of Intermetal Dielectric Deposition and Interconnection Metal Deposition Processes," Ph.D. thesis, Stanford University, 1996.

[9.5]. J. Li, J. P. McVittie, J. Ferziger, K. C. Saraswat, and J. Dong, "Optimization of Intermetal Dielectric Deposition Module Using Simulation," *J. Vac. Sci. Tech. B*, vol. 13, p. 1867, 1995.

[9.6]. M. Ohring, *The Materials Science of Thin Films*, Academic Press, 1992, p. 90.

[9.7]. *Handbook of Thin Film Technology*, L. I. Maissel and R. Glang, eds., McGraw-Hill, 1970.

[9.8]. M. Knudsen, "Die Verdampfung von Kristalloberflachen," *Ann. Physik*, vol. 52, p. 105, 1917.

[9.9]. R. Siegel and J. R. Howell, *Thermal Radiation Heat Transfer*, 3d ed., Hemisphere Pub. Corp., 1992.

[9.10]. S. A. Campbell, *The Science and Engineering of Microelectonic Fabrication*, Oxford University Press, 1996, p. 284.

[9.11]. B. N. Chapman, *Glow Discharge Processes*, John Wiley & Sons, 1980.

[9.12]. M. A. Lieberman and A. J. Lichtenberg, *Principles of Plasma Discharges and Materials Processing,* John Wiley and Sons, 1994.

[9.13]. A. A. Mayo, S. Hamaguchi, J. H. Joo, and S. M. Rossnagel, "Across-Wafer Nonuniformity of Long Throw Sputter Deposition," *J. Vac. Sci. Tech. B.*, vol. 15. p. 1788, 1997.

[9.14]. S. Hamaguchi and S. M. Rossnagel, "Liner Conformality Ionized Magnetron Sputter Metal Deposition Process," *J. Vac. Sci. Tech. B*, vol. 14, p. 2603, 1996.

[9.15]. S. M. Rossnagel, "Ionized Magnetron Sputtering for Lining and Filling Trenches and Vias," *Semiconductor International,*" February 1996, p. 99.

[9.16]. G. A. Dixit, M. F. Chisholm, M. K. Jain, T. Weaver, L. M. Ting, S. Poarch, K. Mizobuchi, R. H. Havemann, C.D. Dobson, A. I. Jeffryes, P. J. Holverson, P. Rich, D. C. Butler, and J. Hems, "A Novel High Pressure Low Temperature Aluminum Plug Technology for Sub – 0.5 Micron Contact/Via Geometries," *IEDM Technical Digest*, 1994, p. 105.

[9.17]. T. Kamins, *Polycrystalline Silicon for Integrated Circuit Applications,* 2d ed., Kluwer Academic Publishers, 1998.

[9.18]. P. Joubert, "The Effect of Low Pressure on the Structure of LPCVD Polycrystalline Si Films," *J. Electrochem. Soc.*, vol. 134, p. 2541, 1987.

[9.19]. R. Falckenberg, E. Doering, and H. Oppolzer, "Surface Roughness and Grain Growth of Thin P-Doped Polycrystalline Si-Films," *Proc. of Fall 1979 Electrochem. Soc. Meeting* (Los Angeles, October 1979), abstract 570, p. 1492.

[9.20]. H. J. Kim and C.V. Thompson, "Kinetic Modeling of Grain Growth in Polycrystalline Silicon Films Doped with Phosphorus or Boron," *J. Electrochem. Soc.*, vol. 135, p. 2312, 1988.

[9.21]. M. M. Manduarh, K. C. Saraswat, and T. I. Kamins, "Phosphorus Doping of Low Pressure Chemically Vapor-Deposited Silicon Films," *J. Electrochem. Soc.,* vol. 126, p. 1019, 1979.

[9.22]. S. Robles, K. Russell, M. Galiano, V. Siva, V. Kithcart, and B. C. Nguyen, "Gap Fill and Film Reflow Capability of Subatmospheric Chemical Vapor Deposited Borophosphosilicate Glass," *J. Electrochem. Soc.*, vol. 143. p. 1414, 1996.

[9.23]. L. Q. Xia, E. Yieh, P. Gee, F. Campana, and B. C. Nguyen, "Process Characteristics for Subatmospheric Chemical Vapor Deposited Borophophosilicate Glass and Effect of Carrier Gas," *J. Electrochem. Soc.,* vol. 144. p. 3208, 1997.

[9.24]. E. Kondoh, Y. Kawano. N. Takeyasu, T. Katagiri, H. Yamamoto, and T. Ohta, "Interconnection Formation by Simultaneous Cu Doping in CVD Al (Al-Cu CVD)," *IEDM Technical Digest*, p. 277, 1993.

[9.25]. D. Pramanik and V. Jain, "Barrier Metals for ULSI: Deposition and Manufacturing," *Solid State Technology,* January 1993, p. 73.

[9.26]. R. Liu, "Metallization," in *ULSI Technology*, ed. C.Y. Chang and S. M. Sze, McGraw-Hill, NY, 1996, p. 388.

[9.27]. J. Li, T. E. Seidel, and J. W. Mayer, "Copper-based Metallization in ULSI Structures," *MRS Bulletin*, vol. 19, no. 8, p. 15, 1994.

[9.28]. Y. Shacham-Diamand, V. Dubin, and M. Angyal, "Electroless Copper Deposition for ULSI," *Thin Solid Films*, vol. 262, p. 93, 1995.

[9.29]. SPEEDIE 3.0 Manual, Stanford University, 1995.

[9.30]. K. Watanabe and H. Komiyama, "Micro/Macrocavity Method Applied to the Study of the Step Coverage Formation Mechanism of SiO_2 Films by LPCVD," *J. Electrochem. Soc.*, vol. 137, p. 1222, 1990.

[9.31]. H. C. Wulu, K. C. Saraswat, and J. P McVittie, "Simulation of Mass Transport for Deposition in Via Holes and Trenches," *J. Electrochem. Soc.*, vol. 138, p. 1831, 1991.

[9.32]. J. C. Rey, L. Y. Cheng, J. P. McVittie, and K. C. Saraswat, "Monte Carlo Low Pressure Deposition Profile Simulations," *J. Vac. Sci. Tech. A*, vol. 9, p. 1083, 1991.

[9.33]. L. Y. Cheng, J. P McVittie, and K. C. Saraswat, "New Test Structure to Identify Step Coverage Mechanisms in CVD of Silicon Dioxide," *Appl. Phys. Lett.*, vol. 58, p. 2147, 1991.

[9.34]. C. Y. Chang, "Experiments and Simulation of Plasma Deposition and Sputter Etching Processes," Ph.D. thesis, Stanford University, 1995.

[9.35]. David Bang, "Modeling and Simulation of Physical Vapor Deposition Systems for VLSI Metallization," Ph.D. thesis, Stanford University, 1997.

[9.36]. C. Y. Chang, J. P McVittie, J. Li, K. C. Saraswat, S. E. Lassig, and J. Dong, "Profile Simulation of Plasma Enhanced and ECR Oxide Deposition and Sputtering," *IEDM Technical Digest*, p. 853, 1993.

[9.37]. J.C. Rey, L.Y. Cheng, J.P. McVittie, and K.C. Saraswat, "Numerical Simulation of CVD Trench Filling Using a Surface Reaction Coefficient Model," *Proc. of Seventh Int'l. IEEE Multilevel Inter. Conf.*, Santa Clara, CA, June 1990, p. 425.

[9.38]. T. S. Cale, M. K. Jain, D. S. Taylor, R. L. Duffin, and C. J. Tracy, "Model for Surface Diffusion of Aluminum-(1.5%) Copper during Sputter Deposition," *J. Vac. Sci. Tech. B*, vol. B11, p. 311, 1993.

[9.39]. W. G. Oldham, An. R. Neureuther, C. Sung, J. L. Reynolds, and S. N. Nandgaonkar, "A General Simulator for VLSI Lithography and Etching Processes: Part II—Application to Deposition and Etching," *IEEE Trans. Elec. Dev.*, vol. ED-27, p. 1455, 1980.

[9.40]. ATHENA is a process simulator from Silvaco Inc.

[9.41]. I. A. Blech, "Evaporated Film Profiles Over Steps in Substrates," *Thin Solid Films*, vol. 6. p. 113, 1970.

[9.42]. TAURUS-TOPOGRAPHY is a topography simulator from Avant! Inc.

[9.43]. E. Meeks, R. S. Larson, P. Ho, C. Apblett, S. M. Han, E. Edelberg, and E. S. Aydil, "Modeling of SiO_2 Deposition in High Density Plasma Reactors and Comparisons of Model Predictions with Experimental Measurements," *J. Vac. Sci. Tech. A*, vol. 16, p. 544, 1998.

[9.44]. "National Technology Roadmap for Semiconductors," SIA, 1997.

[9.45]. S. K. Dew, T. Smy, and M. J. Brett, "Simulation of Elevated Temperature Al Metalization Using SIMBAD," *IEEE Trans. Elec. Dev.*, vol 39, p. 1599, 1992.

[9.46]. L. J. Frederick, S. K. Dew, M. Brett, and T. Smy, "Thin Film Microstructure Modelling Through Line-Segment Simulation," *Thin Solid Films*, vol. 266, p. 83, 1995.

[9.47]. J. A. Thorton, "Influence of Apparatus Geometry and Deposition Conditions on the Stucture and Topology of Thick Sputtered Films," *J. Vac. Sci. Tech.*, vol. 11, p. 666, 1974.

[9.48]. M. Ohring, *The Materials Science of Thin Films*, Academic Press, 1992, p. 223 ff.

9.9 Problems

9.1. What would you expect would happen to the threshold voltage of a MOS transistor if the gate oxide were deposited using $SiH_4 + 2O_2 \rightarrow SiO_2 + 2H_2O$ rather than thermally grown?

9.2. What are the two commonly observed rate limiting steps in silicon epitaxial growth? Under what conditions do they normally dominate the overall deposition rate?

9.3. In an epitaxial deposition, under mass transfer limited conditions, is it more important to control the reactor temperature or the source gas composition in the gas stream to obtain reproducible results? Why?

9.4. In a reactor used for epitaxial growth, the wafers are normally placed flat on the susceptor, and epi grows on the top side only. If the same reactor were used to oxidize wafers, by introducing O_2 rather than SiH_4 (or another Si gas source), SiO_2 would grow on both sides of the wafer. Explain why SiO_2 grows on both sides and epi grows only on the top side.

9.5. For CVD deposition of a film, it is found that the mass transfer coefficient $h_G = 10$ cm sec^{-1} and the surface reaction rate coefficient $k_S = 1 \times 10^7 \exp(-1.9 \text{ eV}/kT)$ cm sec^{-1}. For a deposition at 900°C, which CVD system would you recommend using: (a) a cold-walled, graphite susceptor type: or (b) a hot-walled, stacked wafer type? Explain your answer.

9.6. Calculate the deposition rate for an LPCVD system with the same parameter values given in the example following Eq. (9.10), but at a reduced total pressure so that h_G is increased by 100 times. Assume that the partial pressure of the incorporating species, P_G, remains the same, and C_T decreases by the same factor as the total pressure.

9.7a. Plot the deposition rate (on a log scale) versus $1/T$ (Kelvin), for 600–1200°C, for a CVD system with the following parameter values:

$h_G = 0.5$ cm sec^{-1}
$k_S = 4 \times 10^6 \exp(-1.45 \text{ eV}/kT)$ cm sec^{-1}
Partial pressure of incorporating species = 1 torr
Total pressure = 1 atm
$C_T/N = 1/10{,}000$

Identify the reaction and mass transfer limited regimes.

b. Redo the problem when the total pressure is decreased to 1 torr, so that h_G increases by 100 times. Assume that the partial pressure of the incorporating species remains the same, and C_T decreases by the same factor as the total pressure.

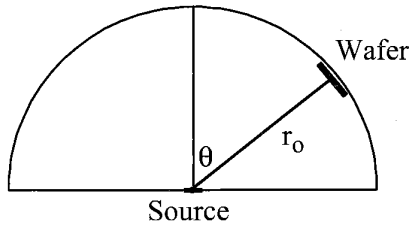
9.8. Show that placing the source in an evaporation system at the inside surface of a sphere facing the center, with the wafers also at the inside surface of the sphere facing the center, leads to uniform deposition wafer to wafer, assuming a small planar surface source.

9.9. Calculate the deposition rate for a small planar surface evaporation source in which $\theta_i = 30°$, $\theta_k = 45°$, the evaporation rate is 1×10^{-3} gm sec^{-1}, the distance

from the source to the wafer is 5 cm, and the density of the material being deposited equal 5 gm cm^{-3}.

9.10. The new deposition engineer installed the company's new evaporation system. Hoping to get uniform depositions on all the wafers mounted on the inside of the spherical wafer holder, he installed the evaporation source crucible at the center of the sphere. If the evaporation source behaves like an ideal small-area planar source, what will be the deposition rate as a function of θ, as defined below? (Let the evaporation rate equal 3×10^{-3} gm sec^{-1}, radius r_o equal 10 cm, and the density of the material being deposited equal 10 gm cm^{-3}.) Sketch a plot of d versus θ for θ from $-90°$ to $+90°$, and specify the deposition rate at a point directly facing the planar source (at $\theta = 0°$) and at $\theta = 90°$. Will uniform depositions result?

9.11. Calculate the mean free path of a particle in the gas phase of a deposition system



and estimate the number of collisions it experiences in traveling from the source to the substrate in each of the cases below. Assume that in each case the molecular collisional diameter is 0.4 nm, the source-to-substrate distance is 5 cm, and that the number of collisions is approximately equal to the source-to-substrate distance divided by the mean free path.

a. An evaporation system in which the pressure is 10^{-5} torr and the temperature is 25°C.

b. A sputter deposition system in which the pressure is 3 mtorr and the temperature is 25°C.

c. An LPCVD system in which the pressure is 1 torr and the temperature is 600°C.

d. An APCVD system in which the pressure is 1 atm and the temperature is 600°C.

9.12. a. Using schematic diagrams, show qualitatively how changing the gas pressure in a standard PVD system (using an appropriate S_c value for this type of system) can affect the deposition profile when depositing a material on a narrow trench structure.

b. In a similar fashion, show how changing the sticking coefficient affects the deposition profile when depositing a material on topography with a step.

9.13. How does the ability to fill the bottom of a narrow trench using sputter deposition change as the target is moved further away from the wafer? Neglect any gas phase collision effects.

9.14. Explain how asymmetric depositions can occur on a wafer in a sputter deposition system. Asymmetric deposition means that thicker deposition occurs on one side of a feature (a step or side of a trench, for example) than the other. Also suggest a way to decrease the asymmetry. Use schematic diagrams in your explanations.

9.15. What value of n in the arrival angle distribution is desired for good step coverage over a step in topography? For good bottom filling of a via? Explain with schematic diagrams. What value of S_C, the sticking coefficient, do you want in each case? Explain with diagrams.

9.16. We have seen how conformal coverage in low pressure systems results from a low sticking coefficient, which causes multiple reemission and redeposition processes within the feature topography. Describe how in atmospheric CVD systems, gas phase collisions by the reactants within the topograhical features can lead to good coverage and filling.

Etching

10

10.1 Introduction

After thin films are deposited on the wafer surface, they are usually selectively removed by etching to leave the desired pattern of the film on the wafer surface. This process was used many times in the CMOS process flow in Chapter 2 and is summarized in Figure 10–1. In addition to deposited films, parts of the silicon substrate itself may be etched, such as in creating trenches in isolation structures. The masking layer may be photoresist, or it may be another thin film such as silicon dioxide or silicon nitride. Oxide or nitride masks stand up better to etching conditions than photoresist and are often called hard masks. But they themselves must be selectively etched, usually using lithographically defined photoresist as the masking layer. The etching of a thin film is usually done until a different layer is reached underneath. Etching only partway through a layer without reaching an underlying material is sometimes done but can be difficult to control uniformly over the whole wafer. Multilayer structures can be etched sequentially using the same masking layer.

Etching can be done in either a "wet" or "dry" environment. Wet etching involves the use of liquid etchants. The wafers are usually immersed in the etchant solution and the exposed material is etched mostly by chemical processes. While such processes are not common in modern VLSI processing for reasons we shall see, they are sometimes used, as in Figure 2–34 in Chapter 2. Dry etching involves the use of gas-phase etchants in a plasma. Here the etching usually takes place by a combination of chemical and physical processes. Because a plasma is involved, dry etching is often called "plasma etching." This is normally the etching method today and was used many times in Chapter 2 (Figure 2–4 for example).

The ideal etch profile in Figure 10–1 has perfectly straight sidewalls exactly under the edge of the mask. What can happen in reality is shown in Figure 10–2. In Figure 10–2(a) the etching is both vertical and lateral. This leads to undercutting of the photoresist mask and nonvertical sidewalls in the film. In Figure 10–2(b) the photoresist is not perfectly rectangular, with rounded top corners and sloped sides. In addition, the etch rate of the photoresist is nonzero, and the mask itself is etched, or eroded, to some degree. This occurs both on its top and on its sides. This leads to even more lateral etching of the film since the photoresist mask gets narrower as the etching continues. If the substrate is attacked by the etchant as well, then its profile will change. Figure 10–2(b) also illustrates this issue, which is related to the selectivity of the etching process.

Etch selectivity is the ratio of the etch rates of the different materials in an etch process. When the etch rates of the mask and the underlying substrate are near zero,

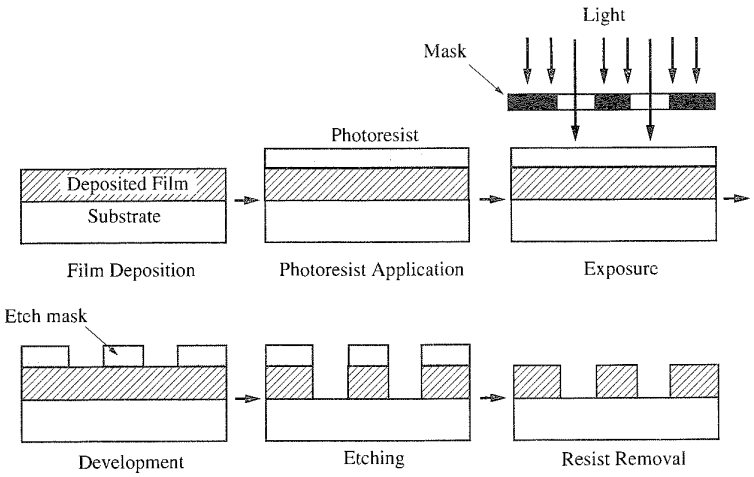


Figure 10–1 General process used in integrated circuit fabrication to define patterned films.

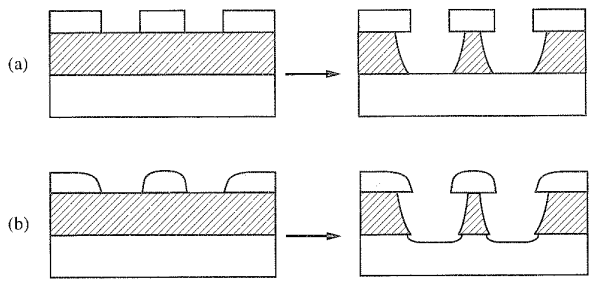


Figure 10–2 Actual etch profiles that can occur. (a) Lateral etching under mask. (b) Rounded photoresist which is further eroded during etching, leading to even more lateral etching. (b) also illustrates etch selectivity.

while the etch rate of the film is appreciable, then the etch selectivity of the film with respect to both the mask and the substrate is high. This is normally a desired situation. If the etch rate of the mask or substrate is significant, then the selectivity is poor. Selectivities, or etch rate ratios, in the range of 25–50 are usually considered reasonable. Different materials have different etch rates usually because of the chemical effects involved with etching, as opposed to the physical effects. Chemical reactions can be very selective whereas physical processes like sputtering tend to be similar for most materials.

Etch directionality is a measure of the relative etch rates in different directions, usually vertical versus lateral, as illustrated in Figure 10–3. Isotropic etching occurs when

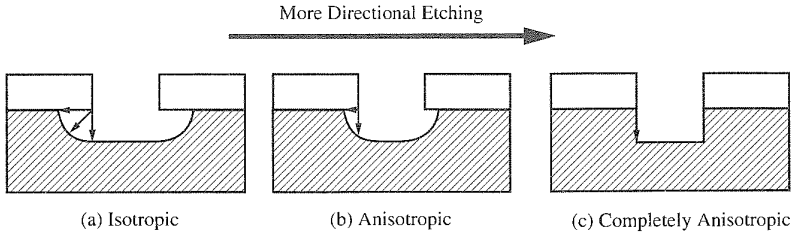


Figure 10–3 Etch profiles for different degrees of anisotropic, or directional, etching: (a) purely isotropic etching; (b) anisotropic etching; (c) completely anisotropic etching.

etch rates are the same in all directions, as in Figure 10–3(a). The etch distance laterally under each side of the mask is the same as vertically under the mask opening. As the opening under the mask approaches zero, the etch profile would approach a perfect semicircle. Figure 10–3(b) shows anisotropic etching, where less etching occurs laterally. Figure 10–3(c) shows completely anisotropic etching, with no etching at all in the lateral direction. Etch directionality is usually related to physical effects in etching such as ion bombardment and sputtering. As we shall see, the more physical an etch process is, the more directional and the less selective it generally is. On the other hand, the more chemical an etch process is, the more selective and the less directional it generally is.

Directionality is often desired in an etch process. Vertical etching maintains lithographically defined feature sizes as they are transferred by etching into underlying films. Vertical etching is also utilized in special applications such as sidewall spacer formation. (Recall Figure 2–30 in the CMOS process flow in Chapter 2.) However, vertical structures that result from vertical etching may also cause problems in subsequent processing. Depositing another layer on top of a structure with vertical steps may cause step coverage problems, as we discussed in Chapter 9. Less steep steps, obtained with less directional etching, would make it easier to obtain good step coverage and would allow easier filling in later deposited layers. Because of shrinking feature sizes, however, modern fabrication methods tend to employ directional etching and vertical steps with little undercutting, and use deposition techniques that achieve good filling and step coverage even for near-vertical structures.

Selectivity is also a requirement. This means that the etch rate of the layer to be etched should be fast compared to the etch rate of the mask and etch rate of the substrate or layer below. Achieving good selectivity and directional etching at the same time is sometimes quite difficult. Other requirements for an etch process are that the gases and liquids used for the etching must be able to be transported to the wafer surface, and the byproducts of the etch process must be transported away easily. The latter means that soluble byproducts are produced in wet etching, and that gaseous or volatile byproducts are produced in dry etching. The etch process should be uniform, safe to use, and should cause minimal damage to structures on the chip. Finally, the etching process should be clean, with low particulate production and little film contamination, and should be cost effective.

10.2 Historical Development and Basic Concepts

There are two main methods of etching used in the semiconductor industry: wet etching and dry, or plasma, etching. In the early days of the industry, wet etching was used exclusively. It was a well-established, simple, and inexpensive technology. Wet etching can also be very selective. Eventually, however, the need for smaller linewidths and more vertical structures required new techniques. Plasma etching methods were developed for integrated circuit fabrication and are used for most etching steps today.

In this section we will describe the basic concepts involved in both of these techniques. We will start with wet etching, which we will use to introduce concepts such as selectivity, overetch, etch bias, and anisotropy that are applicable to both wet and dry etching.

10.2.1 Wet Etching

The first etchants used in the integrated circuit industry were simple wet chemical etchants. By immersing the wafers into baths of liquid chemicals, the exposed films could be etched away, leaving unetched those regions of the film that were masked with photoresist or other films. Wet etches were developed for all steps in the fabrication process. A common wet etch for SiO_2 is hydrofluoric acid, HF. The overall reaction is

$$SiO_2 + 6HF \rightarrow H_2SiF_6 + 2H_2O \tag{10.1}$$

Wet etchants work through chemical processes rather than physical processes. In general, wet etchants work by chemically reacting with the film to form water-soluble byproducts or gases, such as the H_2SiF_6 byproduct which is a water-soluble complex in the above example.

In some cases the etching occurs by first oxidizing the surface of the film or material and then dissolving the oxide. For example, a common etchant for silicon is a mixture of nitric acid (HNO_3) and HF. The nitric acid partially decomposes to nitrogen dioxide, NO_2, which then oxidizes the surface of the silicon by the reaction

$$Si + 2NO_2 + 2H_2O \rightarrow SiO_2 + H_2 + 2HNO_2 \tag{10.2}$$

The HF then dissolves the SiO_2 by the reaction given in Eq. (10.1). The overall reaction is

$$Si + HNO_3 + 6HF \rightarrow H_2SiF_6 + HNO_2 + H_2O + H_2 \tag{10.3}$$

Buffering agents are often added to the etch solutions to keep the etchants at maximum strength over use and time. For example, ammonium fluoride (NH_4F) is added to HF to help prevent depletion of the fluoride ions in the oxide etch. This is called "Buffered HF" or BHF, or more commonly BOE for "Buffered Oxide Etch." The addition of ammonium fluoride also decreases the etch rate of photoresist and helps to minimize lifting of the resist during oxide etching. Acetic acid, CH_3COOH, is often added to the nitric acid/hydrofluoric acid silicon etch to limit the dissociation of the nitric acid.

Wet etches can be very selective because they depend on chemistry. The limiting steps in most etch processes are usually the chemical reactions in which the etch species react

with the film, forming the soluble byproducts. The selectivity, S, of an etch process between two materials, 1 and 2, is simply the ratio of their etch rates, r, in that etchant, or

$$S = \frac{r_1}{r_2} \tag{10.4}$$

Material 1 is usually the film being etched, and material 2 is either the masking material (photoresist or perhaps silicon dioxide) or the material below the film. We often speak of an etch process for material 1 having a certain selectivity over (or "with respect to," or just "to") material 2. For a large selectivity ($S>>1$) we say that the etch process for material 1 has good selectivity over material 2.

Selectivity is important for a number of reasons. First, the selectivity for the material to be etched with respect to the mask determines how thick the mask material must be so that the mask is not completely removed during the etching process. Second, all etch processes generally employ some amount of overetching in order to ensure that the desired etching process has gone to completion. Thus underlying films are normally subjected to the etchant for some period of time toward the end of the etch period. The selectivity for the material being etched with respect to the underlying material determines how much etching of the underlying materials occurs.

Example 1 μm of SiO_2 is being etched on top of a Si substrate. The etch rate of the oxide is 0.40 μm min^{-1}, and the etch selectivity of the oxide with respect to the silicon is 25-to-1. If the etch is done for three minutes, how much of the underlying Si is etched?

Answer From Eq. (10.4), the etch rate of the Si can be calculated:

$$S = \frac{r_1}{r_2} \Rightarrow 25 = \frac{0.40\ \mu\text{m min}^{-1}}{r_{Si}} \Rightarrow r_{Si} = 0.016\ \mu\text{m min}^{-1}$$

The Si is only exposed to the etch after the overlying oxide is completely etched off. The time required to etch away the oxide equals 1 μm/0.40 μm min^{-1}, which equals 2.5 minutes. Therefore, the Si is etched for 0.5 minutes. Hence, the amount of Si etched is 0.016 μm min^{-1} * 0.5 min, which equals 0.008 μm, or 8 nm.

Most wet chemical etchants etch isotropically. The exceptions to this are those which are sensitive to crystallographic orientation. For example, some etchants etch much slower in the <111> direction in crystalline silicon as compared to the other directions. But most commonly used wet etches etch equally in all directions. In Figure 10–3(a) we saw how isotropic etching results in etching underneath the edge of the mask. The amount of undercutting is called the "etch bias." This is shown in Figure 10–4(a). Here we assume that the selectivity of the etch with respect to both the mask and the substrate is infinity (i.e., no etching at all of those materials). The etch thickness or depth is

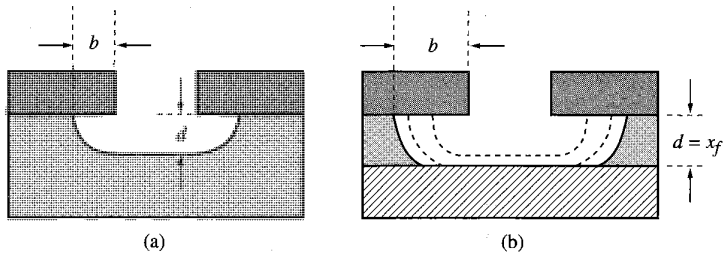


Figure 10–4 Illustration of etch bias and overetch. In (a) the etch bias, b, is shown for a given etch depth, d. In (b), overetching is illustrated where etching is continued even after the etch depth, d, equals the film thickness, x_f, with the result that the etch bias increases.

d and the bias is b. For perfectly isotropic etching b would equal d, but only if etching is stopped before the bottom of the film is reached.

As mentioned earlier, some overetching is normally part of every etch process. If each layer were of uniform thickness, and the etch were perfectly uniform through time and space, then overetching would theoretically not be needed. The etch time required would be based on the previously determined etch rate or on other endpoint detection methods, to be discussed later. It would not even matter what the etch selectivity over the layer beneath is in that case. However, the thickness of films is not perfectly uniform, and they do not always etch exactly uniformly. So some overetch is always done in manufacturing to ensure that the etch goes to completion everywhere. This is illustrated in Figure 10–4(b). In this figure we again assume that the selectivity with respect to the mask and substrate are infinite and that the etching is purely isotropic. The bias b now increases with time as the overetch continues. The amount of overetching needed, usually measured in terms of time or % time, can be determined by estimating the uncertainties in etch rates and in the nonuniformities of the thicknesses and then calculating the worst case etch time needed. Ten to twenty percent overetches are common. This ensures that while some of the structures may be overetched, none are underetched. We will see later that when the etching is anisotropic, overetching is also required to remove residual film from steps in the topography.

Example A 0.5-μm layer of silicon dioxide on a Si substrate needs to be etched down to the Si. Assume that the nominal oxide etch rate is r_{ox} (in μm min^{-1}). There is a ± 5% variation in the oxide thickness and a ± 5% variation in the oxide etch rate. (a) How much overetch is required (in % time) in order to ensure that all the oxide is etched? (b) What selectivity of the oxide etch rate to the Si etch rate is required so that a maximum of 5.0 nm of Si is etched? (Assume the overetch calculated in part (a) is done.)

Answer (a) The nominal etch time = $x/r_{ox} = 0.5/r_{ox}$ (x = thickness in μm.) The overetch is done to make sure all the oxide is etched for the worst case condition; that means for the thickest oxide and the slowest etch rate. The thickest oxide =

0.5 * 1.05 = 0.525 μm. The slowest etch rate = r_{ox} * 0.95 μm min^{-1}. The time to etch the worst case = x/r = 0.525/(r_{ox} * 0.95) = 0.553/r_{ox} min. The amount of overetch, in % time, is just the worst case etch time divided by the nominal etch time = (0.553/r_{ox})/(0.5/r_{ox}) = 1.106 or 10.6% overetch.

(b) The maximum amount of Si that will be etched will occur under the thinnest oxide being etched at the fastest etch rate. The thinnest oxide is 0.95 * 0.5 μm = 0.475 μm. The fastest etch rate is r_{ox} * 1.05 μm min^{-1}. The time to etch through that oxide is x/r = 0.475/(r_{ox} * 1.05) = 0.4523/r_{ox} min. The time that the Si is exposed to the etch is equal to the total etch time (0.553/r_{ox}, as calculated in part (a)) minus the time it took to etch the thinnest oxide, etching at the fastest rate, or

$$(0.553/r_{ox}) - (0.4523/r_{ox}) = 0.1003/r_{ox} \text{ min}$$

We want to etch 0.005 μm or less of the Si, so

$$r_{Si} = x/\text{time} = 0.005/(0.1003/r_{ox}) \text{ μm min}^{-1}$$
$$= 0.050 * r_{ox} \text{ μm min}^{-1}$$

The selectivity of oxide etch rate to Si etch rate is therefore:

$$r_{ox}/r_{Si} = r_{ox}/(0.050 * r_{ox})$$
$$r_{ox}/r_{Si} = 20{:}1$$

Since overetching is usually done, it is important that the selectivity with respect to the layer below be as high as possible or else some etching of that layer will occur during the overetch. On the other hand, if the selectivity is very large, one cannot indiscriminately overetch. This would cause unwanted undercutting or bias.

The above examples assume infinite etch selectivity of the film with respect to the mask; that is, the etch rate of the mask is zero. But there will be some finite etch rate of the mask, which is usually photoresist. What will happen? Obviously the mask will etch at some rate during the etching of the film. This is shown in Figure 10–5(a) for the case of isotropic etching and a rectangular-shaped mask. Here Δm is the amount of mask etched, the same in all directions in this case. This is called mask erosion. Obviously the extent of lateral etching will increase beyond what would be expected without any mask erosion. For completely anisotropic (vertical) etching, etch erosion should not be a problem as long as the mask is perfectly rectangular in shape. The etch will only occur from the top. Only if the mask is etched all the way through would any difficulties arise.

In fact, the mask shape is usually not perfectly rectangular but often has rounded corners or sloped sides. Mask erosion leading to extra undercutting can occur even for completely anisotropic etching, as illustrated in Figure 10–5(b). With a finite etch rate for the mask, there will be equal vertical etching at all points of the mask. At the sloped edge, a finite amount of the mask will be etched away. This will allow lateral erosion of

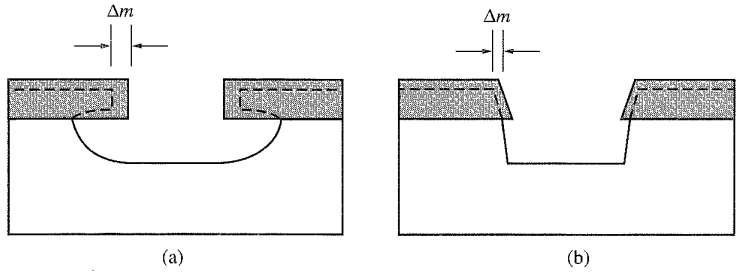


Figure 10–5 Illustration of mask erosion. In (a), there is an idealized rectangular mask and isotropic etching of the mask and the substrate. In (b), there is a sloped mask and anisotropic etching. In both cases more lateral etching of the substrate occurs as a result of the mask etching, or erosion.

the mask, shown by Δm in the figure, and etching of the film in this thin region will occur. This will result in undercutting with respect to the original mask edge, even for perfectly anisotropic, or vertical, etching.

Etching is often neither perfectly anisotropic nor perfectly isotropic, as was shown in Figure 10–3(b). We can define the degree of anisotropy of a film, A_f, as

$$A_f = 1 - \frac{r_{\text{lat}}}{r_{\text{ver}}} \tag{10.5}$$

where r_{lat} and r_{ver} are the etch rates in the lateral and vertical directions, respectively. In Figure 10–4, this would correspond to

$$A_f = 1 - \frac{b}{d} \tag{10.6}$$

Isotropic etching [Figure 10–3(a)] would have an anisotropy of 0, completely anisotropic etching [Figure 10–3(c)] would have an anisotropy of 1, and general anisotropic etching [Figure 10–3(b)] would have an anisotropy between 0 and 1. For a structure etched exactly to the bottom of the film, with no overetch, the anisotropy would be

$$A_f = 1 - \frac{b}{x_f} \tag{10.7}$$

where b is the bias and x_f is the film thickness.

Example Consider the structures shown in Figure 10–6. A 0.5-μm-thick oxide layer is etched to achieve equal structure widths and spacings (S_f). The etch process produces

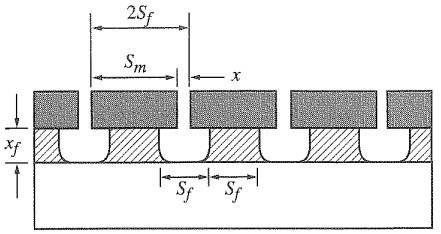


Figure 10–6 Structure and dimensions needed to fabricate features with equal widths and spaces (S_f) for a mask size of S_m, film thickness of x_f, and degree of anisotropy of A_f. x is the spacing between the masks and is usually the minimum lithographic feature. (After [10.1].)

a degree of anisotropy of 0.8. If the distance between the mask edges, x, is 0.35 μm, what structure spacings and widths are obtained? Neglect any overetch.

Answer To obtain equal widths and spacings, S_f, the mask width, S_m, must be made larger to take into account the anisotropic etching. Since

$$S_m = S_f + 2b$$

where b is the bias on each side, and using Eq. (10.7) we can now rewrite S_m as

$$S_m = S_f + 2x_f(1 - A_f)$$

which is the dimension of the mask needed. [This makes sense. For isotropic etching ($A_f = 0$), S_m is a maximum, and for completely anisotropic etching ($A_f = 1$), S_m is a minimum and equal to S_f.] The distance between the mask edges, x, is usually the minimum feature size that can be lithographically resolved. Since

$$x = 2S_f - S_m$$

x can be rewritten, using the expression above for S_m, as

$$x = S_f - 2x_f(1 - A_f)$$

Rewriting this in terms of S_f gives

$$S_f = x + 2x_f(1 - A_f) \tag{10.8}$$

Plugging the above numbers into Eq. (10.8) gives

$$\begin{aligned} S_f &= 0.35\mu m + 2(0.5\mu m)(1 - 0.8) \\ &= 0.55\ \mu m \end{aligned}$$

Eq. (10.8) shows that the structure size approaches the minimum lithographic dimension only when the film thickness gets very small (not always practical) or as the anisotropy gets close to 1. This means completely vertical etching. Unfortunately, wet etching does not result in vertical etching, but rather isotropic etching where A_f is a minimum. These considerations give some insight into the reasons wet chemical etching is not very commonly used in chip fabrication today.

Table 10–1 Common wet chemical etchants for various thin films used in IC fabrication

Material	Etchant	Comments
SiO_2	HF (49% in water) "straight HF"	Selective over Si (i.e., will etch Si very slowly in comparison). Etch rate depends on film density, doping.
	NH_4F:HF (6:1) "Buffered HF" or "BOE"	About $^1/_{20}$ th the etch rate of straight HF. Etch rate depends on film density, doping. Will not lift up photoresist like straight HF.
Si_3N_4	HF (49%)	Etch rate depends strongly on film density, O, H in film.
	H_3PO_4:H_2O (boiling @ 130–150°C)	Selective over SiO_2. Requires oxide mask.
Al	H_3PO_4:H_2O:HNO_3:CH_3COOH (16:2:1:1)	Selective over Si, SiO_2, and photoresist.
Polysilicon	HNO_3:H_2O:HF (+ CH_3COOH) (50:20:1)	Etch rate depends on etchant composition.
Single crystal Si	HNO_3:H_2O:HF (+ CH_3COOH) (50:20:1)	Etch rate depends on etchant composition.
	KOH:H_2O:IPA (23 wt. % KOH, 13 wt. % IPA)	Crystallographically selective; relative etch rates: (100): 100 (111): 1
Ti	NH_4OH:H_2O_2:H_2O (1:1:5)	Selective over $TiSi_2$.
TiN	NH_4OH:H_2O_2:H_2O (1:1:5)	Selective over $TiSi_2$.
$TiSi_2$	NH_4F:HF (6:1)	
Photoresist	H_2SO_4:H_2O_2 (125°C)	For wafers without metal.
	Organic strippers	For wafers with metal.

Before we describe techniques which do result in more anisotropic etching, we will look at some of the wet etch solutions that have been used, and are still used to some extent, in the Si integrated circuit industry. Table 10–1 gives a brief sampling of some of the common solutions and compositions.

The etch rate of a material in a particular etchant depends on the specific composition of the etch solution, and the etch rate can vary by orders of magnitude for different compositions of the same constituents. The rates of chemical reactions are strong functions of temperature, so the etch rates using wet, or chemical, etchants should be functions of temperature as well. For example, the etch rate of silicon dioxide in a 6:1 Buffered Oxide Etch (BOE) can change by a factor of two for a 10°C change in temperature [10.2].

Etch rates also depend on the composition and density of the films. For example, the etch rate of undoped, undensified CVD silicon dioxide in BOE is about 5 nm sec^{-1}, while the etch rate of phosphorus doped (8 atomic %) CVD silicon dioxide is about 9 nm sec^{-1}. On the other hand, the etch rate of densified (annealed at 800–1000°C), undoped CVD silicon dioxide is only about 1.5 nm sec^{-1}, the same as for thermal oxide. Clearly, the composition of the etchant and film, the density of the film, and the etch solution all affect the wet etch rates significantly.

The etch rate can also depend on the crystallographic orientation in some cases. In Table 10–1, the etch rates for two crystallographic directions in Si using KOH are given.

One sees that the etch rate for this etchant in the [100] direction is much faster than in the [111] direction, because the (111) planes are the most closely packed and etch more slowly. This can be used to form v-grooves on the surface of a (100) Si wafer because the {111} planes intersect the surface at a 54.7° angle.

10.2.2 Plasma Etching

While wet chemical etching was successfully used in the early days of integrated circuit manufacturing, it has been largely replaced by plasma, or dry, etching. There are two main reasons for this.

The first reason is that very reactive chemical species are produced in a plasma which can often etch more vigorously than species in a nonplasma environment. One of the first widespread uses of plasma etching was in the early 1970s to etch PECVD silicon nitride, used as the final passivation layer in integrated circuits. This film is difficult to etch with wet chemical etchants. Wet HF can etch it, but the etch rate is quite low, and the process is not selective with respect to SiO_2 which is often underneath. Boiling phosphoric acid was commonly used instead. But this etchant often lifts off photoresist masks so that a mask of SiO_2 has to be used, requiring its own masking and etching steps. A better etch process was sought. It was soon discovered that by using a CF_4/O_2 gas mixture in a plasma atmosphere, atomic fluorine is produced which can etch the nitride film much more easily. In addition, a photoresist mask is not lifted off in this process, making the process much simpler.

The second, and more important, reason why plasma etching has supplanted wet etching is the fact that directional or anisotropic etching is possible with plasma etch systems. Directional etching is needed to minimize underetching and etch bias, which then allows smaller and more tightly packed structures. This directional etching is due to the presence of ionic species in the plasma and the electric fields that direct them normal to the wafer surface.

Plasma etch systems can be designed so that either the reactive chemical components or the ionic components dominate. In many cases, a combination of ionic and reactive chemical species, usually acting in a synergistic manner, are utilized. The net etch rate can be much faster than the sum of the individual etch rates when they are acting alone. In addition, by utilizing both components this way, directionality can be achieved while maintaining an acceptable degree of selectivity.

A basic plasma etch system is shown schematically in Figure 10–7. This is similar to the PECVD and RF sputtering systems shown earlier in Figures 9–14 and 9–28. In these etch systems a low-pressure (1 mtorr – 1 torr) gas is used in the chamber. By applying a high electric field across two electrodes, some of the gas atoms are ionized, producing positive ions and free electrons and creating a plasma, as discussed in Chapter 9. In plasma etching systems, the energy is supplied by an RF generator, usually operating at 13.56 MHz. As we saw in Chapter 9, a voltage bias develops between the plasma and the electrodes. This is due to the difference in mobility of the electrons and the ions. Initially, the more mobile electrons are lost to the electrodes at a faster rate than the slower ions. This results in the plasma being biased positively with respect to the electrodes. For a symmetric RF plasma system, with the two electrodes of equal area, the

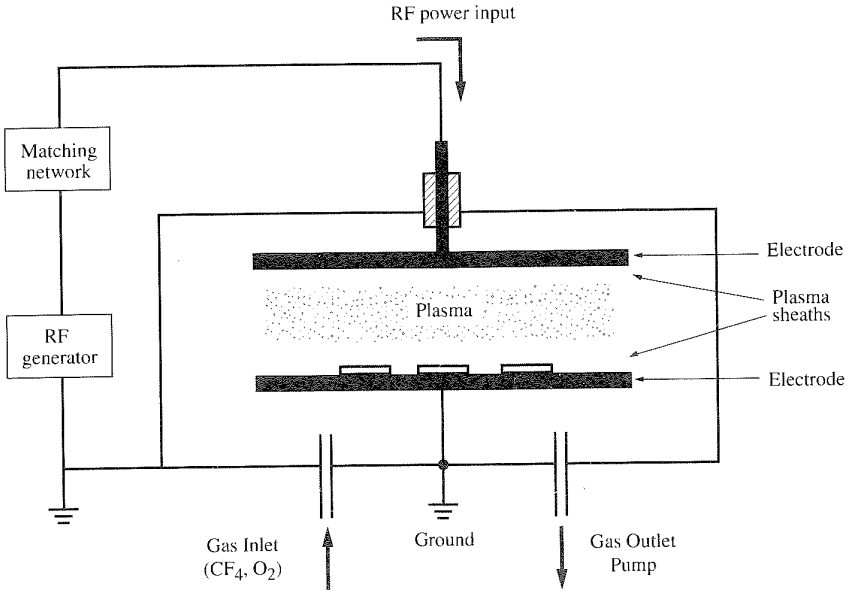


Figure 10–7 Schematic diagram of an RF-powered plasma etch system.

voltage distribution that develops is shown by the solid curve in Figure 10–8 (repeated from Figure 9–29). The sheaths are the regions next to each electrode where the voltage drops occur and correspond to the dark regions of the plasma.

The sheaths form to slow down the electron loss so that it is equal to the ion loss per RF cycle and so the average current to the electrodes is zero. Whereas the heavy ions respond to the average sheath voltage, the light electrons respond to the instantaneous voltage. However, due to the self-biasing, they only cross the sheath during a short period per cycle when the sheath voltage and thickness are near their minimum. During most of the RF cycle the electrons are turned back at the sheath edge, resulting in the sheaths, on the average, being deficient or depleted of electrons. This depletion of electrons in the sheath results in the sheath being dark because of the lack of electron/atom collisions and subsequent relaxation by light emission. In the bulk of the plasma, both ionization and excitation events occur, the latter producing the characteristic glow of the plasma. The lack of electrons in the sheath also increases the impedance of the sheath compared to the plasma bulk resulting in the main voltage drop being across the sheaths. If one of the electrodes is made much smaller in area, then the voltage distribution becomes asymmetric with a much larger voltage drop occurring from the plasma to the smaller electrode, as shown by the dashed curve in Figure 10–8. As described in Chapter 9, this happens because the RF current density must be much higher at the smaller electrode in order to maintain current continuity throughout the system. Therefore the fields must be higher at the smaller electrode.

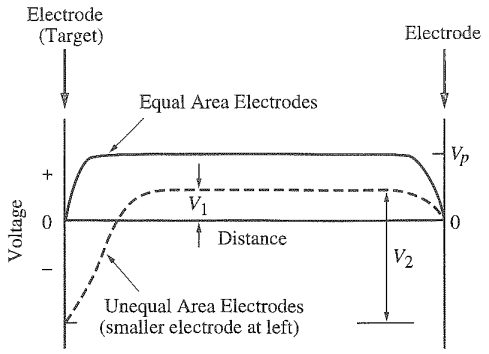


Figure 10–8 Steady-state voltage distribution in RF-powered plasma etch systems. (See Figure 9–27 and the associated text for a more complete discussion of this figure.)

For etching materials other than photoresist (for which O_2 is used) the reactant gases in the plasma are usually halide containing species, such as CF_4, Cl_2, and HBr. Sometimes small amounts of other gases, such as H_2, O_2, and Ar are also added. The high-energy electrons in the plasma can cause a variety of reactions to occur with the reactant gases, including electron-induced ionization, dissociation, recombination, and excitation reactions, as illustrated in Figure 10–9. The plasma will thus contain free electrons, ionized molecules, neutral molecules, ionized fragments of broken-up molecules, and free radicals. For example, in a CF_4 plasma, there will be electrons, CF_4, CF_3 (and other CF_X fragments), CF_3^+ and F. CF_3 and F are free radicals—very reactive species. Typically there will be about 10^{15} cm^{-3} neutral species (1 to 10% of which may be free radicals) and 10^8–10^{12} cm^{-3} ions and electrons. The actual numbers of the different species will depend on the balance between generation and loss or recombination reactions for each.

In standard plasma etch systems, the plasma density (the concentration of ions or electrons in the main plasma, which are virtually equal) and the ion energy as it travels across the sheath to strike the wafer are closely coupled. As the RF power increases, more ions are created, increasing the density, and the sheath voltage increases as well, increasing the ion energy. This coupling comes about because of the capacitive nature of the coupling between the RF current and the plasma [10.3].

10.2.2.1 Plasma Etching Mechanisms

The two main types of species involved in plasma etching are the reactive neutral chemical species and the ions. It is the reactive neutral species—free radicals in many cases, but sometimes other reactive species such as Cl_2—which are primarily responsible for the chemical component in plasma etch processes. It is the ions that are responsible for the physical component. They can work independently, or as we will see, they can work

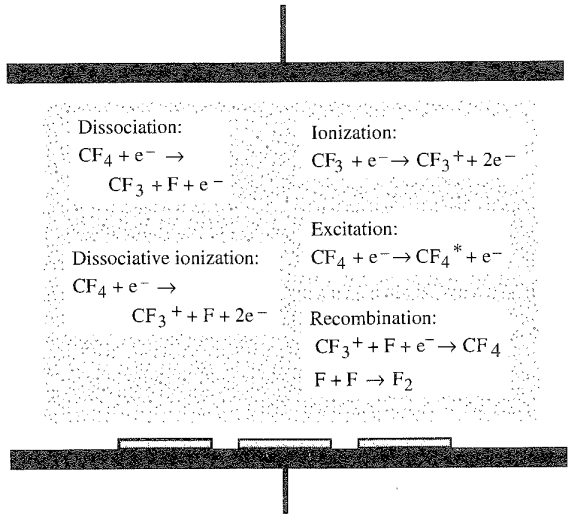


Figure 10–9 Typical reactions and species present in a plasma used for plasma etching.

together in a synergistic manner. When the reactive neutral species act by themselves, the process or mechanism is called chemical etching. Ions acting by themselves can result in physical etching, or sputtering. When the reactive neutral species and ions act in a synergistic fashion, this is called ion-enhanced etching. We will now describe each of these mechanisms.

Chemical Etching Chemical etching of materials in plasma systems is commonly done by free radicals. Free radicals are electrically neutral species that have incomplete bonding; that is, they have unpaired electrons. Examples are the fluorine free radical, F, and the neutral CF_3 free radical. Both of these can be produced by the reaction between free electrons in the plasma and CF_4, for example,

$$e^- + CF_4 \rightarrow CF_3 + F + e^- \tag{10.9}$$

Because of their incomplete bonding structure, free radicals are highly reactive chemical species. Free fluorine, with seven electrons in its outer shell instead of the energetically favorable set of eight, would much rather be bonded to other atoms so that all electrons are paired. It will therefore react quickly with other species to achieve that bonded state. The idea in plasma etching is for the reactive neutral species to react with the material to be etched, Si for example. The byproduct should be a volatile species

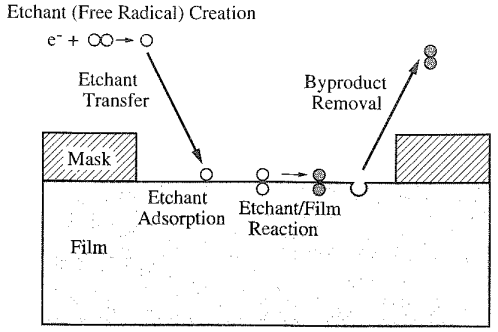


Figure 10–10 Processes involved in chemical etching during plasma etch process.

(that is, easily vaporized or evaporated into the gas phase) that leaves the surface, exposing more Si to be etched. This is illustrated in Figure 10–10. In fact, the fluorine free radical will react with Si via the reaction

$$4F + Si \rightarrow SiF_4 \tag{10.14}$$

to produce the gas SiF_4 (although there may be intermediate steps and species in that reaction). Thus a CF_4 plasma can be used to etch Si. It is important in etch processes that the byproducts of the etch, as well as any unreacted etchants, be removed from the surface for the etching to continue. Figure 10–10 shows a simple etch process, but more complex reactions may occur, involving multiple species and steps.

Gas additives can be used to increase the production of the reactive etch species and thereby increase the chemical etching rate. For example, oxygen is often added to a CF_4 plasma. The oxygen reacts with dissociated CF_4 species (CF_3 or CF_2, for example), which reduces the recombination of those species with F, thereby increasing the amount of free F. This enhances the chemical etching. However if too much O_2 is added, the etchant becomes too diluted or the surface is oxidized by the O_2, and the etching decreases.

We saw earlier that wet chemical etching occurs isotropically. The chemical component of plasma etching, when acting by itself, also occurs isotropically. Two factors lead to this behavior: an isotropic arrival angle distribution and a low sticking coefficient, with the latter usually more important. When we model and simulate plasma etching later in this chapter, we will closely follow the approach we used for deposition processes. That is, we will consider the incoming fluxes of species from the plasma to the wafer surface including the angular distribution of the various fluxes. This will be quantified by the $\cos^n\theta$ distribution, where a value of 1 for n describes an isotropic distribution (that is, coming in at all angles equally so that the normal component to the surface follows a $\cos\theta$ distribution) and a value of n larger than 1 describes a more directed distribution toward the surface. As in the deposition models we assume that all un-ionized

chemical species arrive at the wafer surface at all angles. Since the free radicals are neutral species, the flux of the chemical component of plasma etching has an isotropic arrival angular distribution, or n equals 1.

We also consider the sticking coefficient of the species, as defined by Eq. (9.25), and repeated here.

$$S_c = \frac{F_{\text{reacted}}}{F_{\text{incident}}} \tag{10.11}$$

A high sticking coefficient (equal to 1) means that the reaction, and etching, occur the first time the species strikes the surface. A lower sticking coefficient means that before the species reacts, it can leave the surface—usually assumed to occur at a random angle—and strike the surface somewhere else. One might guess that free radicals would have very high sticking coefficients, since they are so reactive. However, as in deposition processes, the chemical reactions that occur in chemical etching often require multiple steps, and multiple atoms or species, all in the right proportions. In addition, the availability of suitable sites on the surface, competing reactions, and surface recombination of species can be issues. These issues result in most free radicals having effective sticking coefficients that are low, on the order of 0.01 to 0.05 for F etching of Si, for example. Thus the chemical etch species bounce around a lot and tend to maintain random angles of flight even under masks and other surface features. The fluxes for chemical species are illustrated in Figure 10–11(a).

These two properties of the reactive neutral species in plasma systems—isotropic arrival angles and low sticking coefficients—result in the fact that purely chemical etching acts in an isotropic or near-isotropic manner. Only if the sticking coefficient is relatively large would shadowing by the topography and etch mask become important. (Sticking coefficients on the order of 0.1 may actually occur for some purely chemical

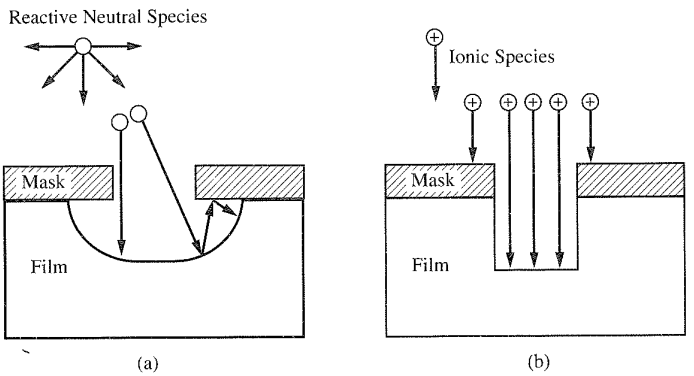


Figure 10–11 Fluxes of species in plasma etching: (a) fluxes of reactive neutral chemical species (such as free radicals), with a wide arrival angle distribution and low sticking coefficient; (b) fluxes of ionic species, with a narrow, vertical arrival angle distribution and high sticking coefficient (assumed equal to 1).

plasma etching cases [10.4].) In that event somewhat more anisotropic etch profiles can result. In general, though, pure chemical plasma etching results in isotropic or near-isotropic etching. We must reiterate that this is only for the chemical component acting independently, or for a purely chemical etch process. As we shall see shortly, the reactive chemical species and the ions can work in a synergistic manner so that the etching is not isotropic at all.

Since free radicals etch by chemically reacting with the material to be etched, the process can be very selective, as discussed earlier for wet chemical etching. Thus any chemical component to plasma etching will be relatively selective, as compared to the physical component.

In some cases, reactive neutral species other than free radicals may be involved in chemical etching. An example is in the plasma etching of Al by the Cl_2 molecule. For the same reasons given above, these reactive neutral species would have the same etching properties as the free radicals, namely isotropic profiles and selective chemical etching.

Physical Etching The other main species that participate in plasma etching are the ions that are present. Because of the voltage drop between the plasma and each electrode and the resulting electric field across each sheath, positive ions will be accelerated toward each electrode. Since the wafers are placed on one of the electrodes, then the ionic species such as Cl^+ (or Ar^+ in a purely physical sputter etch system) will be accelerated to the wafer surface. This striking of the wafer surface results in the more physical component of etching. The flux of ions toward the wafer surface is much more directional than the flux of neutral free radicals because of the directionality of the electric field from the plasma to the surface of the wafer. Thus the etching will be much more directional and more anisotropic. In fact it is usually assumed that all the ions arrive normal to the wafer surface, as shown in Figure 10–11(b), meaning that n is very large in the $\cos^n\theta$ distribution. Because the ions etch by more physical processes, rather than chemical, the etching will be much less selective. It is also assumed that once each ion strikes the surface, it does not strike the surface again somewhere else (or if it does, it does not have enough energy to do anything). This means that the sticking coefficient is effectively one. This also leads to more anisotropic behavior.

Ions can etch material in plasma systems by physical sputtering. Sputtering was discussed in Chapter 9, where we saw how sputtering of elemental, compound, or alloy targets was utilized to deposit material on the surfaces of wafers. We also saw how sputtering of the wafer itself could be done during a simultaneous deposition of material in order to planarize or fill features. Sputtering occurs by ions, accelerated by an electric field, striking the atoms on the surface of a material and physically dislodging them. The sputtering process is often described by a sputtering yield defined in Chapter 9 as the number of atoms or molecules ejected from the target per incident ion and is a measure of the sputter efficiency. As discussed in Chapter 9, the sputter yield is a function of incident angle, and it is this property that allows for planarization and filling in bias-sputtering deposition techniques. Because of the physical nature of sputtering by ion species arriving normal to the wafer surface, very anisotropic etching occurs, with profiles similar to that shown earlier in Figure 10–3(c).

When ionized reactive species such as Cl^+ and CF_3^+ are involved in sputtering, chemistry can also be important in determining the effective sputtering yield since these ions do have the potential of reacting chemically with surface atoms. Such effects can enhance the sputtering yield compared to Ar^+ sputtering, although the magnitude of these effects is usually small compared to ion-enhanced etching, which involves both neutral and ionized species and which we describe next.

Ion-Enhanced Etching Another way that ions participate in plasma etch processes is through ion-enhanced etching. It was recognized many years ago that the ions and reactive neutral species do not always act independently in an etch process. This was easily deduced when it was observed that the etch rate measured is not always just the sum of the two components acting independently; in many cases it is much higher. As a classic example of ion-enhanced etching, Figure 10–12 shows the etch rate of silicon as XeF_2 gas and Ar^+ ions are introduced to the surface. First XeF_2 gas (not a plasma, so there is no physical component at all) is introduced into the chamber and very little etching occurs. Then an Ar^+ ion beam is directed toward the surface and the etch rate increases ten times. Finally the XeF_2 gas is turned off so that only ion sputtering occurs, and the etch rate becomes almost zero. The chemical and physical species are obviously working in a synergistic manner.

This synergistic process is also demonstrated in the shape of the etch profiles that result. In ion-enhanced etching, both chemical and physical components are acting, but the profiles are not just a linear combination of isotropic chemical etching and anisotropic physical etching as illustrated in Figure 10–3(b). Instead, the profile for ion-enhanced etching is much more like the case for physical etching acting alone, as in Figure 10–3(c). If the chemical component in the etch system is increased, the vertical etching is increased but not the lateral etching, which is not what would be expected from chemical etching. In addition, the etch rate in many cases is much higher (>10 times) than that calculated from the measured ion flux assuming normal sputtering yields in the absence of neutral species; hence it is not likely that the observed plasma

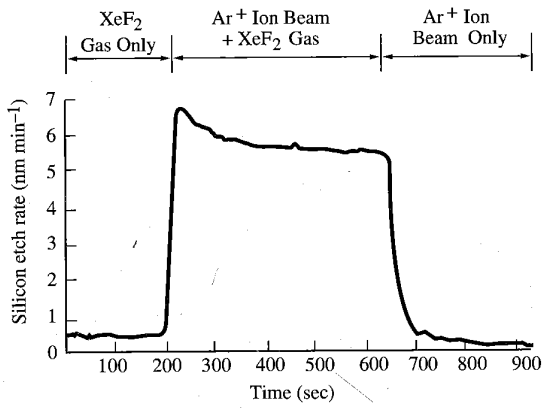


Figure 10–12 Etch rate of silicon as XeF_2 gas (not plasma) and Ar^+ ions are introduced to the silicon surface. Only when both are present does appreciable etching occur, illustrating synergistic etching behavior between chemical and ionic species. (After [10.5, 10.6].)

etch behavior in these cases is due simply to an ionized chemical species, such as CF_3^+. Observed saturation effects, which we will describe later, also support this dual component model involving both neutral and ion species.

In this type of etch process, the chemical components (the free radicals or other reactive neutral species) and the physical components (the ions) in the plasma somehow work together to etch the material, with both components required at each point for the process to work. In addition, because of the required chemical component and the chemical processes that occur, good selectivity can be obtained. So not only can one achieve very anisotropic etch profiles, but also a high degree of selectivity as well. This is something that is very difficult to achieve when the two components work independently. In addition, high etch rates can be achieved. Thus gas chemistries and etch conditions are usually chosen to promote ion-enhanced etching and to suppress independent chemical and physical etching.

There have been many explanations and mechanisms proposed for this type of cooperative etch process. In most of the models the ion bombardment enhances one of the steps of the chemical etch process, such as surface adsorption, etching reaction and formation of byproduct, or removal of byproduct or unreacted etchant. For example, the ion bombardment causes damage (breaking bonds, etc.) which makes the surface more apt to chemically react with the radical. Or the ion bombardment accelerates the formation of volatile byproducts. Or the ion bombardment may dislodge or sputter away etch byproducts which would otherwise tend to stay on the surface and impede the etch process. The removal of byproducts can also include secondary, or indirect, byproducts. It has been found that a chemically inert residue is often formed from etching or sputtering photoresist. In addition, species from the plasma may deposit on the surface. These layers that are formed or deposited, whatever their source, inhibit chemical etching either by physically blocking the chemical etch species or by reacting with them and depleting them. But the inhibitors can be removed, or prevented from forming, by ion bombardment, leading to anisotropic etching. In fact, these inhibitors are often intentionally produced on the surface by providing sources in the gas which enhance their formation so that more directional etching is achieved. In a particular ion-enhanced etch process, one or more of the above mentioned mechanisms (damage induced, inhibitor removal, etc.) may be occurring.

Whatever the exact mechanism for ion-enhanced etching, and it is often difficult to distinguish between them, directional etching occurs because of the directionality of the ions. The enhancement of the process, whether of enhancement of etch reaction or of inhibitor removal, occurs only where the ions strike the surface. Since the ions are striking normal to the wafer surface, the enhancement will occur normal to the wafer surface. This directional enhancement will result in directional, anisotropic etching. This is illustrated in Figure 10–13(a) for the case when the chemical etch reaction is enhanced by ion bombardment. It is assumed in this figure that etching by the neutral (radical) species alone without ion bombardment, the so-called "spontaneous chemical etching," is negligible, leading to no lateral etching at all. In Figure 10–13(b) the process of inhibitor formation and removal is illustrated, which also results in anisotropic etching. Ion bombardment on the horizontal surfaces removes the inhibitor there but leaves the sidewall inhibitor to protect against lateral chemical etching.

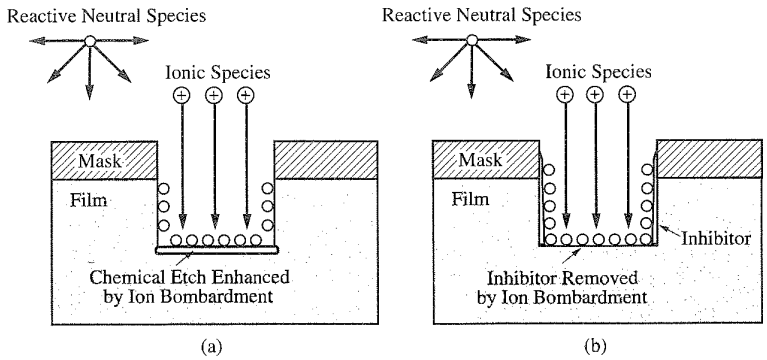


Figure 10–13 Illustration of ion-enhanced etching. In (a) the chemical etch reaction is enhanced by ion bombardment. In (b) an inhibitor is formed which is removed by ion bombardment, allowing chemical etching to proceed. In both cases, anisotropic etching results.

Even with ion-enhanced etching and no spontaneous chemical etching, the slope of the resulting sidewall may not always be exactly vertical. For example, the ion flux may not always be perfectly normal to the wafer surface and can result in slightly bowed sidewalls. Also, when inhibitors form or are deposited during etching, sloped sidewalls can result if the inhibitor formation or deposition rate is very high relative to the etch rate of the inhibitor and substrate, as demonstrated in Figure 10–14. Changing the relative deposition/etch rates by changing the etch chemistry or plasma conditions can be used, in fact, to tailor the desired slope if less than vertical sidewalls, without any undercutting, are desired in specific applications. But the important point still is that very anisotropic etching can be achieved when ion-enhanced etching is occurring.

Selectivity occurs in ion-enhanced etching because chemical reactions are involved in this type of etch process, and chemical reactions are by nature selective. To achieve the needed selectivity, one must choose the proper etchant gas (either F, Cl, or Br based, with or without oxygen or other species added, for example) and other etching conditions which favor the etching of the film over the mask material and the underlying films. With ion-enhanced etching, one can achieve good etch selectivity while still obtaining very directional etch profiles.

10.2.2.2 Types of Plasma Etch Systems

Different types of plasma etch systems and modes of operation have been developed to etch films in semiconductor structures which take advantage of the different species present in a plasma and the different etch mechanisms that can operate. We will now describe these different etch systems or modes, focusing on the advantages and disadvantages of each.

Plasma Etching in Barrel Etchers One of the earliest plasma systems developed was the barrel etcher, named because of its shape, and is shown in Figure 10–15. In this

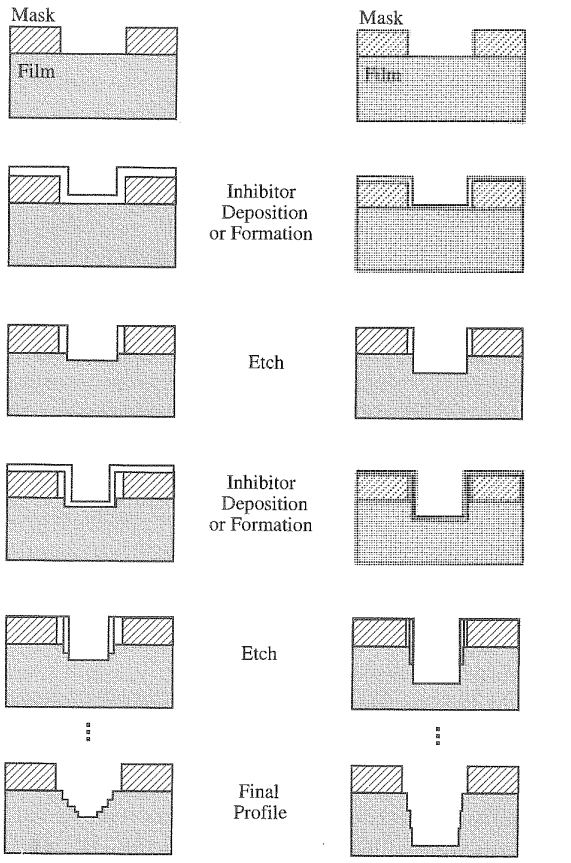


Figure 10-14 Illustration of the effect of the inhibitor deposition rate-to-etch rate ratio on the resulting sidewall slope. The different relative rates of inhibitor deposition and etch rate lead to different slopes of sidewall [10.4].

configuration, the electrodes are curved, wrapping around the quartz tube and the wafers are placed vertically in a substrate holder which sits in the middle of the tube. After evacuating the tube, reactant gases are introduced and RF power is applied to the electrodes. A plasma is generated but is kept away from the wafers themselves by the use of an etch tunnel, a perforated metal shield between the wafers and the electrodes. Reactant chemical species, such as F, diffuse from the plasma where they are generated, through the shield and to the wafers. There they can etch the wafers. Because the plasma and ions are kept away from the wafers by the shield, and the wafers do not sit

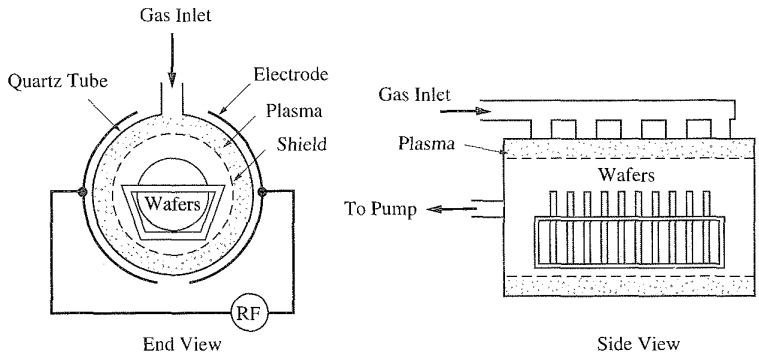


Figure 10–15 Schematic illustration of barrel etch system, in which purely chemical plasma etching occurs.

on either of the electrodes, no ionic bombardment of the wafers occurs. Thus the etching is purely chemical.

Because the etching is purely chemical, it can be very selective. It also causes very little damage to the surface because of the lack of ion bombardment. Also, it is a rather simple process and the throughput can be high with the closely stacked wafers. But because the etching is purely chemical, it is isotropic. This results in significant undercutting or mask bias. In addition, the etch uniformity across each wafer is not very good, especially from the outside of each wafer to the inside, because of the long diffusion path for the reactive species between the wafers.

As a result of the isotropic etching and nonuniformities, barrel etchers are used only for noncritical etch steps. In fact they are mostly limited to photoresist removal, using an O_2 plasma, where masking is not even done. This is called photoresist ashing, or just ashing, since the reaction between the oxygen and the hydrocarbon-based photoresist is the same as burning, producing CO_2 and H_2O. A long overetch is usually done in barrel etchers to compensate for the nonuniformities. This is allowable because of the high selectivity of the chemical etching processes used in these systems.

Similar to barrel etchers in principle but a little different in configuration are "downstream plasma etchers." Microwave or RF power can generate a plasma upstream in one chamber, while the wafers are placed in a separate but connected chamber "downstream" from the plasma. The neutral reactive chemical species can diffuse from the plasma in one chamber to the wafers in another, while the ions do not. Hence purely chemical etching occurs, resulting in good selectivity and little ion bombardment damage, but in isotropic etching. These are also called "afterglow etchers." These are also used for nonmasked etching steps, such as removing silicon nitride masks used for LOCOS processes.

Plasma Etching in Parallel Plate Systems—Plasma Mode Parallel plate etch systems, as shown earlier in Figure 10–7, are commonly used for plasma etching of films. This is very similar to the PECVD system described in Chapter 9, except that etch gases

are used instead of deposition gases. Having this configuration with the wafers sitting on one of the electrodes and directly facing the other electrode gives much better etch uniformity than the barrel etcher design. In addition, it allows for the bombardment of ions on the wafer surface. The voltage distribution that develops from the RF source was shown in Figure 10–8. The voltage distribution is symmetric for equal area electrodes, with voltage drops across each sheath to the electrodes in the range of 10–100 V. As discussed earlier, this voltage drop accelerates the ions to the surface of the wafer, which is sitting on one of the electrodes. This results in a physical component to etching due to ion bombardment, as well as a chemical component due to the neutral reactive species. The physical component means that anisotropic etching is possible.

In symmetric parallel plate systems, with a gas pressure from 100 mtorr to 1 torr, the physical etch component is usually not very significant compared to other parallel plate configurations. In fact, the ion bombardment is even less since the bottom electrode is the grounded electrode in this setup. Since the chamber walls are also usually grounded and thus electrically connected to the bottom electrode, this effectively increases the size of that electrode. This means that the voltage drop from the plasma to that electrode is even smaller, as was shown in Figure 10–8 (the wafers would sit on the right side electrode for the dashed curve in that figure), and weaker ion bombardment would occur. One problem with etching in this mode is that with the large voltage drop to the top electrode, significant sputtering of that electrode can occur, causing contamination problems.

When the electrodes themselves are of equal area and the bottom electrode on which the wafers sit is grounded, making the effective size of the bottom electrode larger, then the etching is said to be operating in the plasma mode. Both chemical and physical components of etching are present in this mode, but as stated earlier the physical component is often rather weak, and highly directional etching is hard to achieve. However, the etching is still rather selective, and better uniformity and etch directionality are achieved compared to barrel etchers.

Plasma Etching in Parallel Plate Systems—Reactive Ion Etching Mode The parallel plate etch system can be modified to enhance the physical component of the etching by increasing the energy of the ions bombarding the wafers. This is achieved by having the electrode on which the wafers sit be smaller than the other electrode. Not only can the lower electrode be made physically smaller than the upper electrode, but the upper electrode can be the one that is grounded and electrically connected to the chamber walls. This makes the upper electrode effectively much larger. The bottom electrode is now much smaller than the upper electrode, and the voltage distribution is as shown in Figure 10–8, with the wafers on the left-hand side in this case. A much larger voltage drop occurs from the plasma to the wafers, in the 100–700 V range, resulting in more energetic ion bombardment of the wafer. With stronger ion bombardment, more directional etching can be achieved.

When the wafers sit on a smaller electrode, the etching is said to be operating in the Reactive Ion Etching or RIE mode. A better name for this might be "reactive and ion etching," since two types of species, the neutral reactive species and the ion species, are involved. In the RIE mode, both the physical ion component and the chemical reactive neutral component are important. They can act separately, but as we described earlier,

gas chemistries and etch conditions are usually chosen so that they act together in a synergistic manner.

Even more directional etching can be attained by lowering the gas pressure. A lower pressure means fewer ion collisions during transit across the sheath in the gas phase, resulting in a more directed ion flux toward the wafer. Decreasing the pressure also increases the sheath voltage (that is, the voltage drop from the plasma to the wafer), which increases the ion energy. But decreasing the pressure can also decrease the plasma density (fewer gas atoms available to be ionized or converted into radicals), so that the pressure cannot be lowered too much. Reactive ion etch systems generally operate in the 10–100 mtorr range.

RIE systems can look very much like parallel plate plasma mode etchers (Figure 10–7) except that the RF power is supplied to the other electrode (to the bottom wafer holding electrode so that the top electrode is grounded with the chamber), and the size of the bottom electrode in the RIE case may be somewhat smaller. Single-wafer systems are often used for better control over each wafer. To increase the batch size other configurations have been used historically, such as with several electrode pairs stacked vertically above one another in a single chamber. Magnetic fields are often added to enhance plasma density and lower operating pressures.

Vertical RIE systems have also been developed. Here the wafers sit vertically against the inside electrode, which is often hexagonally shaped—hence the name, hexode batch etcher. The outside barrel chamber acts as the outer electrode. With its large size compared to the inner electrode, etching in the RIE mode can occur. Triode systems have also been developed. These are similar to the bias-sputtering systems discussed in Chapter 9, and add a second separate RF power supply to control the bias of the wafers. This separates plasma generation from ion energy, the latter being determined by the wafer bias. However, this independent control is only partially achieved using this technique with two capacitively coupled sources, and the plasma can suffer from nonuniformity. Also, the additional internal electrodes can cause contamination problems.

Operating in the RIE mode results in stronger ion bombardment than in the plasma mode, and a more physical component in the etching. While this leads to more directional etching, the higher enegy ions striking the surface can cause radiation or lattice damage, charging, and trenching. These will be discussed in more detail later. In addition, the higher physical component can lead to less selectivity. However, by choosing proper etch chemistries and utilizing ion-enhanced etching, an acceptable degree of selectivity can be achieved in these systems while still being able to etch material with excellent dimensional control. As a result, RIE etching is used extensively in the integrated circuit industry.

Etching in High-Density Plasma Systems A new type of plasma etch system that is becoming more popular is the High-Density Plasma (HDP) etch system. Like triode RIE etchers, these systems separate the plasma density and ion energy by using a second excitation source to control the bias voltage of the wafer electrode. But instead of using two capacitively coupled RF sources as in the triode systems, a different type of source for the plasma is used. The plasma generation source is non-capacitively coupled and produces a very high density plasma ($10^{11} - 10^{12}$ ions cm^{-3}) without itself generating a large sheath bias, utilizing a variety of methods. These include Electron Cyclotron

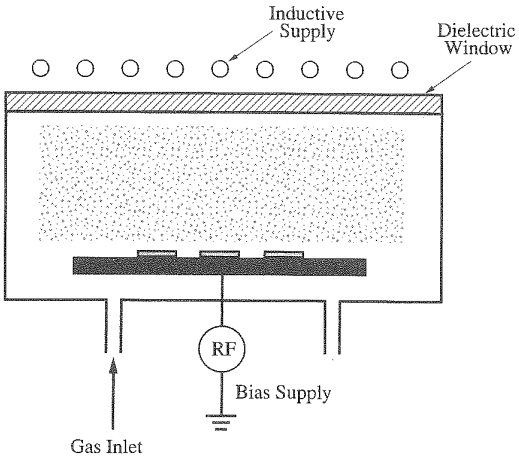


Figure 10–16 Schematic diagram of High-Density Plasma (HDP) etch system. This configuration is powered by an Inductively Coupled Plasma (ICP) source which produces and controls the high-density plasma. The RF wafer bias independently controls the ion energy.

Resonance (ECR) and Inductively Coupled Plasma (ICP) sources. The ICP method is becoming popular due to its much simpler design and equipment. A schematic diagram of an ICP high-density etch system is shown in Figure 10–16.

HDP etch systems are similar to the HDPCVD deposition systems described in Chapter 9. In those systems, the high-density plasma is produced using the same type of power sources, producing the reactive species for deposition as well as ionic species. A separate power source is used to bias the wafer in order to control the ion bombardment on the surface that is used for simultaneous biased sputtering. In Figure 9–45 in Chapter 9 we illustrated an ECR-powered HDPCVD system, which would be similar to an ECR HDP etch system. Similarly, an ICP high-density plasma CVD system would be very similar to the etch system shown in Figure 10–16.

In HDP etch systems, very high plasma densities are produced compared to standard RIE etch systems. Therefore lower pressures can be utilized, in the 1–10 mtorr range, while still achieving high ion fluxes and etch rates. A lower gas pressure means that even more directed etching can be achieved because of fewer gas-phase collisions in the sheath. In standard RIE systems, increasing the power increases the plasma density (and ion flux), allowing for higher etch rates, but also increases the sheath bias and ion energy, leading to more substrate damage. One could raise the pressure in standard RIE systems to produce a higher ion density while keeping the ion energies low, but the etching would become more isotropic due to more gas phase collisions in the sheath. In HDP systems, with the decoupling of the plasma density and ion energy, many ions can be produced at low pressures with low ion energies. As a result, ion bombardment damage can be kept low while maintaining high etch rates and good anisotropic etching. In addition, a lower ion energy usually gives better selectivity than higher ion energy.

Etching in HDP systems is similar to etching in the RIE mode, with a large physical component due to the ions combined with a chemical component involving the reactive neutrals (although the ionized species may themselves also contribute an appreciable chemical component to etching in these systems [10.6]). Therefore both directional etching and reasonable selectivity are possible. But in the HDP systems, the ion flux is

even higher than in standard RIE systems, resulting in higher etch rates at lower pressures. And the higher ion flux, along with the lower pressures, leads to even more directional etching. The separate power sources allow independent control of ion density and ion energy, providing more flexibility in the etch process.

Sputter Etching and Ion Milling The sputter etching and ion milling techniques are at the other end of the spectrum from purely chemical systems; these are purely physical etching methods. We discussed sputtering in detail in Chapter 9, in which sputtering of a target was done to produce species for deposition. We also described how simultaneous sputtering of a film being deposited can be done to help with planarization of the film as well as filling in high aspect ratio spaces or holes. In sputter etching, physical bombardment of films with chemically inert ions, such as argon ions, produced in a plasma is used to etch the film. Figure 10–17 shows an RF sputtering system. The size of the cathode is minimized relative to the anode, which includes the chamber walls, to maximize the ion bombardment. The ion energies in these systems are large, greater than 500 eV, in order to get appreciable etching of films. While these systems are not normally used to etch thin films today, they do illustrate some important points about physical etching, and they have been used in the past.

The main advantages of sputter etching are that all materials can be etched this way and the etching is very directional, or anisotropic. The principle disadvantage is that selectivity is poor because little or no chemistry is involved. The sputtering rate, which is the same as the etch rate, depends on the sputtering yield. This is the efficiency of dislodging or removing an atom from the sputtered material by each impinging ion. The sputter yields of most materials by argon ions are fairly similar. This means that most materials can be sputter etched with about the same efficiency and that this method can often be used when other etch methods do not work. But this also means that the selectivity will not be good. The best selectivity ratios achievable are only about two-to-one. The sputtering yield is usually a function of incident angle. This can lead to faceting, which can be good or bad depending on the desired topography. The fact that

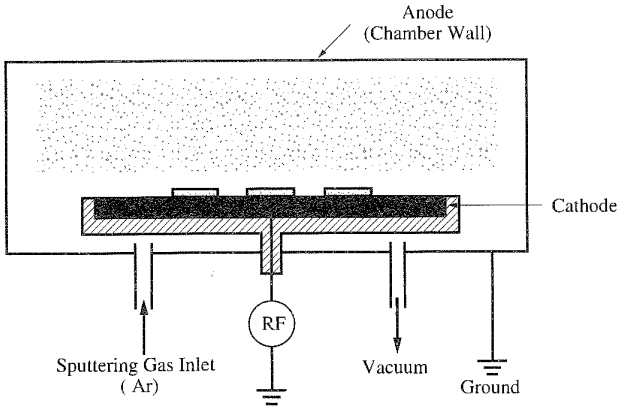


Figure 10–17 Illustration of sputter etch system.

all the etching is done by the highly directional ions means that highly directional etching can occur.

There are a few other problems associated with sputter etching. These would also be problems with any other method that utilizes energetic ion bombardment, such as reactive ion etching, but they are worse for sputter etching which is purely physical. One of these problems is called trenching, or "microtrenching." Along the sides of masking layers, some of the ions will glance off the sides and strike the films at the bottom corner. Hence these corner regions will receive a higher ion flux than if the sidewall were not there. This causes enhanced etching at the edges, resulting in trenches, as illustrated in Figure 10–18(a). Another problem is in redeposition. In a nonselective physical process other layers will be sputtered as well. These sputtered materials can be redeposited elsewhere, including in the region that is being etched. This is illustrated in Figure 10–18(b).

A third problem is lattice or radiation damage. During sputtering, the surface is subjected to bombardment of energetic ions, electrons, and photons. Ion bombardment can lead to displaced atoms and implanted atoms below the surface. Radiation damage can lead to electron traps being produced in gate oxides. Another problem is surface charging. This is associated with the bombardment of the wafer surface with charged ions and electrons. Because of nonuniformities that can exist in the plasma, dielectric materials can be charged by the local differences between ion and electron fluxes. This can lead to tunneling current damage to thin dielectrics. It can also lead to etch profile distortion. This is because the paths of the incoming charged ions will be affected by the presence of the charged regions, either positive or negative, as shown in Figure 10–18(c), and lead to uneven etching.

In ion milling, some of the problems associated with charged species are eliminated. In this technique, a confined plasma is used to generate a broad beam of Ar^+ ions. This is done by using a set of grids to extract and accelerate ions from the plasma chamber toward the target wafer to be etched. A separate electron source is used to supply electrons to neutralize the ion beam. This setup allows for more independent control over

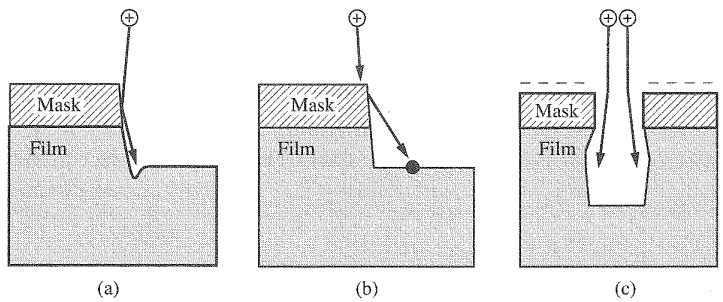


Figure 10–18 Problems associated with sputter etching (or any etching that has a high degree of physical/ionic etching): (a) trenching at bottom of sidewalls; (b) redeposition of photoresist and other material; (c) charging and ion path distortion.

the ion energy and ion density, which are usually closely coupled in a normal plasma system. An amount of reactive species may also be added to, or replace, the Ar in ion milling systems. These species may include CF_4, CCl_4, or O_2. This will add some amount of selectivity to the process. This method is called "Reactive Ion Beam Etching," or RIBE.

Ion beams can also be focused and then used to sputter, or etch, very small areas. This is called the Focused Ion Beam (FIB) technique. It can be used in advanced lithography applications to expose/etch resist instead of electron beams. It is commonly used to repair masks. It can also be used to cut or etch regions of films and substrates to expose cross sections for observation. The ions in this are often He^+, N^+, Ar^+, Xe^+, and most commonly Ga^+.

10.2.2.3 Summary of Plasma Systems and Mechanisms

Figure 10–19 summarizes the different types of plasma etch systems in terms of some of the operating and etch properties. Of course there are wide variations in the properties for different types of systems, and overlaps occur between the different types of etching. For example, the ion energy in HDP etch systems can be the same or lower than in RIE systems, and the selectivity can be better as well. And the selectivity in both HDP and RIE systems can be very good while still achieving anisotropy by choosing the right etch chemistries and utilizing ion-enhanced etching. But the basic trends shown are useful in understanding the differences in the various types of plasma etching systems.

Figure 10–20 summarizes some of the different etch processes that occur in plasma etching which affect the two-dimensional etch profile. Not all the processes would occur simultaneously of course—inhibitor formation and trenching would not just occur on one side with undercutting on the other, for example, but the figure does illustrate the many complex phenomena that can occur during plasma etching.

We have described the similarities between etching and deposition systems several times. PECVD systems are similar to parallel plate etch systems operating in the plasma

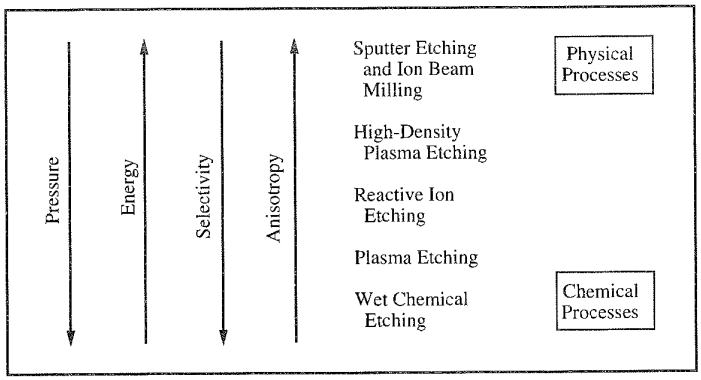


Figure 10–19 Summary of trends of different etch systems.

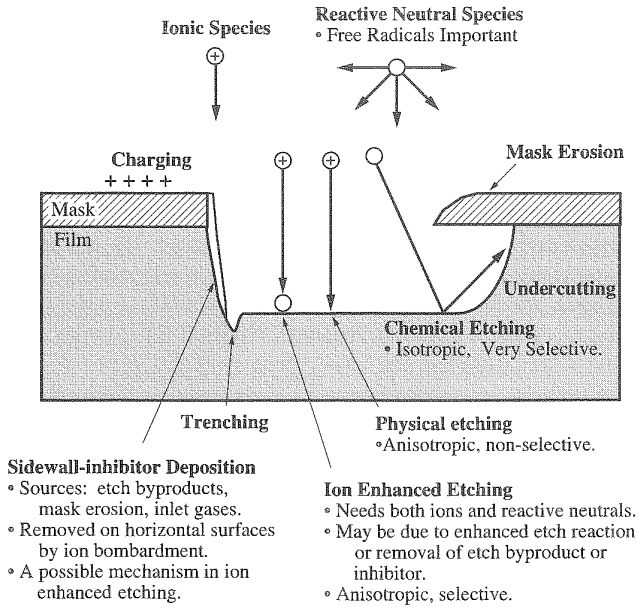


Figure 10-20 Summary of different processes occurring in plasma etching. (After [10.4].)

mode. There are reactive neutral species for either deposition or etching, with a minimum of ionic bombardment in either case. There are also similarities between deposition systems that utilize wafer sputtering, such as bias-sputtering and HDPCVD systems, and reactive ion and HDP etch systems. Bias-sputtering and HDPCVD deposition systems combine deposition of materials and sputter etching of the surface. Both of these processes also occur in reactive ion etching and HDP etching systems. Etching of the wafer surface is done using ion bombardment as well as chemical processes. But "deposition" is also commonly done in these etch systems in which inhibitor layers are "deposited" on the surface to promote anisotropy. These inhibitors form from byproducts of the etch process and the chemistry of the etch gases is often chosen to enhance their formation. This is called inhibitor deposition. So both deposition and etching occur in these deposition systems and etch systems. It is only the degree and type of deposition or etch that is different in specific systems. Many of the same concepts and models apply to both.

10.3 Manufacturing Methods

In this section we will describe some of the specific manufacturing methods used in plasma etching and some of the important issues related to manufacturing. We will limit ourselves to plasma etching here because of its dominance in current integrated circuit

manufacturing. Limited information on wet chemical etching was given in Table 10–1. Most films in current semiconductor manufacturing are etched using parallel plate plasma systems in the RIE mode or using HDP systems.

10.3.1 Plasma Etching Conditions and Issues

The main parameters that can be directly controlled during plasma etching in standard parallel plate etch systems are RF power, pressure, gas compositions, and flow rates. (Some systems allow control of the self-bias, or sheath voltage, and the RF power is adjusted so that a particular bias is achieved.) Power densities of 0.1 to 5 W cm^{-2} are commonly used in parallel plate, capacitively coupled RF plasma systems, usually at 13.56 MHz with excitation voltages of 100–1500 volts RMS. Increasing the RF power increases the plasma density. In standard parallel plate systems, increasing the RF power also increases the self-bias or sheath voltage, which is the voltage drop between the plasma and the electrodes. This increases the ion energy. Ion energies produced are in the 10- to 700-eV range. Higher ion energy results in a higher sputtering yield which results in a higher physical component of etching but can also result in more surface damage.

In HDP systems, with separate power sources for the plasma and wafer bias, one can achieve high plasma densities and high fluxes, without necessarily getting high ion energies. In those systems, the power density on the wafer is in the 0.1–3 W cm^{-2} range. The resulting ion energies can range from 10 to 500 eV. A higher source power is used to generate the plasmas, in the 3- to 10-W cm^{-2} range.

The pressure in plasma etching systems ranges from 1 mtorr to 1 torr, with standard RIE systems operating in the 10- to 100-mtorr range and HDP systems in the 1- to 10-mtorr range. Increasing the pressure causes more gas-phase collisions to occur, decreasing the directionality of the etching. Increasing the pressure also increases the plasma density (since there are more atoms or molecules to begin with), but only up to a point. Above a certain pressure, depending on the etch system and operating parameters, the collisions between the gas molecules and electrons limit the energy of the electrons and thereby limit the ionization rate. Thus the plasma density decreases at higher pressures. In addition, increasing the pressure lowers the sheath voltage [10.7], thereby lowering the ion energy.

The etch rate, to first order, is proportional to the flux of active species, which would be the neutral chemical species and the ionic species. The flux of ionic species equals its concentration times its velocity. In plasma systems, increasing the power increases the plasma density and therefore increases the ion concentration in the plasma. The ion velocity in the plasma (by the electron temperature in the plasma) does not change to first order with a change in power at the sheath edge and determined. (The velocity across the sheath does increase as the power is increased in a parallel plate system, or as the wafer bias power is increased in a high-density plasma system. But the concentration of ions decreases proportionally across the sheath, so as to keep the flux constant across the sheath [10.11].) Thus the ion flux, being equal to the product of the ion concentration and velocity in the plasma at the sheath edge, would increase with power. The effect of increasing the pressure on the ion flux is less clearcut, and different systems show different relationships.

The etch rate due to the ion component would be proportional to the ion flux, as well as to the sputtering yield, or the effective sputtering yield in ion-enhanced processes as we will discuss later. The effective sputtering yield itself can be dependent on ion energy (approximately proportional to the square root of the energy), which for parallel plate systems would be dependent on the RF power. In parallel plate systems, increasing the RF power increases the ion energy which increases the sputtering yield. Likewise, increasing the pressure can decrease the ion energy and sputter yield, decreasing the etch rate with pressure. In HDP systems, one can change the ion energy and thus the sputtering yield, independent of the plasma density and flux, by changing the separate bias voltage control. One can thus increase the physical etch rate by directly increasing the bias voltage.

One might assume that the chemical flux would be increased by increasing the power, which would increase the plasma density, and the generation of reactive neutral radicals. However, perhaps due to the complex chemistry occurring, this is not often observed. Generally the chemical flux increases with pressure. Thus the etch rate which depends on this flux generally increases with pressure. The behavior of the overall etch rate as a function of pressure can get quite complicated when the etching depends on both the chemical and ion components. A more detailed examination of chemical and ion fluxes and how they relate to the etch process will be given in the models and simulation section. In general, decreasing the pressure too far—below about 10 mtorr for standard RIE systems and about 1 mtorr for HDP systems—will result in too low a plasma density and too low an etch rate for practical purposes.

The gas composition—which gases are used and in what proportions—and flow rates can also be controlled and are used to help achieve the desired etch properties. We will discuss different gas compositions used to etch various films below. The flow rate often has a minor effect on etch rate. Increasing the flow rate will initially increase the etch rate by producing more reactive species. But the actual supply of reactant species depends on the balance between generation and loss of species in the plasma, and a steady state is reached where the etch rate will become independent of gas flow. If too high a flow rate occurs, the etch rate may actually decrease since the residence time of the gas may become less than the lifetime of the active species produced in the plasma.

Usually the temperature of the etch system, including the wafer, is not intentionally raised during etching. The plasma supplies the energy for the etch process and heating the gas or wafer would not significantly increase the etch rate or improve the process. An exception to this is in aluminum etching in which the system, including the chamber walls, is heated to 35–65°C to help keep the species volatile and assist in the removal of the byproducts. Without intentional heating, the wafers may still heat up due to the plasma processes and ionic bombardment—temperatures of 90–100°C can be encountered on wafer surfaces. It is important to control and minimize the temperature rise on the wafers. Sidewall inhibitor deposition actually decreases as the temperature goes up in most cases, resulting in less directional etching. The trend in recent years is for better control of temperature and removal of heat from the wafers by developing better wafer chuck systems, including electrostatic chucks, which uniformly extract heat from the wafers.

At the beginning of this chapter we mentioned several desired properties or qualities of etch processes. These included selectivity, etch profile control, uniformity,

minimal substrate damage, and low contamination. We now consider how these issues relate to manufacturing concerns.

Selectivity involves the chemical component of etching. It is achieved primarily by choosing the particular chemical reactants, in the proper proportions, that favor the etching of one material over another. Changing the plasma conditions, such as the wafer bias and sheath voltage, can also affect selectivity in some cases. Achieving selectivity will be discussed later in the context of some common etch chemistries used for films in integrated circuit fabrication.

Etch profile control means obtaining the desired shape of the etch profile. For most applications in today's technology this means obtaining vertical, or near-vertical, sidewalls. Directional etching, either with ion-enhanced mechanisms or physical sputtering, is basically due to the ion or physical component and can be enhanced by increasing the ion directionality (by decreasing the pressure, for example), increasing the ion flux (by increasing the RF power, for example), or by increasing the effective sputtering yield (by increasing the RF power or bias voltage, for example). Less vertical or more sloped sidewalls may sometimes be desired—adequate filling or conformal deposition of subsequent layers may require a slightly sloped underlying topography. Producing less vertical sidewalls could be accomplished by reversing the above mentioned parameter trends. Sloped sidewalls can also be produced by allowing a certain amount of mask erosion to occur and using a tapered mask. While this allows the slope to become less vertical, it does so by increasing the width of the etched region, as we will show later in a simulation.

The etch profile can also be altered by changing the etchant gases. More vertical sidewalls can be obtained by using etchant gas species that have a lower spontaneous chemical component and which have appreciable etching only where there is ion bombardment, such as by ion-induced damage or byproduct removal mechanisms. One can also choose a chemistry that favors inhibitor formation. The inhibitor layer obstructs the spontaneous chemical etching and allows for ion-enhanced etching, resulting in very vertical etching. If the inhibitor forms too quickly relative to the etch rate, nonvertical sidewalls can result, as discussed earlier and illustrated in Figure 10–14. Figure 10–21 shows how the sidewall angle can be changed by changing the relative composition of C_2H_6 in the Cl_2/C_2H_6 etchant chemistry to etch polysilicon. Adding C_2H_6 increases the sidewall polymer formation (or "deposition") relative to the etch rate, causing the sidewall to be less vertical. In this manner, the etch profile can be tailored to the desired shape.

It is interesting that while plasma etching is rather insensitive to temperature, the deposition of an inhibitor layer often shows a negative activation energy with respect to temperature. In a plasma, the bond breaking and forming, which often limit film deposition, is rather temperature insensitive since most of the energy for this is supplied by the plasma energy and not thermal energy. However, the desorption of the deposited inhibitor layers increases with temperature, so that the net temperature dependence of inhibitor deposition is actually negative.

Etch uniformity means uniform etch rates across a wafer and from wafer to wafer. In batch systems, wafer-to-wafer uniformity involves the proper design of equipment so that the plasma and gas flows are uniform across the wafers in the system. Single-wafer systems have been developed to achieve better control over each wafer and to elimi-

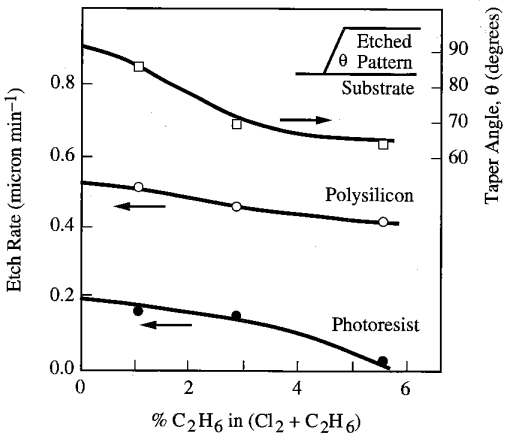


Figure 10–21 Etch rate of polysilicon and photoresist, and resulting taper angle as a function of C_2H_6 in Cl_2/C_2H_6 etch chemistry. More C_2H_6 results in more sidewall inhibitor formation and deposition relative to etching. (After [10.8].)

nate the wafer-to-wafer problems. Uniformity across a single wafer can still be a problem. Sometimes wafers develop "bulls-eye" patterns where the outside etches faster than the inside. This is usually due to depletion effects as the gas flows from the outside of the wafer to the inside.

Another kind of problem related to etch uniformity is called the "loading effect." There are two types of loading effects: macroscopic loading and microscopic loading. Macroscopic loading refers to the phenomenon in which more wafers in a chamber or more area exposed on a wafer to be etched results in a slower etch rate. This occurs because of the depletion of etchant species that occurs as the material is being etched. It is hard to predict to what extent this will occur due to the complex balance between species in the plasma and on the wafer surface. For the same reasons it is often difficult to control—increasing the flow rates does not always help.

Microloading occurs when etch rates vary over small distances on the surface of the wafer. This can occur when differences in the density of the unmasked area occur over small distances, resulting in differences in etch rates, just as in macroscopic loading. Microloading can also occur due to differences in aspect ratios. This effect is known as "Aspect Ratio Dependent Etching" or ARDE. It is also known as RIE "lag." As shown in Figure 10–22, the etch rate is different for different aspect ratios, with a lower etch rate for higher aspect ratio (smaller width) trenches. Different mechanisms have been suggested to explain this, each of which may be important under certain conditions. These include (1) depletion or trapping of the reactant species, or other conductance limitations, as the species travel to the bottom of the trench; (2) distortion of ion paths due to charging; and (3) shadowing effects involving either neutrals or ions. In the last case, collisions in the sheath may cause ions to be off-axis so that the number which reach the bottom is reduced due to simple geometrical shadowing. In each case, a narrower width decreases the chance that the reactant species, either ionic or reactive neutral, makes it to the bottom, thereby reducing the etch rate at the bottom. Etching at a lower pressure would reduce this effect if the shadowing mechanism were important, since reducing

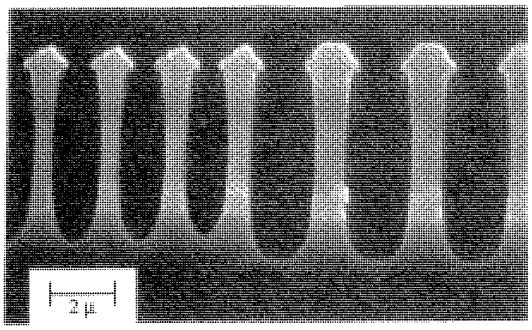


Figure 10–22 Example of Aspect Ratio Dependent Etching (ARDE) in silicon trench etching. The depth of etching decreases as the lateral feature size decreases [10.10].

the pressure would decrease the collisions in the gas phase and hence reduce shadowing effects. This is seen to be true experimentally [10.9] for some conditions, and we will simulate this later in the chapter.

Other uniformity issues involve differences in etching due to local variations in the film itself. The etch rate is a function of the density of the film, as one would expect, but also of the doping of the film [10.11]. Since these can vary locally across the film, the etch rate will vary as well.

Because of nonuniformities in the etch rate, and because the film itself may be of nonuniform thickness across the wafer or wafer to wafer, a certain amount of overetching is done to ensure that complete etching is achieved everywhere on the wafer. This is often 10–20%, in terms of time, past the endpoint detection. (*In situ* measurement and monitoring techniques, including endpoint detection, will be discussed in the measurement techniques sections.) Even more overetching may be required when anisotropic etching is done over nonplanar topography. Figure 10–23 illustrates this for purely vertical anisotropic etching of a conformally deposited film, where an additional time would be required to etch away the corner region. "Stringers" can results if strings of the material are left unetched at the bottom of the corner. Of course, this phenomenon can be taken advantage of in order to produce self-aligned spacers, as we discussed in Chapter 2 (Figure 2–30). In that case, no extra overetch would be done and the remaining oxide at the corner would be left as a spacer. The structure shown in the right side of Figure 10–23 would result, with the very narrow self-aligned spacer having a width approximately equal to the deposited film thickness.

Contamination and particulates can be reduced by reducing the pressure. Fewer contaminants are in the system and there is reduced polymerization of some etch species at lower pressures. Unwanted residues are often due to polymeric byproducts of the etch process. These can often cause problems with incomplete etching. If residues are left in contacts (or vias) after etching the contacts in a dielectric, high contact resistance can result. Reducing residue formation can be done by the proper choice of reactant gases. However, often polymeric formation is desired in order to achieve etch anisotropy, as discussed earlier, and gases are selected to promote such polymeric formation. In that case it is important to eliminate any unwanted residues from the etched surface following the etch process. Overetching can help remove residues in some situations. But in the

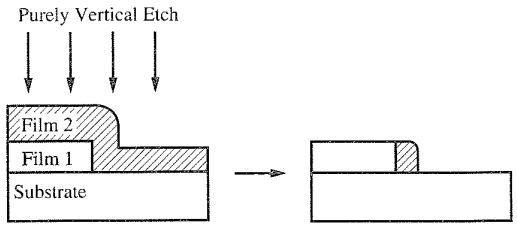


Figure 10–23 Overetching of a film over a step would be required to remove material at corner. In this example, completely anisotropic etching is assumed to emphasize this concept. This phenomenon can be used in producing self-aligned oxide spacers, in which no extra overetching is done so as to intentionally produce the structure at the right.

case of via etching, it may make it worse since sputtering of the underlying metal onto the sidewalls makes the sidewall residue harder to remove. One safe way to remove etch residues is by using a selective chemical etch after the plasma etch.

10.3.2 Plasma Etch Methods for Various Films

Most reactant gases for plasma etching contain halogens, generally Cl and F, and sometimes Br. Free radicals of these species can easily be produced in a plasma which can efficiently etch many films, and volatile etch products commonly result with these species. The exact choice of reactant gases to etch a specific film depends on a variety of factors, the most important being (1) etch selectivity to underlying films; (2) anisotropic etching; and (3) volatility of main etch byproducts.

Since in an etch process it is necessary to remove the byproducts from the surface, keep them off the wafer and pump them out of the system, the volatility of the main byproducts is an important consideration in choosing the gas chemistry. The volatility, or tendency to evaporate, depends on how tightly bound the species is to the surface. Table 10–2 lists the boiling points of typical etch products. The boiling point is an indicator of the volatility of a species—the lower the boiling point, the lower the surface binding energy and the higher the volatility. One can see that if Al is etched with a fluorine chemistry, AlF_3 would be the main byproduct. But the boiling point of AlF_3 (actually the sublimation temperature since it transforms directly from solid to gas without going through a liquid phase) is quite high at 1291°C. This means that the etch byproduct will be tightly bound to the surface with a very low vapor pressure, and even with ion bombardment it would be difficult to remove and keep off the surface. On the other hand, the byproduct of a chlorine-based etchant would be $AlCl_3$ with a boiling point (sublimation temperature again) of 177.8°C. This indicates looser bonding to the surface and a much more volatile species, so that a chlorine-based etchant would be a much better choice to etch Al. Having a volatile, weakly bound main etch byproduct is important when choosing an etch chemistry for a particular film. It is important to note that other, sometimes intermediate, byproducts may form during the etch process, and their removal or conversion to a more volatile species can be a rate limiting step in the etch process.

Table 10–3 lists some common chemistries used in plasma etching films in integrated circuit fabrication. These etchants are chosen mostly based on issues associated with byproduct volatility, etch selectivity, and etch profiles. The exact etch profiles and

Table 10–2. Boiling points of typical etch products (After [10.12]).

Element	Chlorides	Boiling Point (°C)	Fluorides	Boiling Point (°C)
Al	$AlCl_3$	177.8 (subl.)	AlF_3	1291 (subl.)
Cu	CuCl	1490	CuF	1100 (subl.)
Si	$SiCl_4$	57.6	SiF_4	− 86
Ti	$TiCl_3$	136.4	TiF_4	284 (subl.)
W	WCl_6	347	WF_6	17.5
	WCl_5	276	WOF_4	187.5
	$WOCl_4$	227.5		

selectivities will depend on the specific system and plasma conditions. We will next discuss the plasma etching of three films, silicon dioxide, polysilicon, and aluminum, in a little more detail. In doing so we hope to illustrate some of the major issues involved in plasma etching and in choosing the etch chemistries.

10.3.2.1 Plasma Etching Silicon Dioxide

Fluorine-based chemistries are usually used to etch silicon dioxide. The volatility of the main etch product, SiF_4, is very high as indicated by its low boiling point in Table 10–2 and thus the SiF_4 can easily leave the surface via thermal desorption. Fluorine can etch the SiO_2 by reaction of the fluorine free radical, with the overall reaction given by

$$4F + SiO_2 \rightarrow SiF_4 + O_2 \tag{10.12}$$

CF_2 and CF_3 free radicals may also be directly involved in the case of fluorocarbon etchants, in which case CO-containing byproducts would also be produced.

Etching with a significant isotropic component can result using source gases such as NF_3, SF_6, CF_4, or CF_4/O_2 mixtures. CF_4 dissociates to form free F in a plasma via the reaction given in Eq. (10.9). As mentioned earlier, adding O_2 to the CF_4 gas significantly increases the F free radical production and hence increases the etch rate as well as makes the etch profile more isotropic. Too much oxygen, though, can decrease the etch rate due to dilution [or by reversing the reaction in Eq. (10.12) and redepositing oxide on the surface]. These gas sources with relatively high free F concentrations will etch the oxide; however, depending on the conditions, the etching can result in undercutting. In addition, the etch selectivity over silicon or polysilicon, which is often underneath the oxide, is poor. The anisotropy can be enhanced by increasing the ion bombardment and lowering the pressure [10.14], but the selectivity over silicon is still generally poor with these chemistries.

The key to making the etching of silicon dioxide very anisotropic as well as much more selective over Si is to reduce the amount of F free radical production and increase the carbon content. Doing so will decrease any spontaneous isotropic etching by free

Table 10–3. Typical or representative plasma etch gases for films used in IC fabrication (After [10.1, 10.4, 10.13, 10.14].)

Material	Etchant	Comments
Polysilicon	CF_4	Isotropic or near isotropic (significant undercutting); fair to no selectivity over SiO_2.
	CF_4/H_2,	Very anisotropic; nonselective over SiO_2.
	$CF_4/O_2, SF_6$,	Isotropic or near isotropic; good selectivity over SiO_2.
	HBr, Cl_2, $Cl_2/HBr/O_2$	Very anisotropic; most selective over SiO_2.
Single-crystal Si	same etchants as polysilicon	
SiO_2	SF_6, NF_3, CF_4/O_2, CF_4	Can be near isotropic (significant undercutting); anisotropy can be improved with higher ion energy and lower pressure; poor or no selectivity over Si.
	CF_4/H_2, CHF_3/O_2, C_2F_6, C_3F_8	Very anisotropic; selective over Si.
	$CHF_3/C_4F_8/CO$	Anisotropic; selective over Si_3N_4.
Si_3N_4	CF_4/O_2	Isotropic; selective over SiO_2 but not over Si.
	CF_4/H_2	Very anisotropic; selective over Si but not over SiO_2.
	CHF_3/O_2, CH_2F_2	Very anisotropic; selective over Si and SiO_2.
Al	Cl_2	Near isotropic (significant undercutting).
	$Cl_2/CHCl_3$, Cl_2/N_2	Very anisotropic; BCl_3 often added to scavenge oxygen.
W	CF_4, SF_6	High etch rate; nonselective over SiO_2.
	Cl_2	Selective over SiO_2.
Ti	Cl_2, $Cl_2/CHCl_3$, CF_4	
TiN	Cl_2, $Cl_2/CHCl_3$, CF_4	
$TiSi_2$	Cl_2, $Cl_2/CHCl_3$, CF_4/O_2	
Photoresist	O_2	Very selective over other films

fluorine. The etching that then occurs involves other fluorine species and is strongly ion enhanced. In addition, polymer inhibitor formation will be increased, also suppressing lateral etching. Furthermore, selectivity over Si will be greatly enhanced.

One way to reduce the relative amount of F free radicals is by adding H_2 to the CF_4 gas. The hydrogen reacts with the fluorine to produce HF which reduces the amount of atomic F. Another possibility is to use more fluorine-poor and more carbon-rich gas species, such as CHF_3, C_3F_8, or C_2F_6. With fewer F free radicals, any spontaneous (non-ion-enhanced) chemical etching by these species is reduced. Appreciable etching happens only where ion bombardment occurs, perhaps involving an ion-induced damage or bond-breaking process, leading to anisotropic etching. CF_2 or CF_3 free radicals, rather than free F, are believed to be involved in the etch mechanism in these fluorine-deficient gas mixtures [10.14]. Increasing the carbon-to-fluorine ratio with these etchants also increases the amount of polymer inhibitor formed, or deposited, during the etching [10.6]. With a CF_4 or CF_4/O_2 gas, little or no polymer forms on the surface. However with a higher carbon concentration relative to fluorine in the gas mixture,

more polymer formation will occur relative to etching. With a thick polymer inhibitor layer present, etching can not occur on vertical surfaces where there is little ion bombardment to remove the polymer inhibitor layer. Any lateral etching is suppressed, and near vertical sidewalls can result.

Of course, as even more carbon is added to the system and the deposition of the polymer becomes faster relative to the etching of the polymer and film, then sloped sidewalls (without mask undercutting) can result, as discussed earlier and illustrated in Figures 10–14 and 10–21. CHF_3 has a higher ratio of C to F than CF_4 does so that more polymer forms, in fact enough so that sloped sidewalls often result. If pure CHF_3 produces excess inhibitor deposition for particular plasma conditions, then O_2 could be added to react with the carbon and reduce the polymer deposition and change the etch profile. Other species such as CO_2 or CF_4 may also be added to CHF_3 for this purpose. Figure 10–24 conceptually illustrates the different etch profiles in SiO_2 that could result when different amounts of inhibitor formation occur.

Reducing the fluorine-to-carbon ratio also greatly improves the etch selectivity of SiO_2 over Si. This is because polymer formation preferentially occurs on Si surfaces compared to SiO_2 surfaces during the etch process. This may be due to the fact that carbon in the polymer reacts with the oxygen from the oxide, forming carbon monoxide, which results in much less accumulation of polymer on the oxide surface. With more polymer on the Si, its etching is slower, resulting in a higher SiO_2 to Si etch selectivity. Figure 10–25 shows how the etch rates of Si and SiO_2, and hence the selectivity, vary as H_2 is added to a CF_4 gas mixture [10.15]. The etch selectivities of oxide over $TiSi_2$ and TiN, which are important when etching contact or via holes in multilayer structures, are enhanced in this same manner by adding H_2 to the fluorocarbon etchant.

Changing the etchant chemistry may also change the etch rate of the mask. As seen in Figure 10–25, the etch rate of photoresist changes significantly as the H_2 content in the H_2/CH_4 gas mixture changes, with a lower H_2 content (and less F production) resulting in a higher etch rate. Adding O_2 to a CHF_3 mixture similarly increases the etch rate of the photoresist. The resulting mask erosion can cause the etch profile of the SiO_2 to change, as we will illustrate in the models and simulation section (Figure 10–33), and may alter the profiles shown above in Figure 10–24.

A relatively large bias voltage is often used for SiO_2 etching (in the 400- to 500-V range). This is needed to help remove the polymers. After the SiO_2 etching is complete, including a suitable overetch, any remaining polymer residue must be removed, espe-

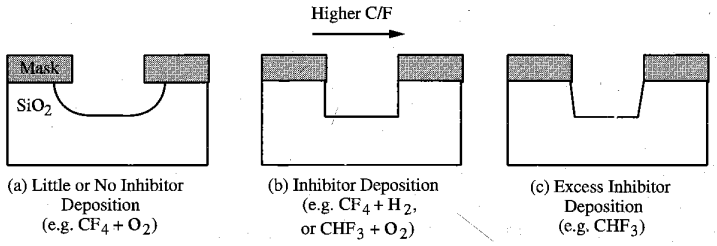


Figure 10–24 Effect of inhibitor deposition on resulting etch profile. The profiles shown assume no mask erosion, which may occur for different etch chemistries. The actual profiles obtained will also depend on the specific etch system and plasma conditions.

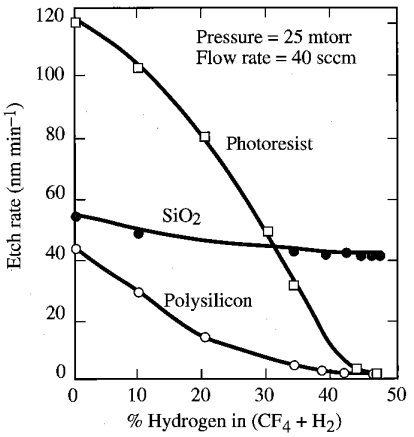


Figure 10–25 The etch rates of Si and SiO_2 (and photoresist) as H_2 is added to a CH_4 gas mixture. (After [10.15].)

cially in contact holes. The polymer can be removed by an O_2 or CF_4 plasma treatment. CF_4 is better at removing all the polymer and even some damaged substrate, but it cannot be over done since it has poor selectivity over Si and there can be lateral etching. A lower energy is used for the cleanup step so as not to damage the exposed material below. Also helpful is an anneal treatment; the polymer appears to break down at temperatures higher than $\approx 280°C$.

It should be clear from the discussion in this section that there are many details and subtleties associated with developing a particular manufacturing recipe for plasma etching. We will return to some of these issues in the models and simulation section.

10.3.2.2 Plasma Etching Polysilicon

The choice of etch chemistry for polysilicon depends on the same factors as etching silicon dioxide: byproduct volatility, selectivity, and anisotropy. Another issue is the presence of the ubiquitous native oxide layer on the film's surface, requiring an initial "breakthrough" etch step to remove it.

F-based etchants could be used for polysilicon etching because of the volatility of SiF_4. However, etching polysilicon with SF_6, CF_4, or CF_4/O_2 has a relatively large isotropic component due to the high concentration of F free radical species produced in such plasmas combined with little polymer formation. We saw that with etching SiO_2, anisotropy can be enhanced by reducing the relative F component and increasing the carbon component, thus decreasing the spontaneous etching, increasing polymer formation, and enhancing ion-induced etching. Adding H_2 to CF_4 can accomplish this. But doing so also increases the etch rate of SiO_2 relative to Si by preferentially forming a polymer on the Si. This is good when etching SiO_2 but bad when etching polysilicon when a high selectivity over the underlying SiO_2 is required. (Recall the structure in Figure 2–26 in the CMOS process flow in Chapter 2 for an example of why good selectivity over SiO_2 might be important.) Thus with a fluorine chemistry it is difficult

to achieve both anisotropic etching and good selectivity of polysilicon over silicon dioxide at the same time.

If anisotropy is not important, then adding O_2 to CF_4 allows etching polysilicon with good selectively over silicon dioxide since the etch rate of the Si is increased by O_2 much more than that of SiO_2. But the etch will be fairly isotropic. One approach with fluorine chemistry when both anisotropy and selectivity with respect to an underlying oxide are desired is to switch the gas chemistry during the etching. One would start with an anisotropic etch, with a CF_4/H_2 mixture for example, and then when the etch is almost done, switch to an etch chemistry that is selective over SiO_2, using CF_4/O_2 for example. This will, however, result in some undercutting of the polysilicon.

Anisotropic and selective etching of polysilicon can be achieved simultaneously using Cl-based etchants. The byproduct volatility is a little lower with Cl than with F, but is still acceptable. Cl_2, HCl, or $SiCl_4$ can be used. These have a lower purely chemical component to etching than the F-containing etchants, and less lateral etching of Si occurs than with F. The etch rate in the vertical direction is significantly increased by ion bombardment, perhaps by surface bond breaking or by enhancing byproduct removal. Thus the etching of Si with Cl can be fairly anisotropic even without a polymer inhibitor layer present. Along with being anisotropic, the etch selectivity of Si to SiO_2 is high. For one thing, there is no polymer inhibitor layer that may form preferentially on the Si, as in the fluorine case, and degrade selectivity. Also, chlorine etches SiO_2 rather slowly. Furthermore, removing or minimizing the carbon in these systems is seen to significantly lower the etch rate of SiO_2 by chlorine—carbon evidently plays a role in the SiO_2-Cl etch reaction and enhances the etching at low carbon concentrations. Therefore the etch selectivity of Si to SiO_2 can be quite high, typically over 100, for these anisotropic, chlorine-based etchants. The anisotropy can be enhanced even more by adding a small amount of O_2. The O_2 reacts with the Si film, or with etch reaction byproducts that have adsorbed on the sidewall to form a nonpolymer, SiO_2-based inhibitor layer. The O_2 also helps with selectivity over oxide by removing any carbon that comes from photoresist erosion. Keeping the wafer temperature low, by providing helium backside heat conduction for example, also improves the anisotropy by enhancing inhibitor deposition.

Cl-based etchants exhibit a large doping sensitivity. High n^+ levels can increase the polysilicon etch rate by up to 25 times, and can increase the spontaneous etching and undercutting. An SiO_2-based inhibitor can block this. Cl-based chemistries etch the polysilicon slower than F-based chemistries, and mixtures of the two types have been used. The addition of SF_6 also reduces trenching. Trenching at gate edges can be a problem with chlorine etching, with the underlying gate oxide being damaged. Chlorinated flourocarbons, such as C_2ClF_5, have also been used for etching polysilicon but are out of favor since they are believed to cause environmental damage by depleting the ozone layer.

The most popular etchants used currently for polysilicon are chemistries based on bromine, the next halogen after F and Cl in the periodic table. HBr and Br_2 are commonly used. These act in a similar fashion to the Cl-based etchants, etching with a low purely-chemical component and without a polymer inhibitor, and are even more anisotropic and selective over oxide. Trenching of the underlying oxide is less of a problem as well. But the etch rates of Br-based chemistries are even slower than Cl-based

chemistries, and they are often used in a mixture with Cl_2 in current polysilicon etch processes. As with etching with Cl, adding O_2 in these noncarbon containing gases promotes inhibitor formation, perhaps by forming SiO_2 from the Si film, and thereby increases the anisotropy. The O_2 also helps with selectivity over oxide by removing any carbon that might result from photoresist erosion. One drawback to the use of HBr and Cl_2 is their high toxicity which puts more stringent requirements on equipment and gas handling systems.

Polysilicon, like single-crystal silicon, usually has a small native oxide layer present on the surface, even after cleaning. This must be removed prior to etching. Since the etchants used for the main polysilicon etch have low-oxide etch rates, the etch must be initially modified to increase the oxide etch rate. This is called a breakthrough step. During the first 5 to 10 seconds of the poly etch, one can lower the selectivity with respect to oxide by adding an etchant such as CF_4 or by increasing the RF power to enhance the physical component of etching and etch the thin oxide layer from the surface.

An overetch step after the main etch is also important for polysilicon etching. This not only allows for deposition and etch nonuniformities, but also for the presence of stringers at steps, as discussed earlier and illustrated in Figure 10–23. This overetch step is often done at a lower RF power to increase the selectivity over oxide. A final postetch "cleanup" step is done to remove any sidewall layer material. While polymeric (carbon containing) layers would need an O_2 plasma strip, followed by an HF dip, inorganic layers usually need only an HF dip.

To produce shallow trench isolation structures, etching of the silicon is a critical process step as described in Chapter 2 (Figures 2–6 to 2–9). Achieving the desired etch profile and selectivity to mask layers is important, and many of the same ideas regarding polysilicon etching would apply. Using a Cl- or Br-based etchant can give good selectivity over silicon dioxide or silicon nitride mask layers and provide good anisotropy, as described above. Using Cl etchants alone can result in harmful ion-induced trenching at the bottom corners, so Br etchants are often used—either by themselves or in combination with Cl etchants. The slope of the etch profile can be tailored by including gas additives, such as O_2, to promote inhibitor deposition, and multistep etch processes may be utilized to achieve the desired profile.

10.3.2.3 Plasma Etching Aluminum

Plasma etching of aluminum interconnect lines has been a challenge for process engineers, but successful strategies have been developed to overcome the many problems associated with this film. Again, byproduct volatility, selectivity, and anisotropy are issues, as well as native oxide on the surface. In addition, a potential corrosion problem as well as Cu doping in the film lead to more difficulties.

Because of the presence of a native oxide, Al_2O_3, on the Al film surface, an initiation or breakthrough etch step must be done before the main Al etch step. Ion bombardment, or sputtering, is usually done to remove this film in conjunction with chemical scavenging. Argon sputtering can be used, but gases such as BCl_3, $SiCl_4$, CCl_4, or BBr_3 are often used to scavenge the O_2 and H_2O and remove Al_2O_3 on the Al surface by both physical and chemical means. A load lock system is required to control the H_2O level.

The byproduct of etching Al with F-based chemistries, AlF_3, is not very volatile as discussed earlier, so that a Cl-based chemistry is most often used for the main Al etch step. Cl_2 is commonly used, and Cl_2 (not a free radical, but still a reactive neutral species) can directly etch the Al. However Cl_2 etches Al isotropically, producing an undercut, and ion bombardment has little effect on the etch rate (unlike the etching of Si by Cl_2). To suppress lateral etching of the Al and obtain vertical profiles, sidewall inhibitor formation is needed. Therefore inhibitor sources or promoters are often added to the gas mixture. These include $CHCl_3$, $CFCl_3$, CCl_4, and N_2. Resist erosion plays a significant role in supplying inhibitor material in Al etching, and N_2 enhances the inhibitor formation from photoresist, perhaps forming a CN polymer. BCl_3 also seems to promote inhibitor formation. The etch rate and selectivity over silicon dioxide are acceptable for Cl-based chemistries.

Al interconnect lines usually contain Si and Cu, which can complicate the etching. While Si is readily etched by Cl, Cu is not (the volatility of the Cu byproducts by halogen etching is quite low). Ion bombardment and often higher temperatures are needed. We will describe in Chapter 11 how Ti and/or TiN layers are often used above and below Al layers in the interconnects. These are usually etched at the same time as the Al, and Cl-based chemistries can readily etch them as well.

An overetch step is especially important for Al etching to remove the residues, both from the inhibitor layers and the Cu precipitates that can remain. The difficulty in Cu precipitate and residue removal depends on the Al deposition conditions, with a higher deposition temperature giving larger precipitates and more residues.

Corrosion of the etched Al line can be a problem when it is exposed to the ambient. Residual chorine on the sidewall and in the resist can react with water to form HCl. The HCl can then etch the Al. Cu in the films enhances this electrochemical reaction. It is important to passivate the Al film after the etching and before it is exposed to air. The wafer can be heated to 100°C or 150°C to drive off the Cl. Or the Cl can be buried under a CHF_3 polymer, which is then wet etched. But the common method has been to expose the wafer to an F-containing plasma, such as a CF_4 plasma. The chlorine is replaced by fluorine on the Al surface, passivating it. An O_2 plasma then removes the photoresist, and a deionized H_2O bath removes any remaining chlorine. A wafer rinse may also be done before the ashing to help remove the water-soluble chlorides before they are converted by the ashing to insoluble metal oxides [10.16]. A newer process uses an H_2O plasma to both pacify the surface and to remove the resist.

We end this section with a short comment regarding the use of freons, or fluorocarbons. Chlorinated flourocarbons without any hydrogen, such as C_2ClF_5 or $CClF_3$, are believed to deplete the ozone layer in the upper atmosphere. There has been a trend to replace these with safer source gases in the semiconductor industry.

10.4 Measurement Methods

Measurement techniques for etching can be separated into two types: post-etch measurements and *in situ* measurements. Post-etch measurements are generally performed outside the etch chamber. *In situ* measurements are done inside the chamber as the

etching is occurring for diagnostic and monitoring purposes. The most important type of monitoring is that done for endpoint detection.

Post-etch measurements for etching are performed to characterize the etch process. The resulting line width can be measured using an optical microscope. Observing the wafer under a microscope can also indicate whether some of the film still remains in the patterned areas due to underetching. However optical observations and measurements are difficult for submicron lines. A much more accurate method for line width measurement is the Scanning Electron Microscope (SEM), discussed in earlier chapters. More automated versions of SEMs have been especially designed for linewidth observation and measurement for both lithography and etching. Measurements from above the wafer looking down assume that the sides of the etched film are vertical. Obviously this may not be so. To observe and measure the actual profile of the etched film, cross-sectional SEM images are used. An example was shown in Figure 10–22. This method requires cleaving the sample and is done on test wafers, usually when developing the process. Linewidth measurements of etched films using electrical testing methods were discussed in Chapter 5.

Other post-etch measurement techniques include composition and contamination measurements. Small amounts of polymeric residues may remain which can have detrimental affects on later processing and circuit behavior. Surface analysis techniques such as Auger Electron Spectroscopy (AES) and X-ray Photoelectron Spectroscopy (XPS) can be used to detect and measure chemical species left on the film or in the near-surface region. XPS is often used to give bonding information and can distinguish whether carbon detected on the surface is bonded to fluorine in C-F_2 structures or to the silicon in the substrate with C-Si bonding, for example. This helps to determine the actual source of the contaminant or residue and the mechanism for its production.

Measuring damage induced by plasma processing is important. Physical damage, in terms of dislocations, trenching, and surface roughening, can be measured by TEM or SEM. Electrical damage, including charging and electron traps, is measured using device or electrical measurements, often using MOS capacitors as discussed in Chapter 6.

In situ measurements are performed inside the etch chamber while the etch process is taking place. The most important type of *in situ* measurement is that done for endpoint detection. Due to the physical component in most plasma etching, selectivity is not 100 percent. Therefore it is important to end the etching soon after the film has been etched to the bottom (allowing for a certain amount of overetch, of course). Some sort of *in situ* measurement is required to do this.

One way to detect the endpoint is to measure the thickness of the film as it is being etched. If one does not want to etch all the way through the film, the endpoint thickness may be some finite value. This method would also give etch rate information as well. An *in situ* thickness measurement technique for optically transparent films, such as silicon dioxide, is interferometry. This method determines the thickness of the film by measuring the interference of optical beams—using either light from a laser or from the plasma itself—reflected from the top and bottom of the etched layer. As the material is etched, oscillations in the reflected signal occur as constructive and destructive interference occurs. The period of oscillation is related to the change in film thickness, x_f, via the relationship

$$\Delta x_f = \frac{\lambda}{2n} \tag{10.13}$$

where λ is the wavelength of the laser light and n is the refractive index of the film. Therefore, for each complete cycle in the reflected signal, the film has been etched Δx_f. When the film has been etched completely, the oscillations cease. For nontransparent films, such as metals, techniques using reflectivity differences between the film and the layer underneath can be used. A major disadvantage in film thickness techniques is that the information is obtained from only a small area of the wafer which may not be fully representative of other regions. Recent single-wafer systems use an array of sensors to measure the thickness change at many points on each wafer. One advantage to these systems is that they can accurately measure the ongoing etch rate during an etch process, which can be valuable information concerning the process.

The most common methods for endpoint detection are those which monitor the reacting species or reaction products of an etch process. Optical spectroscopy or mass spectroscopy can be used, but the former is more common. In optical emission spectroscopy a photosensor detects and measures the emitted radiation, through a window into the chamber, from the various species in the plasma as a function of wavelength. In a plasma, light is emitted by exited atoms and molecules as electrons relax from one energy state to another (which is why the plasma glows). Each species emits light in a characteristic spectrum which can be used to identify the particular atom or molecule. (See Figure 4–18.) For example, the fluorine free radical emits light at 703.7 nm (among other wavelengths) as it relaxes from a plasma-induced excited state. Monitoring this wavelength can be used to determine the relative concentration of the fluorine free radical in the plasma. Endpoint detection is done by either monitoring the reacting species or the reaction products.

As an example, consider the etching of SiO_2 on Al by CF_4 chemistry. The main etch reaction occurring here involves the fluorine free radical which is produced in the plasma. As the etching begins, F is consumed by the etch process, forming SiF_4. Therefore the signal from F emission is decreased. When the etching finishes, the concentration of F increases, and its signal goes back up. The endpoint is thus indicated by the increase of the reactant species. Alternately, the endpoint can be indicated by the increase or decrease in the emission signal of an etch product. For example, in etching SiO_2 by CHF_3, one reaction that can occur is between the carbon on the SiO_2 surface and oxygen, producing CO. Monitoring CO at its characteristic 483.0-nm wavelength can be used to detect the presence of CO. When the CO signal decreases below a certain point, the etch is determined to be over. This method, in contrast to thickness measurements, takes its information from the gas phase and thus averages over a much larger region. Therefore, a more average endpoint is determined. Loading effects can occur, since the signal from either the reactants or products will depend on how much material is being etched. In addition, the exact quantification of the signal to concentrations is difficult and some calibration is necessary to determine at what point the etch is completed. However, this method is relatively simple and usually a good indicator of when the etching is complete, and is the most common method in current technology.

The technique of optical emission spectroscopy, in addition to being used as an endpoint determination method, is also used in diagnostics of plasma etch processes. By following the emission signals, the relative concentrations of the various species present during the etching process can be measured throughout the etch process. This can give valuable information concerning the various mechanisms occurring and the roles of the different species.

In situ measurements can also be utilized for process control applications. For example, loading effects could be monitored and changes in equipment parameters could be made automatically and instantaneously, based on the feedback from such measurements. Such measurements could also be used to determine when chambers must be cleaned due to buildup of etch residues on the chamber walls.

10.5 Models and Simulation

We have described the different types of underlying mechanisms important for plasma etch processes: chemical etching by reactive chemical species; physical etching by bombardment or sputtering of ionic species; and ion-enhanced etching involving both types of species acting in a synergistic manner. As in Chapter 9, we will describe physically based models that predict the shape, or topography, that results from specific etching technologies. These models are implemented in a variety of simulation tools including SPEEDIE [10.17], TAURUS-TOPOGRAPHY [10.18], and ATHENA [10.19], as well as other topography simulation tools mentioned in the previous chapter. We will show examples from all three of these simulators. In this section we will first describe how physically based simulators model plasma etching. Since there are a number of similarities to the models described in Chapter 9 for simulating deposition of thin films, we will refer the reader to that chapter for some of the specific details.

10.5.1 Models for Etching Simulation

As in the simulation of deposition, plasma etching can be simulated by considering the fluxes of species that come from the plasma and strike the surface of the wafer. In Figure 10–20 we schematically illustrated the fluxes of species coming from the plasma and the various processes that occur on the wafer surface during etching of the films. In physically based topography simulators, the flux of each type of species is calculated, in particular the component of the incoming flux normal to the surface at each point. Because of the relatively low pressures used in etch systems and large mean free paths compared to the feature size on the wafer surface, it is usually assumed that free molecular flow is the dominant gas-phase transport mechanism near the surface, that is, there is no boundary layer diffusion process. The net etch rate at each point is then calculated using an expression that relates the etch rate to the different fluxes. The exact form of that expression depends on the model that is used. We will describe some of these models and expressions in the next section. With the etch rate determined at each location,

the appropriate amount of material is removed, the boundary is moved, and the profile can evolve as the etching continues. The etch rate at each location can change with time as the etching progresses because the fluxes that reach each location on the surface can change due to changing geometrical effects, such as flux shadowing.

As in deposition modeling, the flux of species arriving from the plasma and striking the wafer surface can be calculated using Monte Carlo techniques. For example, the distribution of ions in terms of arrival angle at the surface and of energy can be calculated as a function of system parameters, such as gas pressure, composition, and RF settings. This explicitly takes into account the collisions that occur in the plasma and in the sheath, and the acceleration across the sheath to the wafer. There have been several simulation tools developed, including a version of SPEEDIE, that take this approach. However, we again will utilize a simpler approach and define the incoming fluxes analytically on a plane just above the surface. We specify the maximum, unobstructed flux of each species—ions or chemical, for example—and the angular distribution of each.

The angular distribution is assumed to be of the form

$$F_{\text{direct}}(\theta) = F^{\circ} \cos^{n}\theta \tag{10.14}$$

where F_{direct} is the normal component of the flux coming in at angle θ, F° is the flux at θ = 0 degrees, and θ is the angle from the vertical direction. By adjusting n from 1 to large numbers, one can obtain isotropic to anisotropic distributions of arrival angles, respectively. This was illustrated in Figure 9–25 in Chapter 9. With $n = 1$, the distribution is isotropic, with fluxes coming from all directions equally. Higher pressure, more collisions in the gas phase, and shorter mean free path would all lead to a more isotropic arrival distribution. In etching, neutral chemical species are usually assumed to arrive at the top surface isotropically, and $n = 1$ is usually used. For $n > 1$, a tighter distribution of arrival angles would occur. This means that the fluxes arriving at each point on the surface are more directed, or more anisotropic. In etch models, the ionic species—accelerated across the sheath to the wafer surface due to the electric field—are assumed to arrive at the surface approximately perpendicular to the wafer and very anisotropically, and n = 10 to 80 is usually used.

Figure 10–26 illustrates the species and fluxes that occur in plasma etching that are included in physically based etch simulators. This is analogous to the fluxes used in deposition modeling (see Figure 9–35). The ionic species are involved in the physical component of etching, and the reactive neutral species would include those species involved in the chemical component of etching, and would primarily be the free radicals. Another type of species is the deposited, or deposition, species. These would include material that is deposited in the form of inhibitor layers, often carbon polymeric material. These would be analogous to the neutral species that deposit in the deposition models. As we mentioned earlier in this chapter, the same processes (deposition, chemical etch, ion bombardment, etc.) often occur in both deposition and etch processes.

As in deposition modeling, the total flux of each type of species arriving at each position will be the sum of that coming directly from the plasma (a direct flux) plus that coming from somewhere else on the surface of the wafer (an indirect flux). The indirect fluxes may be a result of reemission or resputtering from another location. In Chapter 9 many of those processes and the equations used to describe them were presented.

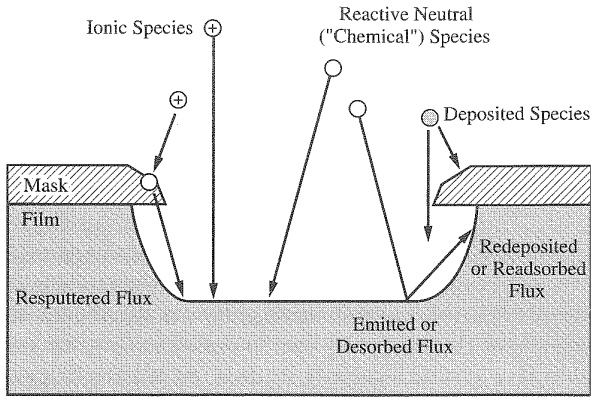


Figure 10–26 Illustration of fluxes, both direct and indirect, that are used in physically based etch simulations.

One of the important sources of indirect flux is that due to reemission of the etch reactant before it can react and etch. This is described by the sticking coefficient. The sticking coefficient is defined, as in Chapter 9 and earlier in this chapter, as the ratio of the flux of species that actually deposit or react at each location on the surface, to the flux of incident species:

$$S_c = \frac{F_{\text{reacted}}}{F_{\text{incident}}} \tag{10.11}$$

Therefore S_c times the chemical flux hitting the surface at each point is the net flux that actually etches, and $(1 - S_c)$ is the fraction of the flux that is reemitted. These reemitted atoms or molecules can then travel to other locations on the surface and etch material there. If S_c is set to a low number ($S_c << 1$) for the chemical species, as is usually the case, then isotropic etching can occur for purely chemical etch processes. Most of the chemical species move in all directions from place to place, as was illustrated in Figure 10–11(a), with a net result that etching is isotropic.

It is usually assumed that the angular distribution of the desorbed or emitted flux follows an ideal cosine distribution ($n = 1$ in $\cos^n\theta_i$ emitted distribution, where θ_i is the angle relative to the surface normal); that is, the emitted species are emitted in an ideal fashion from a small area on the surface. In SPEEDIE it is currently assumed that only the neutral species can be emitted from the surface; for ionic species the sticking coefficient is assumed to be one. This does not mean that all ionic species are successful in etching, or sputtering, an atom or atoms from the surface. It just means that the ions that do not etch any material are not reemitted to etch elsewhere.

In Chapter 9, we described how it has been experimentally determined that surface diffusion for SiO_2 deposited by both LPCVD and PECVD methods was not important —the results could be modeled better with the reemission and redeposition process. A similar experiment has been done to distinguish between the two processes in plasma

etching of silicon [10.20, 10.21]. By modeling the undercut profiles, it was determined that using the reemission/readsorption model, with a low sticking coefficient, works better than using the surface diffusion process in the simulation. Thus surface diffusion of the species is often not included as a model in etch simulators.

Physical etching occurs by sputtering when ions strike the surface. Sputtering occurs when the incoming ion transfers enough energy to the wafer surface to break the bonds holding a surface atom in place. The sputtering yield, Y, is defined as the number of atoms or molecules ejected from the target per incident ion and is a measure of the sputtering efficiency. Therefore Y times the ion flux striking the surface at each point determines how much physical etching occurs for a given ion flux .

The sputter yield, Y, is a function of ion energy, the mass of the incoming ion and the target material (the wafer surface material, in this case), and the incident angle, θ, at point i, with a normal incidence being 0°. As we discussed in Chapter 9, ions coming in at low angles are reflected from the surface and do not transfer significant energy to the surface atoms, while ions coming perpendicular to the surface knock surface atoms further into the film and not off. The maximum sputtering yield occurs usually at somewhere from 40–70°. This is usually modeled as was described in Figure 9–39.

In some of our models we will use K_i, the ion-induced etch yield, or the effective sputtering yield. Like Y, it is a measure of how many atoms or molecules of substrate are etched per incident ion but can include the synergistic effects with the chemical component, resulting in higher values for it than normal physical sputtering yields. K_i generally has similar properties as Y, angle and energy dependencies for example. For cases of purely physical sputtering, K_i would just equal Y.

All of the geometrical issues involved in the etching process are taken into account in physically based etch simulators. That is, the net flux of each type of species at each point on the surface is calculated based on the relative orientation of that point to all sources of fluxes. This includes fluxes from the plasma source (the direct flux) and from all other points on the surface (the indirect fluxes). The view angles to all other sources of fluxes, as well as the angular distribution emitted from each source, are taken into account in the calculation of the arriving flux at each point.

Once the net flux of each type of species is calculated at each point on the surface, the etch rate can then be calculated at each point. This is determined by the specific etch model used in the simulation. We will next consider some of the models and corresponding etch rate expressions that etch simulators use to model plasma etching.

10.5.2 Etching Models — Linear Etch Model

Different types of etch models have been developed and are available in simulators. Some of these models are for specific types of etching equipment—for high-density plasma etch systems or for sputter etch systems, for example. However, there are also some more general-purpose models. These were developed to model the different types of mechanisms that occur during plasma etching: physical, chemical, and ion-assisted etching, and which can occur in any of the plasma etch systems. We will concentrate on those models, and two in particular: the linear etch model and the saturation/adsorption etch model, both of which are used in the SPEEDIE simulator.

The simplest model for etching assumes that the chemical and ionic components act independently of each other and combine in a linear fashion. This is the basis of the linear etch model in SPEEDIE. Other simulators include similar models. It is assumed in this model that the etch rate due to each component is linearly proportional to the flux of that component. Studies of chemical etching do show a linear dependence of etch rate on the chemical flow rate or flux for certain conditions [10.5, 10.22]. Likewise, the etch rate in the presence of ionic bombardment is seen to be linearly dependent on the ion current or flux over a range of conditions [10.23]. The two terms are then added together, so that

$$\text{Etch rate} = \frac{(S_c K_f F_c + K_i F_i)}{N} \tag{10.15}$$

The density, N (in cm^{-3}), is that of the film being etched. F_c is the chemical (reactive neutral, free radical) flux and F_i is the ion flux at each point on the surface. S_c is the sticking coefficient, which is between 0 and 1. K_f and K_i are the relative rate constants for the two processes and can include any stoichiometric factors. In practice, K_f is often used as a fitting parameter for the purely (non-ion-enhanced or "spontaneous") chemical etching. As described above, K_i is the ion-induced etch yield and is related to the number of ions needed to remove one atom or molecule for the material being etched. It is essentially equivalent to the physical sputtering yield, Y, but can be greater than Y in some cases as we will discuss. K_i can be a function of incident angle and ion energy. Like K_f, it is often used as a fitting parameter.

One can see that the net etch rate at each point is just the linear combination of purely chemical etching and ion-driven etching, with each term linearly dependent on the appropriate flux. The etch profile would likewise be a linear combination of the two components, with chemical etching occurring equally in all directions plus the physical, ion-driven etching occurring primarily in the vertical direction.

While being simple, with no coupling between the components, the model may be appropriate for certain conditions. It would be valid when the etch process is purely chemical, in which case the ion flux would be zero, and the etch rate is simply proportional to the chemical flux. Purely chemical etching would occur in a barrel etcher or in an after-glow system, for example, where the ions are prevented from striking the wafer surface. It could even occur in a parallel plate system if little or no inhibitor layer is formed and there is a very high concentration of free radicals in the plasma and only spontaneous chemical etching occurs. For purely chemical etching, the etch profile would be completely isotropic and look like Figure 10–3(a). This model would also be valid when the etch process is only dependent on the ion flux, as in solely physical etching, in which case the first term would be zero. This would occur, for example, in a sputter etch system where only high-energy, chemically inert Ar^+ ions are used. When the etch rate is simply proportional to the ion flux with no chemical term, the profile would look like Figure 10–3(c), with vertical or near-vertical etching and no undercutting.

The model would also be applicable when the etch process does have both chemical and physical etching components, but the two act independently of each other. Or at least they each appear as if they are independent in that they do not limit the action of the other for a given range of conditions. For example, the physical, ionic component in

this model (the second term in the rate expression) does not, in fact, have to be due to purely physical sputtering. It could actually be ion-enhanced etching, but operating in a regime where the chemical component in that mechanism is high enough, due to a relatively high flux of neutrals, for example, and is not limiting the ion-driven etching, as we will describe later. In that regime the ion-enhanced etching would act as if it were purely physical and not dependent on the chemical flux, that is, simply proportional to the ion flux. In that case, K_i is often much greater than the normal physical sputtering yield. [And if there is no spontaneous chemical etching, the profile would look like that for purely physical etching, Figure 10–3(c).] In addition, any adverse effect the ion bombardment might have on the spontaneous chemical etch term, by depleting the surface chemical species for example, may not be significant at lower ion fluxes or higher neutral fluxes [10.21]. When two (seemingly) independent etch components are present, isotropic chemical and directional ionic, the etch profile would look like Figure 10–3(b), a linear combination of the profiles for spontaneous chemical etching and ion-driven etching. So in general, the simple linear etch model would be appropriate when one or two independent, or seemingly independent, etch components are present, resulting in a range of etch profiles.

Example Silicon is plasma etched for 5 minutes and the process follows the linear etch model. The flux of chemical species, F_c, on an unobstructed flat surface is equal to 2.5×10^{18} atoms cm^{-2} sec^{-1}, the sticking coefficient is 0.01, and K_f is 0.2. The flux of ionic species, F_i, on a flat surface is equal to 1×10^{16} atoms cm^{-2} sec^{-1} and K_i is 1. The density of Si is 5.0×10^{22} atoms cm^{-3} A photoresist mask is used so that a wide trench is etched in the Si. (There is no etching of the photoresist.) How far is the Si etched in the vertical direction (away from the mask edge), and how far is the Si etched in the lateral direction right under the mask for these conditions?

Answer For the linear etch model, Eq. (10.15):

$$\text{Etch rate} = \frac{(S_c K_f F_c + K_i F_i)}{N}$$

In the vertical direction, there would be both chemical and ion fluxes, and the total etch rate in this model would simply be the sum of those components. Away from the mask edge we assume no shadowing effects, so F_c and F_i are the unobstructed chemical and ion fluxes, respectively, given above. Therefore

$$\text{Etch rate(vert)} = \frac{(0.01 * 0.2 * 2.5 \times 10^{18}\text{cm}^{-2}\text{sec}^{-1} + 1 * 1 \times 10^{16}\text{cm}^{-2}\text{sec}^{-1})}{N}$$

$$= \frac{1.5 \times 10^{16}\text{cm}^{-2}\,\text{sec}^{-1}}{N}$$

The density of Si, N, is 5.0×10^{22} atoms cm^{-3}, so the Si etched in the vertical direction is

$$\text{Etch depth(vert)} = \frac{1.5 \times 10^{16} cm^{-2} sec^{-1}}{5.0 \times 10^{22} cm^{-3}} * 300 \text{ sec} = 9.0 \times 10^{-5} cm = 0.9 \mu m$$

In the lateral direction, right under the mask edge, there is virtually no ion flux ($F_i \approx 0$). Therefore only the chemical flux term is used in the etch rate expression and it is assumed that the chemical flux is the same in all directions (due to the isotropic arrival distribution and, especially, the low S_c), so that

$$\text{Etch depth(lat)} = \frac{(0.01 * 0.2 * 2.5 \times 10^{18} cm^{-2} sec^{-1} + 0)}{5.0 \times 10^{22} cm^{-3}} * 300 \text{ sec}$$

$$= 3 \times 10^{-5} cm = 0.3 \mu m$$

There is more etching in the vertical direction than in the lateral direction, and a profile such as that in Figure 10–3(b) occurs in this case.

This model has been used to accurately simulate the etching of Si under certain conditions. For example, Singh and colleagues [10.20] etched Si using a CF_4/O_2 plasma. As we noted earlier, this can result in isotropic etching, with a very strong chemical component. They used the etch system in the plasma mode, with symmetric electrode areas, which minimizes any ion bombardment. In addition, an oxide etch mask was used rather than photoresist so that the formation of a carbon-containing inhibitor layer was minimized. The etch profiles obtained were virtually isotropic and they were able to model them with the linear etch model in SPEEDIE and assuming no ion component. A value of 0.04 was used for S_c.

This same group [10.21] etched Si using an SF_6 plasma, but now observed a mixture of chemical and physical etching. Etch profiles similar to that in Figure 10–3(b) were obtained, with notable undercutting but not completely isotropic. They were able to accurately model the profiles again with the linear etch model, with a chemical-to-physical component ratio of 2:1. They actually identified the physical component as an "ion-assisted mechanism" as opposed to purely physical sputtering but modeled it with a simple linear dependence on ion flux, independent from the isotropic chemical flux. Ulacia and colleagues [10.24] did similar experiments etching Si using a SF_6/C_2ClF_5 plasma and modeled it with the linear etch model. (Again the physical component was identified as ion-enhanced, but following a linear relationship with ion flux.) They were able to vary the chemical-to-physical ratio from 4:1 (near-isotropic) to approximately 1:1 and then to 1:20 (almost perfectly anisotropic) by changing the SF_6:C_2ClF_5 ratio and the amount of carbon in the system, and model the resulting etch profiles quite well with the linear etch model for a specific set of conditions.

We now describe a series of plasma etch simulations using the SPEEDIE simulator and the linear etch model just described. As with the SPEEDIE deposition simulations in Chapter 9, we will use analytical $\cos^n\theta$ distributions for arrival angles (as opposed to Monte Carlo calculations of gas-phase processes). A value of n equal to 1 results in an isotropic arrival angle distribution, relative to the vertical direction. A value of n higher than 1 results in a more anisotropic arrival distribution, with a more directed downward distribution such as occurs with ionic species. In all the simulations presented in this section, we etch a silicon substrate using a photoresist mask. The top layer is 1-μm thick photoresist, with a slightly sloped window defined in it.

The first simulation uses the linear etch model, and Eq. (10.15) to govern the etch rate at each point. Pure chemical etching is illustrated first in Figure 10–27(a). The simulation time is 240 seconds, and each of the contours corresponds to a 30-second time step in the etching. The arrival angle distribution parameter, n, for chemical is set to one, indicating an isotropic arrival angle distribution. The maximum unobstructed direct flux of neutral species is set to 5×10^{18} cm^{-2} sec^{-1}. This corresponds to the total integrated direct flux of chemical species arriving from all directions and hitting a flat, unobstructed surface. Knowing the angular distribution of arriving atoms and the geometry of the surface, SPEEDIE can convert this to the direct chemical flux normal to the surface, $F_{c(direct)}$, at each point. The value of the maximum direct chemical flux used in the simulations is usually determined by knowing the chemical rate on a flat surface, in which case it is directly equal to F_c, or is determined by fitting experimental profiles to the simulations. The ion flux is set to zero in this simulation, so that the etch is purely chemical. A low value for the sticking coefficient, S_c, is used for the chemical species, equal to 0.01, ensuring that F_c is equal everywhere and the chemical component to etching occurs isotropically. A value of 0.2 is used for K_f. Negligible etching of the photoresist occurs in these simulations.

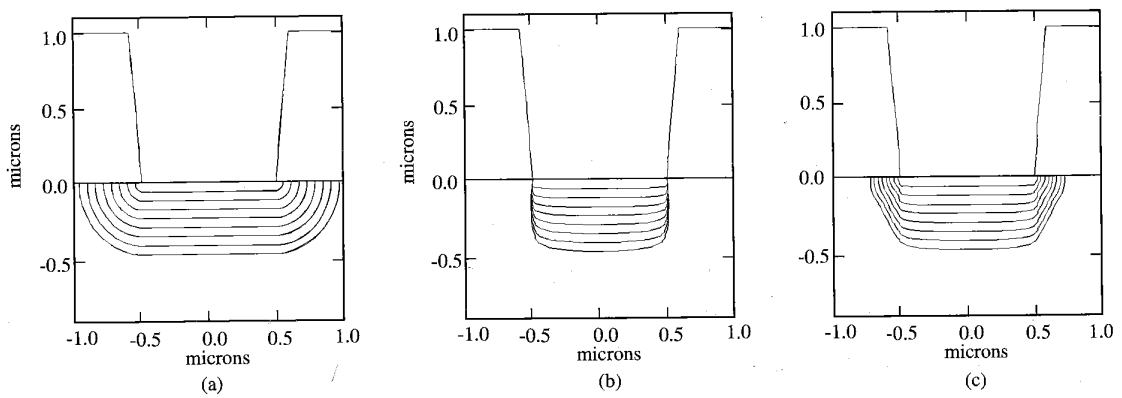


Figure 10–27 SPEEDIE simulations for etching of Si with linear etch model: (a) all chemical etching, showing isotropic etching; (b) all physical or ionic etching, showing very anisotropic etching; (c) half-chemical, half-ionic etching, showing a combination of isotropic and anisotropic etching. The different curves correspond to different times in the etch simulation.

Next only physical, ionic etching is illustrated using the same linear etch model. The chemical flux is set to zero while the maximum ionic flux is set to 1×10^{16} cm^{-2} sec^{-1}. Setting $n = 80$ means that nearly all the ionic flux will arrive from the vertical direction. K_i is set to 1. Figure 10–27(b) shows the results. Very anisotropic etching results as expected. Purely physical etching would occur in a sputter etch system (although one would expect etching of the resist as well in that case). This etch behavior could also occur, and be modeled in the same manner, for an ion-enhanced mechanism for conditions where the chemical component is not limiting the ion-driven etching anywhere in the structure, and where there is no spontaneous chemical etching.

Now we combine both chemical etching and physical/ionic etching in the linear etch model. As we discussed this model assumes that the two components are acting independently, or else in a regime where neither is limiting the action of the other. Here we use 2.5×10^{18} cm^{-2} sec^{-1} for the chemical maximum flux and 0.5×10^{16} cm^{-2} sec^{-1} for the ionic maximum flux, and the same values for S_c, K_f, and K_i we used above. This results in relatively equal contributions of the ionic and chemical components since the $S_cK_fF_c$ and K_iF_i products [see Eq. (10.15)] are equal. The resulting etch profile is shown in Figure 10–27(c). Different degrees of anisotropy can be achieved by varying the relative amounts of the two components. This behavior could occur in a plasma system in which there is both chemical etching and ionic sputtering acting separately. It could also occur where there is some spontaneous chemical etching, not affected by ion bombardment, as well as ion-enhanced etching for conditions in which the chemical component is not limiting the etching.

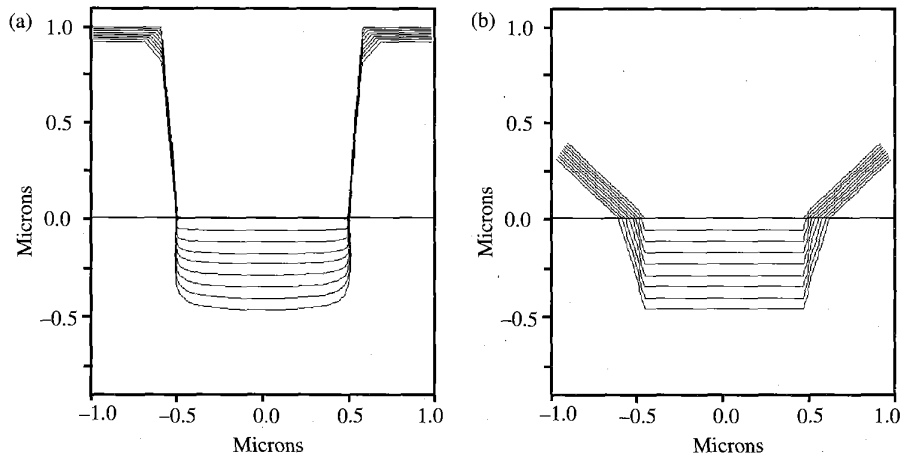


Figure 10–28 SPEEDIE simulations for etching of Si with linear etch model for physical-only etching with etching of photoresist included: (a) angle-dependent sputtering of photoresist included, showing faceting; (b) initial taper of photoresist changed, allowing for mask erosion due to vertical physical etching. A wider area of silicon is etched, and sloped sidewalls in the Si result.

Next we allow for some physical etching of the photoresist. We first allow for some angle-dependent physical sputtering of the photoresist during physical-only etching. We do this by increasing the value of K_i for the photoresist from 0 to 0.3. In Figure 10–28(a) we see the faceting that occurs at the top corners. Next we change the initial shape and thickness of the photoresist mask. We start with a thin mask, only 0.4 μm thick at the sides, which slopes diagonally to the opening. When physical etching occurs, the edge of the mask will be etched away, illustrating the phenomenon of mask erosion. Figure 10–28(b) shows the result. The etching away of the mask results in a wider etched area in the silicon, and nonvertical sidewalls. Changing the etch chemistry to change the amount of mask erosion can be utilized to obtain a desired etch profile.

We next illustrate the concept of Aspect Ratio Dependent Etching (ARDE). As discussed earlier this is the phenomenon in which the depth of etching decreases as the aspect ratio increases. One possible mechanism for this is that ion-neutral collisions in the sheath cause ions to be off-axis, making their arrival angle distribution wider. The less directed flux means that fewer ions reach the bottom of the etched trench due to view angle or shadowing considerations. With fewer ions getting to the bottom for the narrower openings, less vertical etching occurs, whether it is purely physical etching or ion-enhanced etching. This is simulated in the following two figures using the linear etch model. First we repeat the physical-only etching of Figure 10–27(b), except that we change the arrival angle distribution factor, n, for ions from 80 to 10, as might occur if the gas pressure were higher and more gas phase collisions take place, for example. This widens the arrival angle distribution a little, but the ions are still fairly directed downward. Figure 10–29(a) shows the result. More bowing of the sides results as compared to the $n = 80$ example in Figure 10–27(b), and this type of profile is seen in some etch

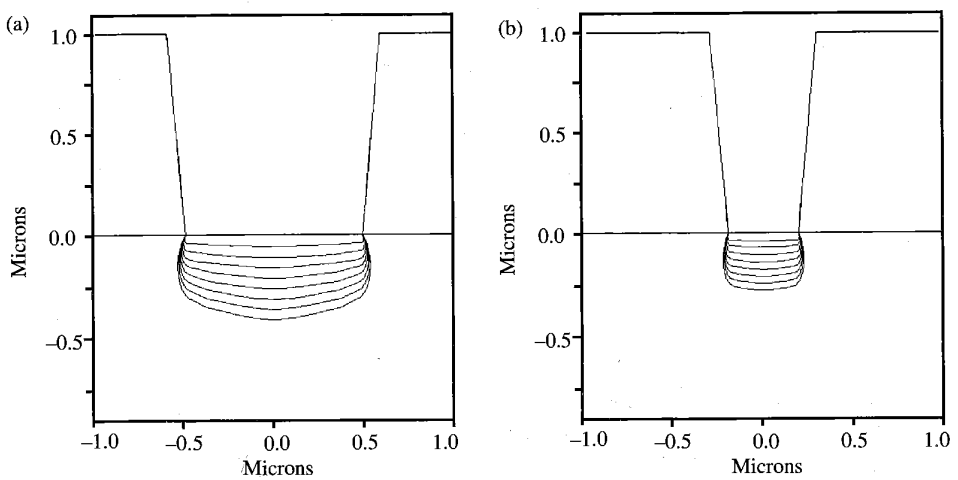


Figure 10–29 SPEEDIE simulations for etching of Si with the linear etch model illustrating Aspect Ratio Dependent Etching (ARDE). Physical-only etching (with arrival angle distribution factor $n = 10$) is done for two different mask opening widths: (a) mask opening = 1 μm; (b) mask opening = 0.4 μm. Deeper etching occurs for the wider mask opening.

systems. Now we decrease the window in the photoresist down to 0.4 μm. The resulting etch profile is shown in Figure 10–29(b). The depth of etching is clearly reduced. (The same effect is not seen if n were left at 80 for the two widths.) With the slightly wider angular distribution of arrival angles for the ions, the etch depth is less for the narrower mask opening than the wider opening even though all the other etch conditions are the same. Therefore, the etching could be different for different feature sizes on the same wafer, even if the features are right next to each other.

Interestingly enough, in some etch systems and under certain conditions, a reverse ARDE effect is sometimes observed where wider etch features etch less than narrower ones. Both forward and reverse ARDE has been simulated in SPEEDIE using the "Simuletchdepo" model, to be described later, which includes a modified saturation/adsorption etch term combined with a polymer deposition term [10.25]. In that modeling, the transport of the neutral species down the trench is important for the aspect ratio dependence.

In summary, the linear etch model is relatively simple, combining the chemical and physical components in a linear fashion. It can be used in a physically based simulator such as SPEEDIE to simulate etch profiles with different degrees of anisotropy, for a variety of etch systems. It can also model effects such as ARDE and mask erosion. While simple, the model may in fact be adequate for particular plasma etch situations, even for ion-enhanced etching under limited conditions. But it does not model the coupling between the chemical and physical components that can occur, particularly the synergistic interactions in ion-enhanced etching.

10.5.3 Etching Models — Saturation/Adsorption Model for Ion-Enhanced Etching

To account for the physical and chemical components acting together, and not independently, another model is needed. As was described earlier, the etching is often not just simply the sum of the two individual components, as in the linear etch model. In ion-enhanced etching the two components act in a synergistic manner, and the etch rate can be much higher than the sum of the two separate rates. Likewise, the etching can be limited by the deficiency or lack of either of the two components. Ion-enhanced etching can be due to a variety of mechanisms. Any one of the steps of the chemical etch process may be enhanced by ion bombardment, such as the reaction of the etchant and the surface material, or the removal of etch byproducts. Inhibitor layers, formed as a result of the etch process, may require ion bombardment to remove them.

What is needed in the model is an expression that allows for the etch rate to be larger than the sum of the two components acting independently. In fact, it may be the case that the presence of both components is needed for the etching to occur. Without any ion bombardment, for example, a byproduct layer may not be removed and no chemical etching can proceed. Without any chemical processes, the ion sputtering alone may be negligible. This means that the two components are acting in series, as opposed to in parallel. It also means that the deficiency of one component can limit the total etch rate, even as the other component is increased. This leads to saturation effects—increasing the flux of one of the components will increase the etch rate to a point, but then the rate will saturate due to the lack of enough of the other

component. These results have all been observed experimentally [10.6, 10.23, 10.26, 10.27, and 10. 28].

A model for ion-enhanced etching that captures the above behavior is the saturation etch model. This model is also commonly called the Langmuir adsorption or the saturation/adsorption etch model and includes surface coverage issues involving species adsorbing on and leaving the surface in its formulation. While we will derive a simple expression below based on one type of ion-enhanced etching mechanism (ion-enhanced byproduct removal), its results are generally applicable to many different mechanisms in which ion bombardment enhances one or more of the steps in the etch process. It is really a phenomenological model which can incorporate many simultaneous reactions and kinetics into its simple form. Many versions of this basic model have been developed to include a variety of processes and chemistries, but they are all similar in idea and approach [10.26, 10.28, 10.29, 10.30, 10.31, and 10.32].

For this simple analysis we consider both reactive neutral species and ions striking a surface during an etch process. The neutral species are adsorbed on the surface and chemically react with the surface material, forming etch products which contain both reactant species and film atoms. The ions striking the surface can enhance the removal of the etch byproducts from the surface, perhaps by sputtering or dislodging the loosely held byproducts or by inducing their conversion to more volatile species which then spontaneously desorb [10.6, 10.32]. A steady state will develop when the net flux of the neutrals to the surface is zero, or equivalently when the rate of byproduct formation equals the rate of byproduct removal. The idea is that the number of neutral species absorbed and reacted on the surface per time, according to simple Langmuir adsorption behavior, is proportional to how many sites are available: one minus the surface coverage. And the number of sites available depends on the neutral flux and on the ion bombardment which removes the byproducts and increases the sites for reactive neutral species adsorption. At steady state, the etch rate is then just proportional to the removal rate of byproducts, or equivalently to the neutral species adsorption rate, and depends on both the neutral and ion fluxes in a synergistic manner.

According to Langmuir adsorption theory, the number of reactive neutral species that adsorb (per unit time per area) on the surface to react with the surface material is

$$\#\text{ reactive neutrals adsorbed per time per area} = S_c(1 - \Theta)F_c \qquad (10.16)$$

where S_c is the sticking coefficient of the available area, Θ is the fraction of sites covered by the byproducts, so that $(1 - \Theta)$ is the fraction of sites available, and F_c is the flux of neutral species onto the surface. A chemical reaction constant, K, can be included, but it is just assumed to equal one here. We assume in this derivation that all neutral atoms that adsorb and stick on the surface will react quickly with the surface and form etch byproducts.

The reactive neutrals (in the etch byproducts) that are removed by ion bombardment per unit time per area from the surface is proportional to the byproduct coverage and equals

$$\#\text{ reactive neutrals (contained in byproducts) removed per time per area} = K_i\Theta F_i \qquad (10.17)$$

F_i is the incoming flux of ions and K_i is the ion-induced etch yield, or the effective sputtering yield (in general, a measure of how many atoms or molecules are removed per incident ion in a given process). K_i can be much larger than the normal sputtering yield of surface atoms—removing reaction products by ion bombardment may be much easier than removing film atoms by direct physical sputtering of the surface.

We assume a steady-state condition will develop in which the number of neutrals adsorbed equals the number removed. Thus equating Eqs. (10.16) and (10.17) gives

$$S_c(1 - \Theta)F_c = K_i\Theta F_i \tag{10.18}$$

$$\therefore \Theta = \frac{1}{1 + \dfrac{K_iF_i}{S_cF_c}} \tag{10.19}$$

The etch rate is then just the flux of removed species, equal to Eq. (10.17) [or equivalently, the flux of adsorbed and reacted species given by Eq. (10.16)], divided by the density of the film being etched, N. (Any stoichiometric coefficients needed to relate the ratio of film atoms to the neutral species in the reactant or byproduct molecules can be added or just included in S_c or K_i). Therefore by using Eq. (10.19) for Θ we obtain

$$\text{Etch Rate} = \frac{1}{N}\frac{1}{\left(\dfrac{1}{K_iF_i} + \dfrac{1}{S_cF_c}\right)} \tag{10.20}$$

By writing the etch rate expression in this form, the two components (analogous to electrical conductances) are seen to be acting in series rather than in parallel.

We plot Eq. (10.20) in Figure 10–30. For simplicity we assume the density, N, is unity. Here the x-axis is ion flux times the ion yield, or F_iK_i, and the equation is plotted for various values of the chemical flux times the sticking coefficient, F_cS_c. However, due to the symmetry, we could interchange F_cS_c and F_iK_i and obtain the same results. One sees that for no ion flux, the etch rate is zero, no matter what the chemical flux is. Similarly, for no chemical flux, the etch rate is zero, no matter what the ion flux is. One also sees that as the ion flux is increased for a given chemical flux, the etch rate initially increases linearly but eventually saturates. The value it saturates at, or approaches, is equal to F_cS_c/density. Similarly, when increasing the chemical flux at a constant ion flux, the etch rate increases linearly and then saturates. The process is limited by the slower of the two component processes.

The same type of saturation behavior is seen with respect to the effective sputtering yield term, K_i. Increasing the effective sputtering yield by increasing the ion energy will result in an increase in etch rate, which will then saturate as the ion energy is further increased.

The initial linear behavior is made clear by noting that when F_iK_i is small and F_cS_c is large, then by Eq. (10.20) the etch rate becomes approximately equal to just F_iK_i/density. This is the same as the physical term in the linear etch model and shows why the simple linear etch model can be used in some cases for ion-enhanced etching, as discussed earlier.

Since both ion bombardment and chemical reaction by the reactive neutral species are required in the saturation model, this ion-enhanced etching mechanism results in

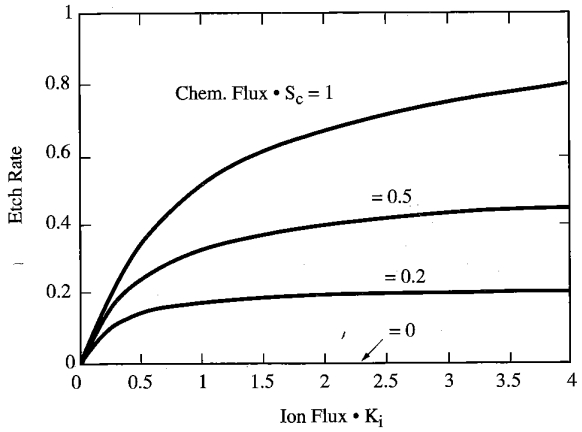


Figure 10–30 Etch rate versus ion flux for various values of the chemical flux, as calculated by Eq. (10.20). An identical plot can be obtained by interchanging F_c S_c and $F_i K_i$.

mostly vertical etching. This is because the ion bombardment occurs mostly in the vertical direction, with no reemission. On sidewalls (vertical surfaces), F_i is close to zero, and hence the etch rate according to Eq. (10.20) would be close to zero in lateral directions.

Example Silicon is plasma etched for 5 minutes using a photoresist mask and this time the process follows the saturation/adsorption, or ion-assisted, etch model. The flux of chemical species, F_c, on an unobstructed flat surface is equal to 2×10^{18} atoms cm^{-2} sec^{-1} and the sticking coefficient is 0.01. The flux of ionic species, F_i, on an unobstructed flat surface is equal to 2×10^{16} atoms cm^{-2} sec^{-1} and K_i is 1. There is no etching of the photoresist, and no spontaneous chemical etching of the Si. How far is the Si etched in the vertical direction (away from the mask edge), and how far is the Si etched in the lateral direction right under the mask?

Answer In the vertical direction, away from the mask edge, there are both chemical and ion fluxes, and the total etch rate would be given by Eq. (10.20). The density of Si is 5.0×10^{22} atoms cm^{-3}. Therefore

$$\text{Etch Rate (vert)} = \frac{1}{5 \times 10^{22}\ \text{cm}^{-3}} \frac{1}{\left(\dfrac{1}{1 * 2 \times 10^{16}\ \text{cm}^{-2}\text{sec}^{-1}} + \dfrac{1}{0.01 * 2 \times 10^{18}\ \text{cm}^{-2}\text{sec}^{-1}}\right)}$$

$$= 2.0 \times 10^{-7}\ \text{cm sec}^{-1}$$

For a 5-minute etch, the Si etched in the vertical direction is

$$\text{Etch depth(vert)} = 2.0 \times 10^{-7}\text{cm sec}^{-1} \times 300 \text{ sec} = 0.6\ \mu\text{m}$$

In the lateral direction, right under the mask edge, there is virtually no ion flux ($F_i \approx 0$), only chemical flux. Therefore the etch rate expression goes to zero. (And there is no spontaneous chemical etching.) Thus the Si etched in the lateral direction, right under the mask edge is virtually zero. A very anisotropic etch profile results.

The absence of lateral etching in this type of etching assumes that there is no spontaneous chemical etching, that is, no chemical etching without any ion bombardment. However, there may in fact be a degree of either spontaneous chemical or physical etching occurring in some systems. If the ion-enhanced etching occurs because the ion bombardment transfers energy to the surface atoms which speeds up the chemical reactions, there still may be appreciable etching without the enhancement. Or if the ion-enhanced etching is due to the presence of a byproduct layer whose removal is enhanced by ion bombardment, there may be significant spontaneous desorption of the byproduct layer. Likewise, there may be a finite component of spontaneous etching by ionic species for which neutral chemical species are not required, such as purely physical sputtering.

To account for any spontaneous etching, terms corresponding to pure chemical or pure physical etching can be included in the expression for ion-enhanced etching. In the simplest such model, terms the same as those used in the linear etch model are added to the ion-enhanced etching rate, with those terms scaled by some factor relative to the enhanced rate. For example, a ($S_C K_f F_c$/density) term can be added to Eq. (10.20) to provide for a spontaneous chemical etch component. Adding spontaneous terms would allow for some etching even if one of the two components were absent and could result in etch profiles that show some lateral etching in addition to the vertical etching [as in Figure 10–3(b)]. In more advanced models the spontaneous etch terms may be coupled with other terms, such as the surface coverage [10.26, 10.28, 10.30, 10.31, 10.32].

While Eq. (10.20) is a rather simple expression, and other more detailed versions of it have been developed, it captures the main features observed in ion-enhanced etching for many different mechanisms. Expressions the same as or very similar to this, most based on surface coverage or site considerations, have been used for a variety of ion-enhanced mechanisms. These include etching mechanisms where ion bombardment enhances the chemical etching by damaging the surface, where the sputtering yield of the surface is increased by the chemical reaction between the reactive neutral species and the surface atoms, as well as different byproduct removal mechanisms. (And it is often difficult to distinguish between the different mechanisms, both experimentally and conceptually.) In each of these cases, the analysis generally follows that as outlined above, with the ion bombardment term [Eq. (10.17)] corresponding to different ion-induced etch processes (damage creation, byproduct removal, etc.). But the end result is an expression for etch rate very similar to that given in Eq. (10.20), with the two components

acting in series and in a synergistic manner. The saturation/adsorption model, or variations of it, has been used to model a variety of etch systems, including Si, SiO_2, and W thin films, under a variety of conditions and chemistries [10.26, 10.27, 10.28, 10.29, 10.30, 10.32, and 10.33].

Something that this expression does not account for is an inhibitor mechanism in which an inhibitor/byproduct forms or deposits independently of the main etch species and reaction, such as when etching Al with Cl_2 and a separate inhibitor component. Nor does it model thick or excess inhibitor formation and the highly sloped etch profiles that can result. Modeling which explicitly includes these processes will be described later.

We now describe some SPEEDIE plasma etch simulations using the saturation/adsorption etch model, as described by Eq. (10.20). Here the chemical components (neutral reactive species, such as free radicals) and physical components (ions) work together. Only when and where both are present does etching occur. In the first simulation, approximately equal contributions from the chemical and ionic etching are specified by having the K_iF_i and S_cF_c products equal. Figure 10–31 shows the simulation, with each etch contour representing a time increment of 30 seconds in the 240-second total etch time. As expected in this model, very anisotropic etching is achieved even with a strong chemical component. This is because ion bombardment is necessary in this ion-enhanced mechanism, and the ions only strike in a near-vertical direction. Etch chemistries are often chosen that produce this synergy between chemical and ionic components.

The shape of the etched sidewall can be changed in the saturation/adsorption model by changing the n factor for the ion arrival angle distribution. A smaller value for n (closer to 1), corresponding to a less directed flux, gives a more bowed sidewall. A less directed flux can occur for a higher pressure or a lower wafer bias.

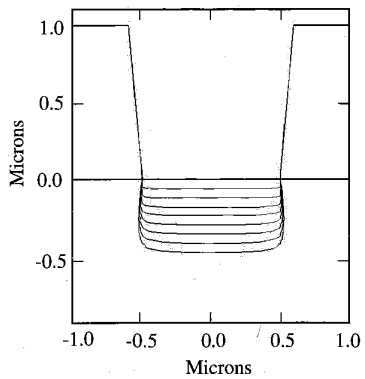


Figure 10–31 SPEEDIE simulations for etching of Si with saturation/adsorption etch model. Approximately equal components of chemical and physical etching are used.

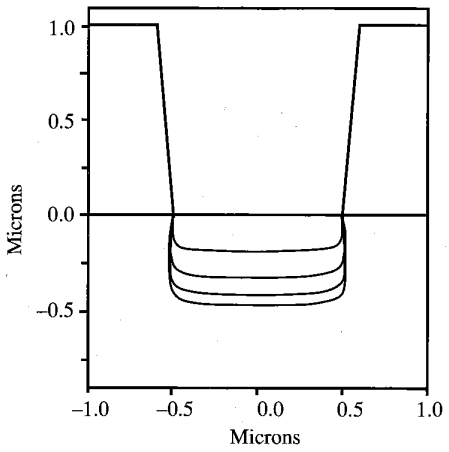


Figure 10–32 SPEEDIE simulations for etching of Si with saturation/adsorption etch model. The chemical flux is kept constant (at 2×10^{18} cm^{-2} sec^{-1}) while the ion flux is linearly increased from 0 to 1×10^{16} cm^{-2} sec^{-1} in equal increments. Each contour here corresponds to the final etch profile for the different ion fluxes. The etch rate is initially linear with ion flux but then starts to saturate.

Next we do a series of simulations with this same model in which the chemical flux is kept constant, but the ion flux is increased. Figure 10–32 shows the simulation profiles. The maximum chemical flux is kept at 2×10^{18} cm^{-2} sec^{-1} while the maximum ion flux is varied from 0 to 2×10^{16} cm^{-2} sec^{-1} in equal increments. S_e and K_i equal 0.01 and 2, respectively. Unlike the plots in the previous simulation figures which showed the different time interval profiles, only the final 240-second etch profile is shown for each simulation in this figure. In this case the different profiles correspond to the different maximum ion flux values used. One can see that for zero ion flux, no etching at all takes place. As the ion flux increases, the etching increases. At first the etch rate is nearly linear with ion flux, but as the ion flux is increased more, the etch rate starts to saturate and the increase in the etch rate slows down. Eventually the etch rate will not perceivably increase at all with an increase in ion flux. This illustrates the saturation effect we discussed previously for this model and illustrated in Figure 10–30.

A result very similar to Figure 10–32 would be produced if the chemical flux were increased while the ion flux remains constant. With no chemical flux, no etching at all occurs. As the chemical flux increases, the net etch rate increases linearly at first, then not as fast, and eventually saturates. The etching of silicon and silicon dioxide is often done with a high chemical flux, operating in or near the ion-limited, chemical species saturated regime. However at the bottom of high aspect ratio features, chemical-limited etching may occur due to inadequate transport of the neutrals in some cases [10.25].

10.5.4 Etching Models — More Advanced Models

There are other models used in topography simulators—some of them variations of the saturation/adsorption model and some of them quite complex—which attempt to

capture details of specific etch processes. Some of these more advanced models also try to directly address issues such as the sidewall inhibitor mechanism in ion-enhanced etching, and physical sputtering in sputter etch systems. We will not discuss these in detail but will briefly describe a few of them.

One such model in SPEEDIE attempts to separate the sidewall inhibitor effect from other ion-enhanced etch mechanisms in case two different effects may be dominating in different regions. For instance, there may be an ion-enhanced chemical reaction occurring at the bottom of the structure being etched where there is little inhibitor remaining, while the sidewall inhibitor may be affecting things on or near the sidewalls. This is implemented in the Sidewallpass model in SPEEDIE [10.17, 10.34] by modifying the saturation/adsorption model. The result is that the etch rate versus ion flux curve in Figure 10–30 is changed so that the left-hand portion, below some threshold level, drops faster than normal. The idea is that on or near the sidewalls, where the ion flux is low to begin with, it may be even lower due to the ions being consumed by the sputtering of a thin sidewall inhibitor layer. Zheng and McVittie [10.34] showed that in the case of etching polysilicon with a SF_6/Cl_2ClF_5 chemistry this model better matches the actual etch profile, with even less etching and bowing on the sidewalls where the ion flux is low. (However, this model does not explicitly account for thick or excess inhibitor formation, as we shall discuss below.) Other models have been developed which similarly consider multiple ion-enhancement mechanisms.

In other more advanced models, the formation and sputtering of a thick sidewall inhibitor layer are explicitly included. In these the sloped etch profiles due to excess inhibitor formation (as illustrated in Figure 10–14) are modeled, something that the simpler saturation models do not do. Examples are SPEEDIE's "Simuletchdepo Thick Sidewall Etch" model [10.17, 10.35] and Avant!'s TAURUS-TOPOGRAPHY's dry etch model [10.18], which include simultaneous deposition and etching of a polymer inhibitor layer. In SPEEDIE's model the polymer deposition is controlled by a CVD-like process with the inhibitor flux (the "deposited flux" in Figure 10–26) coming from the gas phase above the wafer, separate from the flux of reactive neutral etching species. Ion sputtering of the inhibitor layer then allows for etching of the substrate or film in the vertical direction. Such models can include ion-enhanced deposition and sputtering of the deposited material [like Eq. (9.58) in Chapter 9 for high-density plasma CVD deposition], as well as purely chemical etching of the inhibitor-free surfaces (for chlorine etching of Al, for example) or ion-enhanced chemical etching of the inhibitor-free surfaces (for etching of SiO_2 or Si, for example). Han and colleagues [10.35] have shown that this approach can model high-density plasma etching of silicon dioxide, using an ICP plasma source and C_2F_6 etchant gas. Ion-enhanced deposition and sputtering of the polymer with an ion-flux-dependent sticking coefficient, along with ion-enhanced etching of the SiO_2 (utilizing the saturation/adsorption expression), are used in this modeling.

Another advanced model similar to these was developed by Tuda and colleagues [10.36] for plasma etching of Si with Cl_2/O_2. This model includes passivation layer formation from surface oxidation and from etch byproducts arriving from the plasma as well as being directly redeposited from the bottom surface onto sidewalls. Other models include ion reflection, trenching, and localized surface charging. For HDP systems, a contribution from ionized chemical species to the chemical component of etching

might also be included. All these advanced models can obviously be quite complicated, with many parameters.

Finally, advanced simulators may also include a separate model for pure physical etching, or ion sputtering, which includes redeposition of the sputtered material [10.37]. The expression used for this in SPEEDIE [10.17] is

$$\text{Etch rate} = \frac{K_{sp}Y(\theta)F_i - K_{rd}S_{rd}F_{rd} - S_{cd}F_d}{N} \tag{10.21}$$

The first term is due to the sputtering, and the sputtering yield can be angle dependent. The second and third terms are due to redeposition of sputtered material. In particular the second term is due to the direct redeposition from every other point on the structure where sputtering occurs. The third term is from material that is sputtered off the surface but goes into the plasma and then drops back down to the wafer surface. The third term can be very difficult to calibrate. N is the density of the film being etched.

10.5.5 Other Etching Simulations

As we mentioned in Chapter 9, commercial suppliers of integrated circuit simulation tools offer both empirical and more physically based models. Here we present examples of both types. The first is from Avant!'s TAURUS-TOPOGRAPHY simulator [10.18], utilizing the dry etch model with polymer deposition. This model is similar to SPEEDIE's physically based Simuletchdepo model described above. Using direct and indirect flux calculations similar to those in SPEEDIE, TAURUS-TOPOGRAPHY simulates the deposition, reemission, and sputtering of polymer material that occurs in the formation of a sidewall inhibitor layer. In the example described here, SiO_2 is plasma etched over Si using a photoresist mask.

Figure 10–33 shows the simulated results of the sidewall polymer etch model. Figure 10–33(a) shows the simulated results after 0.9 minutes of etching. One can see the sidewall polymer that has formed. This layer inhibits lateral etching of the oxide. No polymer layer is present on the trench bottom because of strong ion bombardment and sputtering there, and also because of less polymer deposition due to shadowing. Figure 10–33(b) shows the simulation after 1.8 minutes of etching. The sidewall layer is now thicker, affecting the etch profile as described earlier and illustrated in Figure 10–14. A sloped etch profile, with no etch bias, results due to the high-polymer deposition-to-etch ratio.

Finally, some commercially available simulation tools provide simpler, more empirical models. Silvaco's ATHENA simulator (using the ELITE topography tool) [10.19] is an example and we will use it here to simulate etching of silicon through a photoresist mask. In these simpler models, the user specifies the etch rates of the different components and various parameters to obtain the shape of the etch profile, rather than specifying fluxes. There is both a "wet etch model" and a "RIE model" in ATHENA. The wet etch model is completely isotropic, while the RIE model contains elements of both the linear etch and saturation/adsorption models. We will briefly describe the RIE model here and provide some examples.

In the RIE model in ATHENA, the etching process is separated into three adjustable etch rate components: the isotropic, the directional, and the chemical. Using the

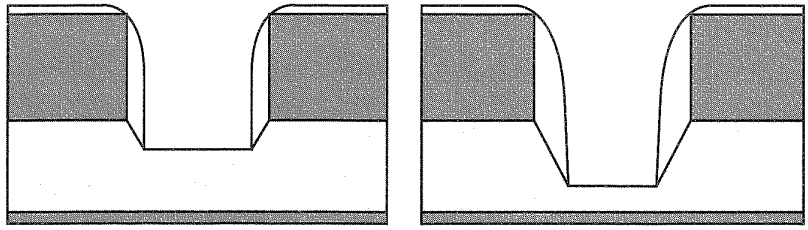


Figure 10-33 Avant!'s TAURUS-TOPOGRAPHY [10.18] simulation using their dry etch model with simultaneous polymer deposition. (a) shows the result of etching silicon dioxide over silicon with a photoresist mask after 0.9 minutes. Notice the sidewall polymer layer that has formed, inhibiting lateral etching. (b) shows the etch profile after 1.8 minutes. The sidewall layer is thicker and the etch profile in the oxide is sloped due to the high deposition-to-etch ratio.

first two is similar to using the linear etch model. The isotropic component, designated by r_{iso}, etches equally in all directions. The directional component, designated by r_{dir}, etches only in the vertical direction. For a sloped surface, where the etch rate normal to the surface is important, the etching due to r_{dir} is $r_{dir} * \cos\alpha$. Here α is the angle between the surface normal and the vertical direction. By using one or both of these two components, one can simulate purely chemical/isotropic etching, purely physical/directional etching, or a linear combination of the two. (As with the linear etch model, the physical component here may indeed be due to an ion-enhanced etching mechanism, but for which the chemical component is not limiting the process.) One can control the degree of etch anisotropy by changing the relative values of these two parameters, such that the degree of anisotropy is

$$A_f = \frac{r_{dir}}{r_{iso} + r_{dir}} \tag{10.22}$$

[One can show that this is equivalent to our definition of anisotropy in Eq. (10.5) if one observes that $r_{lat} = r_{iso}$ and $r_{vert} = r_{iso} + r_{dir}$).] No sticking coefficient is used, nor are arrival angle distributions. For these components, it is assumed that the etching is purely isotropic for the one component, and purely vertical for the other. One just specifies the etch rates for each component. The Taurus topography simulator from Avant! has a similar etch model in which one simply specifies the degree of anisotropy desired.

The third component of the ATHENA RIE model is called the chemical component. This implicitly assumes an ion-enhanced etch mechanism. When one specifies the chemical component etch rate in the RIE model, one must also specify the "divergence" parameter. The divergence parameter determines the standard deviation of the ion's Gaussian angular distribution function. It is assumed that this chemical etching only occurs where an ion strikes. In this simple way both the chemical and ionic components of the ion-enhanced mechanism are taken into account in this model; however saturation

effects are not accounted for. By varying the divergence, different shaped sidewalls are obtained, very similar to changing the n factor in the $\cos^n\theta$ ion arrival angle distribution used earlier. In ATHENA, all three components—the isotropic, the directional, and the chemical—can be specified, resulting in almost any shaped profile imaginable.

In the first series of ATHENA etch examples, silicon is etched through a photoresist mask using each of the three RIE model components. Figure 10–34(a) shows the initial structure. For the first example the silicon is etched with a purely isotropic component. An isotropic etch rate, r_{iso}, of 5 nm per second is specified. The time of the etch is set at 2 minutes, which should etch the silicon 600 nm deep. The resulting profile is shown in Figure 10–34(b). In the next simulation, pure physical or directed etching is done. This is done by specifying r_{dir} to be 5 nm per second and r_{iso} to be zero. The simulated profile is shown in Figure 10–34(c). In the final simulation, the ion-enhanced chemical etching component is used. This is done by designating the chemical component to be 50 and the divergence to be 5. The resulting profile is in Figure 10–34(d). Changing the divergence parameter is similar to changing the arrival angle distribution (the n value) of ions in the earlier models, due to changes in gas pressure, for example.

The final series of ATHENA simulations are shown in Figure 10–35. Here different combinations of isotropic, directional, and chemical (ion-enhanced) components are

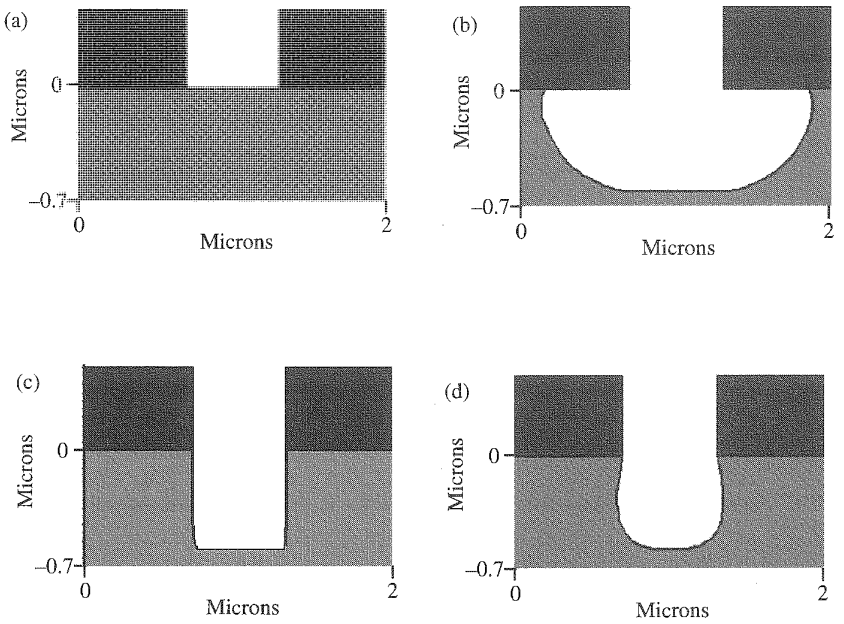


Figure 10–34 Silvaco's ATHENA [10.19] topography simulation of silicon etching using different etching components in the RIE model: (a) initial structure with photoresist mask over Si substrate; (b) purely isotropic etching; (c) purely directional; (d) ion-enhanced etching .

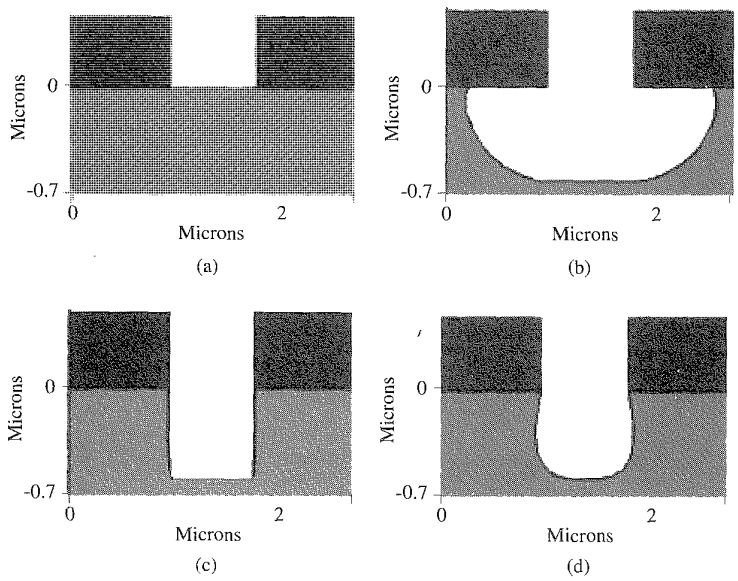


Figure 10–35 Silvaco's ATHENA [10.19] etch simulation using different etching component combinations in the RIE model: (a) chem = 50, div = 10, iso = 0, dir = 0; (b) chem = 30, div = 10, iso = 0, dir = 20; (c) chem = 30, div = 10, iso = 20, dir = 0; (d) chem = 30, div = 10, iso = 10, dir = 10. These illustrate the variety of etch profiles that are obtainable with this easy-to-use tool.

used to obtain a wide variety of etch profiles. These simulations demonstrate the usefulness of such simulation models. It is relatively simple to use them and calibrate them against actual etch profiles. These types of models are often used to characterize a particular etch system and etch chemistry used to plasma etch a particular film. One can build up a library of etch equipment, materials, etch chemistry, and plasma conditions to use in future simulations. This is sometimes done by commercial suppliers of these simulation tools and a model parameter library for various commonly used pieces of deposition and etch equipment with model parameters for various films and normal operating conditions is included with the simulator. One can then use these models and parameters to optimize an etch or deposition process without being involved with all the details. These types of simulations also run very quickly.

Of course the main advantage of the more physically based models is that they try to be more closely linked to actual physical processes. This is especially true for the models currently being developed which actually model the etch-enhancement mechanism, such as inhibitor layer formation and removal. As a result, these types of models are more apt to be valid for conditions outside those for which they are calibrated and tested, but both types of models are useful for different applications.

10.6 Limits and Future Trends in Technologies and Models

With the ever-decreasing feature sizes in future generations of integrated circuit technology, the challenges in etching will certainly continue. Etching features with smaller and smaller dimensions will obviously pose problems. Damage to the underlying films and active regions will also be issues as the dimensions decrease and plasma processing is used. High-density plasma etch systems will surely help in these regards. A high-density plasma at low gas pressures and the decoupling of the plasma density and ion energy mean that fast, anisotropic etching can be achieved while minimizing the ion-induced damage. Development of this type of etch technology will surely continue.

With ever-increasing wafer sizes in the future, uniformity and loading problems will be harder to control. Improvements in equipment design will have to continue to overcome these issues. More *in situ* control will be required. In addition to endpoint detection, automatic feedback will be needed to control equipment and process parameters such as gas flow, pressure, and power to the plasma sources.

As with deposition equipment, a trend in etch equipment will be the use of cluster tools. By connecting to multiple etch (and deposition) chambers and passing the wafers from one chamber to another, one can perform several processing steps sequentially without breaking vacuum. This minimizes exposure of the wafers to atmosphere, reducing particulate contamination and surface oxidation between steps. For example, one can have chambers to etch TiN and Al, do a fluorine passivation, and strip the photoresist. This becomes more important as the trend continues toward multilevel stack structures. One can also combine deposition and etch chambers in such a tool. Most of these tools are single wafer, as opposed to batch systems. More control over each wafer is possible, but the etch rates must be high since each wafer must be processed individually.

There will be more challenges in etching specific films as the dimensions decrease and aspect ratios increase. For example, in Al metal etching the issues of corrosion, resist and residue removal, Cu byproduct removal, loading effects (both macro and micro), and of course profile control and selectivity will have to be addressed each time the structure dimensions are decreased. As we have seen, there are many issues and considerations involved in developing an etch process for each film, and it will be harder and harder to satisfy all the requirements.

New plasma etching techniques may be used to meet the increasing demands as the technology progresses. Recently a new plasma etch process, called the Bosch process, has been developed in the MEMS (Micro-Electro-Mechanical Systems) industry for etching very deep Si trenches. This uses fluorine chemistry in a high-density plasma, and both excellent anisotropy and selectivity with respect to SiO_2 masks are reported [10.38]. In this process, polymer deposition and etching steps are sequentially alternated in an HDP etch system. Whether this can be adapted in the future for etching isolation structures or polysilicon gates in VLSI structures is unclear, but new plasma etching methods along these lines may be developed and used.

Etching and patterning processes other than plasma etching may be used more in the future. Chemical-Mechanical Polishing (CMP) has found widespread use in the past few years and is now a mainstream process. As described in the next chapter, this uses a combination of wet chemicals and mechanical grinding to polish or etch the top surface

of a film, leaving a very planar surface. Dielectric films and W plugs are now often planarized by this technique. When CMP is combined with the "damascene" fill and etch-back process, also described in the next chapter, materials such as copper that are difficult to plasma etch can be patterned to integrated circuit specifications.

New materials will be added in the future. As we discuss in the next chapter, new metal interconnect materials such as copper and new dielectrics such as low-k dielectrics will soon be required. There has already been extensive development on integrating these types of materials into the fabrication process, including the patterning of these films. Etching some of these films, especially Cu, has been found to be very difficult, so much so that other methods for patterning these films such as CMP are more attractive.

The models and simulation tools for etching processes will also undergo changes and improvements. As we mentioned earlier in the chapter, we need to model the actual processes more deeply, such as explicitly modeling the sidewall inhibitor formation and removal and phenomena such as ion reflection and charging. Work in these directions has begun. And as with the deposition models, physically based models and simulation tools need to be combined with user-friendly and efficient software tools.

10.7 Summary of Key Ideas

We have seen in this chapter how etching is used to pattern films by selectively removing material with the use of mask layers, usually photoresist but sometimes other thin films. Etching can be either wet or dry, the former using liquid chemical etchants and the latter using gas-phase, plasma-produced etch species. Some of the issues that are important in etching are selectivity with respect to masks and underlying films, directionality (anisotropic versus isotropic etching) for profile control, mask erosion, etch bias or undercutting, etch uniformity, residue removal, and damage in the underlying layers caused by etching.

In wet etching, chemical reactions dominate the etch process. This leads to very selective but isotropic etching. In dry or plasma etching, two types of species are present that are important in the etch process: reactive neutral chemical species (such as free radicals) and ionic species. The first etches by chemical processes and is isotropic or nearly isotropic and can be very selective, while the second etches by more physical processes and is very directional, or very anisotropic, but not very selective. The three main types of etch mechanisms that occur in plasma etching are chemical etching (involving the neutral chemical species), physical etching or sputtering (involving the ion species), and ion-enhanced etching (involving both in a synergistic way). Ion-enhanced etching can occur by the ions enhancing the chemical etch process or by removing inhibitor layers. Directional etching can be achieved in this manner, as well as selectivity, a combination usually desired in an etch process. Many types of plasma etching equipment have been developed that utilize the different etching mechanisms, including barrel etchers, plasma mode etchers, Reactive Ion Etching (RIE) systems, High-Density Plasma (HDP) systems, and sputter etching equipment. These systems range from purely chemical to purely physical etching, with many systems operating in between these two. Where both chemical and physical components are present, as in RIE and HDP systems, plasma conditions and etch chemistries are usually selected so that the

two etch components act synergistcally, and ion-enhanced etching occurs giving directional etching and good selectivity.

Most simulation models focus on the two components of etching, the chemical and physical. Advanced simulators model the etch process based on the fluxes and surface processes of the chemical and ion species. As in deposition modeling, the ion species have a narrow arrival angle distribution (being very vertical to the wafer) and a high sticking coefficient. The chemical species have a wide arrival angle distribution and low sticking coefficient. The two components can act independently or synergistically. In the linear etch model, the chemical and physical components of etching are combined in an independent linear fashion, and a wide range of etch profiles can be produced for different plasma etch systems and for certain conditions. In the saturation/adsorption model for ion-enhanced etching, the two species act synergistically, as described by the expression for two processes acting in series. Very directional anisotropic etching results, along with saturation effects. Variations can occur, such as when some spontaneous, non-ion-assisted chemical etching occurs in parallel, multiple ion-enhanced mechanisms are present, or when excess inhibitor material forms, and modifications or additional models can be utilized to capture these features.

In the future, structures will get even smaller and new materials will be used. New etching tools and processes will also evolve. The models and simulation tools will evolve as well. As with deposition models, they will become based even more on the actual physical processes.

10.8 References

[10.1]. C. J. Mogab, "Dry Etching," in *VLSI Technology*, ed. by S.M. Sze, McGraw-Hill, 1983, p.307.

[10.2]. W. S. Ruska, *Microelectronic Processing*, McGraw-Hill, 1987, p. 206.

[10.3]. C. Chang, "Experiments and Simulation of Plasma Deposition and Sputter Etching Processes," Ph.D. thesis, Stanford University, 1996, p. 93.

[10.4]. J. P. McVittie, private communication.

[10.5]. J. W. Coburn and H.F. Winters, "Ion and Electron Assisted Gas-Surface Chemistry—An Important Effect in Plasma Etching," *J. App. Phys.*, vol. 50, p. 3189, 1979.

[10.6]. J. W. Coburn, "Surface-science Aspects of Plasma-assisted Etching," *Appl. Phys. A*, vol. A59, p. 451, 1994.

[10.7]. R. A. Morgan, *Plasma Etching in Semiconductor Fabrication*, Elsevier Science, 1985.

[10.8]. M. Kimizuka, Y. Watanbe, and Y. Ozaki, "Pattern Profile Control in Magnetron Reactive Ion Etching of Poly-Si," *J. Vac. Sci. Tech. B.*, vol. 10, p. 2192, 1992.

[10.9]. H. C. Jones, R. Bennett, and J.Singh, "Size Dependent Etching of Small Shapes" in *Proc. of the Eighth Symp. on Plasma Prcoessing*, G. S. Mathad and D.W. Hess, eds. the Electrochem. Soc., 1990, p. 45.

[10.10]. J. P. McVittie, unpublished data.

[10.11]. M. A. Leiberman and A.J. Lichtenberg, *Principles of Plasma Discharges and Materials Processing*, J. Wiley & Sons, 1994.

[10.12]. B. Gorowitz and R. J. Saia, "Reactive Ion Etching," in *Plasma Processing for VLSI*, N. G. Einspruch and D. M. Brown, eds., Academic Press, 1984, p. 307.

[10.13]. W. R. Runyan and K. E. Bean, *Semiconductor Integrated Circuit Processing Technology*, Addison-Wesley, 1990, p. 276.

[10.14]. D.L. Flamm, in *Plasma Etching, an Introduction*, D.M. Manos and D. L. Flamm, eds., Academic Press, 1989.

[10.15]. L. M Ephrath and E.J. Petrillo, "Parameter and Reactor Dependence of Selective Oxide RIE in CF_4 + H_2," *J. Electrochem. Soc.*, vol. 129, p. 2282, 1982.

[10.16]. P. Singer, "Plasma Etch: a Matter of Fine Tuning," *Semiconductor International*, December 1995, p. 65.

[10.17]. SPEEDIE is a 2D etch and deposition simulator developed at Stanford University.

[10.18]. TAURUS-TOPOGRAPHY is a commercially available topography simulator developed by Avant! Inc.

[10.19]. ATHENA and its ELITE topography simulator are commercially available from Silvaco Inc.

[10.20]. V. I. Singh, E. S. G. Shaqfeh, and J. McVittie, "Study of Si Etching in CF_4/O_2 Plasmas to Establish Surface Re-Emission as the Dominant Transport Mechanism," *J. Vac. Sci. Tech. B*, vol. B12, p. 2952, 1994.

[10.21]. V. I. Singh, E. S. G. Shaqfeh, and J. McVittie, "Simulation of Profile Evolution in Si Reactive Ion Etching with Re-Emission and Surface Diffusion," *J. Vac. Sci. Tech. B*, vol. B10, p. 1091, 1992.

[10.22]. D. L. Flamm, D. M. Donnelly and J. A. Mucha, "The Reaction of Flourine Atoms with Silicon," *J. Appl. Phys.*, vol. 52, p. 3633, 1981.

[10.23]. B. A. Heath, "Selective Reactive Ion Beam Etching of SiO_2 over Polycrystalline Si," *J. Electrochem. Soc.*, vol. 129, p. 396, 1982.

[10.24]. J. I. Ulacia and J. P. McVittie, "A Two-Dimensional Computer Simulator for Dry Etching Using Monte-Carlo," *J. Appl. Phys.*, vol. 65, p. 1484, 1989.

[10.25]. J. S. Han and J. P. McVittie, "Modeling of Aspect Ratio Dependent Etching in an Inductively Coupled Plasma," *Mat. Res. Soc. Symp. Proc.*, vol. 389, p. 197, 1995.

[10.26]. V. M. Donnelly, D. L. Flamm, W. C. Dautremont-Smith, and D. J. Weber, "Anisotropic Etching of SiO_2 in Low-frequency CF_4/O_2 and NF_3/Ar Plasmas," *J. Appl. Phys.*, vol. 55, p. 242, 1984.

[10.27]. R. Petri and D. Henry, "Tungsten Etching Mechanisms in Low-Pressure SF_6 Plasma," *J. Appl. Phys.*, vol. 72, p. 2644, 1992.

[10.28]. T. M. Mayer and R. A. Barker, "Simulation of Plasma-Assisted Etching Processes by Ion-Beam Techniques," *J. Vac. Sci. Tech.*, vol. 21, p. 757, 1982.

[10.29]. J. Pelletier and M. J. Cooke, "Microwave Plasma Etching of Si and SiO_2 in Halogen Mixtures: Interpretation of Modeling Mechanisms," *J. Vac. Sci. Tech.*, vol. B7, p. 59, 1989.

[10.30]. E. Zawaideh and N. S. Kim, "Plasma Etching Model Based on a Generalized Transport Approach," *J. Appl. Phys.*, vol. 62, p. 2498, 1987.

[10.31]. M. A. Lieberman and A. J. Lichtenberg, *Principles of Plasma Discharges and Materials Processing*, John Wiley & Sons, 1994.

[10.32]. D. C. Gray, I. Tepermeister, and H. H. Sawin, "Phenomenological Modeling of Ion-Enhanced Kinetics in Flourine-Based Plasma Etching," *J. Vac. Sci. Tech. B*, vol. 11, p. 1243, 1993.

[10.33]. J. P. Chang, A. P. Mahorowala, and H. H. Sawin, "Plasma-Surface Kinetics and Feature Profile Evolution in Chlorine Etching of Polysilicon," *J. Vac. Sci. Tech. A*, vol. 16, p. 217, 1998.

[10.34]. J. Zheng and J. P. McVittie, "Modeling of Side Wall Passivation and Ion Saturation Effects on Etching Profiles," *Proc. of IEEE Int'l. Mtg. Numerical Modeling of Process and Device for I.C.: NUPAD-V*, p. 37, 1994.

[10.35]. J. S. Han, J. P. McVittie, and J. Zheng, "Profile Modeling in High Density Plasma Oxide Etching," *J. Vac. Sci. Tech*, vol. 13, p. 1893, 1995.

[10.36] M. Tuda, K. Ono, and K. Nishikawa, "Effects of Etch Products and Surface Oxidation on Profile Evolution During Electron Cyclotron Resonance Plasma Etching of Poly-Si," *J. Vac. Sci. Technol. B*, vol. 14, p. 3291, 1996.

[10.37]. C. Y. Chang, J. P. McVittie, and K. C. Saraswat, "Backscattering in Ar Sputtering of Oxide," *Appl. Phys. Lett.*, vol. 63, p. 2294, 1994.

[10.38]. J. Bhardwaj, H. Ashara, and A. McQuarrie, "Dry Silicon Etching for MEMS," *Proc. of Symp. on Microstructures and Microfab. Sys.*, Electrochem. Soc., 1997.

10.9 Problems

10.1. Why might pure chemical etching, such as in wet etching, be adequate for patterning the silicon nitride layer used to define the field oxidation area in the CMOS process flow given in Chapter 2?

10.2. If the etch anisotropy is 0, what is the undercut or etch bias when etching a 0.5-μm thick film? What is the undercut when the anisotropy is 0.75? Assume no overetch in each case.

10.3. In a certain process, it is desired that the pitch of metal lines be equal to or less than 1.0 μm (the pitch equals one metal linewidth plus one spacing between metal lines, measured at top of features). Assume that the metal linewidth and spacing are equal (that is, 0.5 μm each). The height of such structures is also 0.5 μm, and the minimum lithographic dimension is 0.25 μm.

a. What minimum degree of anisotropy is needed in an etch process in order to produce such a structure?

b. What minimum pitch could be obtained for such a structure with *wet* etching? (Again with minimum lithograph dimension of 0.25 μm, thickness of 0.5 μm, and equal metal width and spacing.)

10.4. What are the advantages and disadvantages of reactive ion etching versus sputter etching? Cite a hypothetical example of when you might want to use sputter etching rather than RIE.

10.5. Explain how loading effects can affect endpoint detection.

10.6. It is found that a certain plasma etch chemistry in a certain RIE etch system produces vertical sidewalls with zero etch bias when etching a particular film. Adding chemical A to the etch chemistry results in nonvertical sidewalls, and an etch bias. Adding chemical B to the original etch chemistry results in nonvertical sidewalls, but with zero etch bias. Explain what may be going on.

10.7. a. In a particular etch process, if selectivity is the biggest concern, which type(s) of etch equipment should be used?

b. If the biggest concern is ion bombardment damage, which type(s) of etch equipment should be used?

c. If the biggest concern is obtaining vertical sidewalls, which type(s) of etch equipment should be used?

d. If the biggest concerns are selectivity *and* vertical sidewalls, which type(s) of etch equipment should be used?

e. What about selectivity *and* vertical sidewalls *and* damage, while maintaining reasonable etch rate?

10.8. It is observed that the sidewall slope in an etch process becomes more sloped as the temperature is reduced. Why?

10.9. If the anisotropy of an etch process is 0.45, sketch the etch profile. What percentage of the etch rate in the vertical direction is due to the spontaneous chemical etching and what percentage is ionic/physical, assuming a linear etch mechanism? State all assumptions.

10.10. For a particular plasma etch process in which the linear etch model is applicable, a degree of anisotropy of 0.8 or better is desired. If the unobstructed ionic flux on a flat surface is 1×10^{16} atoms cm^{-2} sec^{-1} (with K_i equal to 1), what unobstructed chemical flux would result in an anisotropy of 0.8? For this process S_C is 0.01 and K_f is 0.1.

10.11. We want to see how the etch rate in the vertical direction might depend on pressure assuming that the etch follows the saturation/adsorption model. Assume that for a particular etch system that the chemical flux is directly proportional to the pressure, while the ion flux is inversely proportional to the pressure. That is $F_c = F_c'*P$ and $F_i = F_i'/P$. (P is normalized to 1 atm and unitless.) Also assume that density = 1 atom nm^{-3}, and that $K_iF_i' = S_cF_c' = 1$ atom nm^{-2} sec^{-1}.

a. Plot the vertical etch rate versus pressure, P, from $P = 0$ to 10.

b. Repeat with $K_iF_i' = 40$ atoms nm^{-2} sec^{-1} and $S_cF_c' = 1$ atom nm^{-2} sec^{-1}.

10.12. In an etch process, there is a finite amount of purely chemical etching without any ion bombardment (i.e., spontaneous chemical etching). In addition, ion bombardment greatly increases the etch rate by facilitating the breaking up of the etch precursor. At high ion flux the etch rate saturates. No etching occurs when there is only ion bombardment with no chemical component.

a. Write a generalized etch rate equation that can describe this behavior.

b. Sketch an etch rate versus ion flux curve for this process for some nonzero chemical flux.

c. Sketch what the etch profile might look like for this process (i.e., etching through a window in a mask as in Figure 10–3).

Back-End Technology

11

11.1 Introduction

Back-end technology refers to the interconnect layers, contacts, vias and dielectric layers that wire the active devices into specific circuit configurations. Figure 11–1 is a schematic diagram taken from Chapter 2 showing these components in a typical integrated circuit structure. Figure 11–2 is a SEM micrograph of an IBM circuit showing the back-end structure with the dielectric layers etched away, revealing the interconnects, contacts, and vias. Another SEM of this same chip was shown in Figure 1–12. The relative importance of back-end structures has greatly increased in recent years because the circuit delays associated with interconnects have not kept pace with the faster device speeds provided by scaled technologies. Thus much recent effort has been devoted to improving back-end technology and it is widely believed that these efforts will only accelerate in the future.

Interconnects can be either local or global. In general, local interconnects are the first, or lowest, level of interconnects. They usually connect gates, sources, and drains in MOS technology, and emitters, bases, and collectors in bipolar technology. In MOS technology a local interconnect, polycrystalline silicon, also serves as the gate electrode material. Silicided gates and silicided source/drain regions and materials like TiN and W can also act as local interconnects. Local interconnects can tolerate higher resistivities than global interconnects since they are not very long. But they must also, in most cases, be able to withstand higher processing temperatures because they are deposited earlier in the process flow than global interconnects. Global interconnects, mostly made of Al, are generally all the interconnect levels above the local interconnect level. They often cover large distances between different devices and different parts of the chip and therefore are always low-resistance metals.

Ohmic contacts connect an interconnect with active regions or devices in the silicon substrate. A high-resistivity dielectric layer, usually silicon dioxide, separates the active regions from the first-level global interconnect, and electrical contact is made between the interconnect and the active regions in the silicon through openings in that dielectric layer. At the same time in the processing, contact can be made between the first-level global interconnect and the local interconnects, since they are also separated by that same dielectric layer. Connections between two levels of global interconnects are usually given a different name—vias. They are made through openings in the different levels of the intermetal dielectrics (IMDs), which separate interconnects from each other and from the active areas and devices. The bottom dielectric layer through which the contacts are made is usually called the first-level dielectric. Examples of these structures are shown in Figures 11–1 and 11–2.

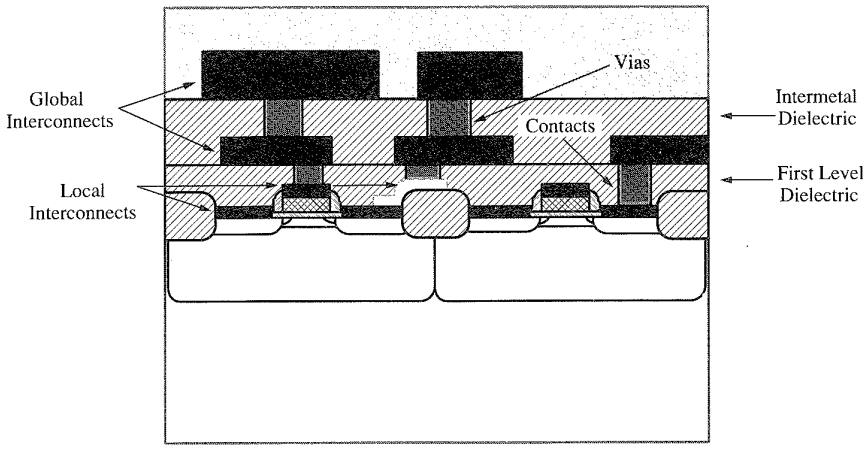


Figure 11–1 Schematic cross section of back-end structure, showing interconnects, contacts, and vias, separated by dielectric layers (cross-hatched regions).

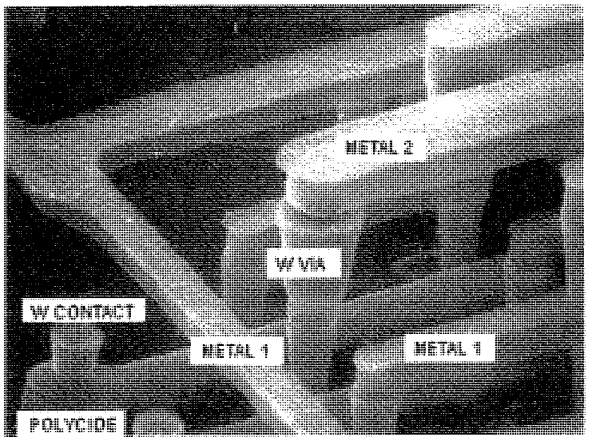


Figure 11–2 SEM micrograph of an IBM circuit showing the interconnects with the dielectric layers etched away. Photograph courtesy of Integrated Circuit Engineering Corporation.

It is easy to see why interconnect structures are becoming more important in overall chip performance. Consider the structure in Figure 11–3 [11.1]. The RC time delay of a signal propagating along such a line is to first order obtained by treating it as a distributed, unterminated transmission line. For such a system the time delay, τ_L, is approximately equal to $0.89RC$ where R is the line resistance and C is the total capacitance associated with the line [11.2]. A more accurate analysis would include such elements as the load capacitance, driver resistance, and line inductance [11.3, 11.4, 11.5]. However, the $0.89RC$ expression for the time delay which considers only the interconnect RC component is sufficient for our purposes.

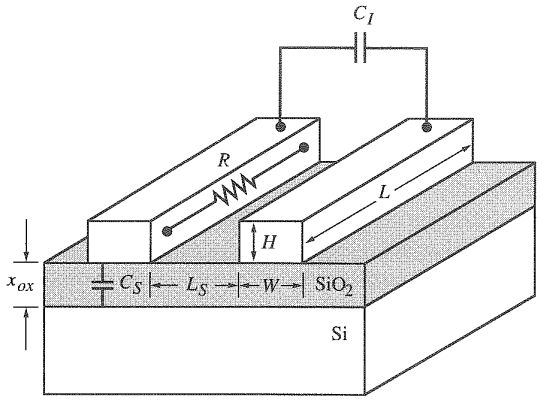


Figure 11-3 Interconnect structure for *RC* analysis. The two lines on top are the metal interconnects of dimensions *W*, *L*, and *H*, sitting on a SiO_2 layer. There is also SiO_2 between and above the metal lines. (After [11.1].)

The line resistance, in Ω, of one of the interconnects is given by

$$R = \rho \frac{L}{WH} \tag{11.1}$$

where ρ is the interconnect resistivity, and L, W, and H are the interconnect length, width, and height, respectively. The total capacitance associated with the line is

$$C = K_{ox}\varepsilon_o \frac{WL}{x_{ox}} + K_{ox}\varepsilon_o \frac{HL}{L_S} \tag{11.2}$$

where x_{ox} and K_{ox} are the oxide thickness and dielectric constant, respectively, and ε_o is the permittivity of free space. The first term represents the line to substrate capacitance, C_S, and the second term the coupling capacitance between adjacent lines, C_I. It is assumed that the lines are surrounded by oxide on all sides.

The total *RC* delay associated with the line is thus

$$\tau_L = 0.89 \cdot K_I K_{ox}\varepsilon_o \rho L^2 \left(\frac{1}{Hx_{ox}} + \frac{1}{WL_S}\right) \tag{11.3}$$

where K_I is added to empirically account for fringing fields and other interconnects above and below the line in multilayer interconnect systems. K_I is often taken to be ≈ 2.

Now we must consider what happens to the dimensions of the structure as the technology progresses. First, L_S and W are often close to the minimum feature size F_{min}. As the technology progresses, F_{min} gets smaller. The values of L_S and W may in fact be larger than the minimum feature size for some interconnect levels, but they still commonly scale with F_{min}. Both x_{ox} and H also shrink as F_{min} shrinks in this ideal scaling scheme, keeping the aspect ratio, *H/W*, constant. In the original analysis by Saraswat and Mohammadi [11.1], it was assumed that x_{ox} and H equaled $0.35F_{min}$ and $0.25F_{min}$, respectively. However, these thicknesses have not decreased as quickly as this in recent years, especially for the global interconnects. To keep the analysis simple, we assume

that x_{ox} and H, as well as L_S and W, are all equal to F_{min}. Thus Eq. (11.3) with K_I equal to 2, becomes

$$\tau_L = 3.56 \cdot K_{ox}\varepsilon_o\rho \frac{L^2}{(F_{min})^2} \tag{11.4}$$

For local interconnects, L usually also shrinks as F_{min} shrinks. Therefore, the net result is that the RC time delay for local interconnects stays approximately constant according to this scaling scheme. However, for global interconnects the length usually increases rather than shrinks. This is because the chip area of each new technology generation usually keeps increasing, forcing the global interconnects to increase in length to connect all the devices. The average length of the longest global interconnect in a circuit can be approximated by [11.1]

$$L_{max} = \frac{\sqrt{A}}{2} \tag{11.5}$$

where A is the chip area. Inserting this into Eq. (11.4) finally leads to the following expression for the delay associated with global interconnects:

$$\tau_L = 0.89 \cdot K_{ox}\varepsilon_o\rho \frac{A}{(F_{min})^2} \tag{11.6}$$

Example Calculate the interconnect delay time according to Eq. (11.6) for the following conditions: Al interconnect ($\rho = 3.0 \times 10^{-6}\ \Omega$ cm), SiO_2 dielectric ($K_{ox} = 3.9$), area $= 100\ mm^2$, and the minimum feature size of 0.35 μm.

Answer

$$\tau_L = 0.89 \cdot K_{ox}\varepsilon_o\rho \frac{A}{(F_{min})^2}$$

$$= 0.89 \cdot 3.9 \cdot 8.86 \times 10^{-14}\ \text{Fcm}^{-1} \cdot 3.0 \times 10^{-6}\ \Omega\,\text{cm} \frac{1\text{cm}^2}{(0.35 \times 10^{-4}\ \text{cm})^2}$$

$$= 7.5 \times 10^{-10}\ \text{sec} = 0.75\ \text{nsec}$$

As technology progresses, F_{min} shrinks and A usually gets larger. Both of these lead to an increase in the RC time delay in global interconnects. Even if A stays constant, the time delay will increase due to F_{min} shrinking. Physically this occurs because as the interconnect cross-sectional area decreases, R increases, but C does not decrease in proportion because the dielectric thicknesses surrounding the line also decrease. Hence RC

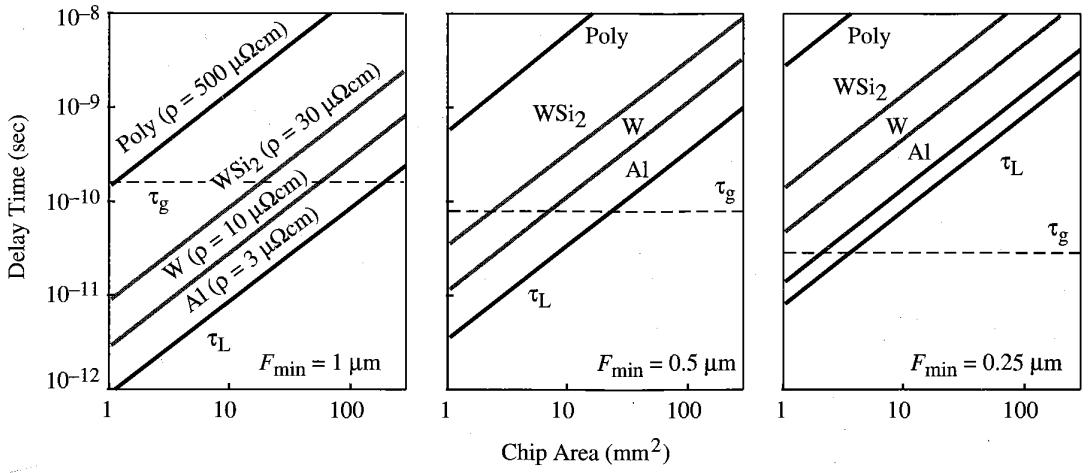


Figure 11-4 Global interconnect and gate time delay versus chip area and minimum feature sizes, F_{min}, for various interconnect materials. Interconnect delay corresponds to longest global interconnect in the circuit. (Updated from [11.1].)

increases. Having A increase only exacerbates this problem. Figure 11–4 plots Eq. (11.6) as a function of chip area for different values of F_{min} and different interconnect materials, updated from the original figure in [11.1] to use the new expressions for x_{ox} and H and to include the F_{min} = 0.25 μm case. Cu is also included as a possible interconnect material in the 0.25-μm technology. The interconnect delay here is based on the longest global interconnect using Eq. (11.6), with minimum feature size dimensions and assuming that the driver or load elements do not contribute significantly to the overall time delay. Also plotted for comparison is a typical gate delay, τ_g, for each minimum feature size, as measured in ring oscillator circuits. The gate delay is dependent upon F_{min} but not A and decreases as the minimum feature size decreases. It is obvious that for larger F_{min}'s, the gate delay is generally large relative to the delay for Al global interconnects. (And one can also see why one would not want to use polysilicon as a global interconnect.) But for smaller minimum features and larger chip areas, the global interconnect delay becomes quite large compared to the gate delay and can have a large impact on the circuit performance.

In 0.25 μm (=F_{min}) manufacturing technology, chip areas are typically 200–350 mm^2 (corresponding to a maximum global interconnect length of 7–9 mm), and gate delays as measured in ring oscillators are generally about 30–70 picoseconds. The RC delay of the longest global interconnect line, with minimum feature size dimensions, by this analysis is about 3–5 nanoseconds for Al lines, or about 80 times the gate delay. This does not necessarily mean that circuit performance is totally dominated by the interconnects in today's integrated circuits. The specific circuit architecture, the exact heights, widths and lengths of the interconnect lines on the different levels, the driver and load components, and other circuit factors will determine the speed of the circuit and the relative importance of the front-end and back-end elements. What is important

are the trends that these and other analyses predict: While the gate delays will keep getting smaller, the local interconnect delays will stay about the same, and the global interconnect delays will get larger and have a larger impact on the circuit performance. Recent analysis by Bohr based on Intel's technology agree with the qualitative predictions of this simple analysis [11.6, 11.7].

Each new technology generation in the NTRS represents a $\sqrt{2}$ reduction in feature size. This corresponds to a 2x increase in RC interconnect delay per generation for a given line length [Eq. (11.4)]. However, the metal line aspect ratio is not typically kept constant but is increased with each technology generation. This is primarily due to the fact that the metal line thickness has stayed rather constant in recent times rather than shrinking in order to keep the resistance lower. This means that the interconnect delay does not increase quite as fast as Eq. (11.4) predicts, but it is still increasing while the gate delay is decreasing. It has been estimated [11.6, 11.8] that in current CMOS technology with Al lines the delay due to the interconnects can be about 30–40% of the total circuit delay, up from about 15–20% just a generation previously. In the next technology generations, the interconnect delay could be more than 50% of the total circuit delay, emphasizing the importance of the back-end structures.

As feature sizes shrink, the size of the contacts is also getting smaller. Since to first order the contact resistance, R_c, is inversely proportional to the contact area, the importance of the contacts increases as the circuit dimensions are shrunk.

The specific projections for back-end technology given by the NTRS are shown in Table 11–1. A number of the qualitative observations made earlier are apparent in this table. Note the increasing numbers of wiring levels in the future and the increasing aspect ratios associated with interconnect lines as well as contacts and vias. Finally, notice the decreasing metal resistivity requirement which cannot be met by aluminum in the future, and the decreasing dielectric constant required for intermetal dielectrics, which cannot be met by SiO_2. Clearly new materials will have to be introduced into back-end

Table 11–1 Future projections for back-end technology taken from the SIA NTRS [11.9]

Year of First DRAM Shipment	1997	1999	2003	2006	2009	2012
Minimum Feature Size, F_{min} (nm)	250	180	130	100	70	50
DRAM Bits/Chip	256M	1G	4G	16G	64G	256G
DRAM Chip Size (mm^2)	280	400	560	790	1120	1580
MPU Chip Size (mm^2)	300	360	430	520	620	750
Wiring Levels—Logic	6	6–7	7	7–8	8–9	9
Min Metal CD (nm)	250	180	130	100	70	50
Min Contact/Via CD nm	280/360	200/260	140/180	110/140	80/100	60/70
Metal Aspect Ratio	1.8	1.8	2.1	2.4	2.7	3.0
Contact Aspect Ratio (DRAM)	5.5	6.3	7.5	9	10.5	12
Via Aspect Ratio (logic)	2.2	2.2	2.5	2.7	2.9	3.2
Metal Resistivity ($\mu\Omega$-cm)	3.3	2.2	2.2	2.2	<1.8	<1.8
Interlevel Metal Dielectric Constant	3.0–4.1	2.5–3.0	1.5–2.0	1.5–2.0	<1.5	<1.5

structures to meet these objectives. Cu interconnects and "low K" dielectrics are the likely answer to these requirements.

Metallization structures are also receiving increased attention because of failure and reliability issues. Failures are due to a variety of causes: stress-induced void formation in interconnect lines; hillock growth; unwanted chemical reactions and interdiffusion between layers; dopant diffusion and segregation; excessive roughening of films; corrosion; electromigration; and others. Many of these will be discussed in more detail later in the chapter.

Finally it should be pointed out that in the schematic figures in this section we illustrated each metal or dielectric level as simple single-layer structures. In practice, each level or layer is often a multilayered structure itself. The trends to multilayered back-end structures and the use of new processing techniques have been due to the ever-increasing demands on both processing stability and electrical performance as the technology progresses.

11.2 Historical Development and Basic Concepts

The earliest integrated circuits in the 1960s consisted of aluminum lines connecting active regions—transistors and resistors—in silicon. Silicon dioxide passivated the silicon surface and isolated the aluminum interconnects from those regions in the substrate where electrical contact was not wanted. Where contact was desired, openings in the oxide were made so that the aluminum could come in contact with the silicon. Figure 11–5 shows an early structure.

Pure aluminum was thus used for both the contact material and the interconnect. Why aluminum? First, it has a low electrical resistivity at room temperature. (See Table 11–2 for a listing and properties of commonly used interconnect materials). Second, it adheres well to silicon and silicon dioxide. Third, it makes good electrical contact to heavily doped Si. The latter two characteristics are related to the fact that the oxide of Al is very stable, more stable than SiO_2. This means that if Al is deposited on top of SiO_2, it will react with it even at low temperatures, forming a thin layer of Al_2O_3 at the interface. This forms a glue layer, binding the Al to the SiO_2. In the case of Al-to-Si contacts, the Al will reduce any native Si oxide on top of the Si. Without this reaction, the ubiquitous native oxide on the silicon could prevent ohmic contact. Fortunately, the Al_2O_3 that is formed allows Al to diffuse through it into the Si, so that good ohmic contact between the Al and Si is obtained. A fourth benefit of using Al is that its presence

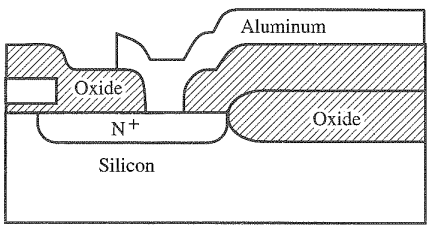


Figure 11–5 Early back-end structure with Al interconnect making contact to highly doped Si active region.

Table 11–2 Properties of interconnect materials

Material	Thin film resistivity (μΩcm)	Melting point (°C)
Al	2.7–3.0	660
W	8–15	3410
Cu	1.7–2.0	1084
Ti	40–70	1670
PtSi	28–35	1229
$TiSi_2$	13–16	1540
WSi_2	30–70	2165
$CoSi_2$	15–20	1326
NiSi	14–20	992
TiN	50–150	~2950
$Ti_{30}W_{70}$	75–200	~2200
polysilicon (heavily doped)	450–1000	1417

assists the annealing out of interface traps at the Si-SiO_2 interface, presumably by converting H_2O to free H, which passivates the traps. (See Chapter 6 where this process is described in more detail.) For all these reasons, contacts and interconnects made of aluminum were instrumental in the development of the first integrated circuits.

It was not long, though, before problems arose. As the circuits became more complicated, the structures became smaller, and the processing got more involved, new demands were placed on the metallization schemes. As a result, new metallization strategies, materials, and structures were developed. Even today the evolution continues, as evidenced by the changing requirements for interconnects in Table 11–1. We will now follow this evolution from the early days of integrated circuit fabrication, first with the contacts, then with the interconnects, and finally with the dielectrics.

11.2.1 Contacts

Contacts provide low-resistance connections between the metal interconnects and active device regions. Metal/semiconductor contacts can in general be rectifying (Schottky) or ohmic. Figure 11–6 shows the I-V characteristics and band diagrams for these modes of current transfer through a contact. The barrier ϕ_B, in the ideal case is strongly dependent on the work function of the metal. Nature provides a wide variety of metal work functions so in principle it should be possible to choose a metal which produces a small ϕ_B contact to silicon device regions. In practice, this is not generally possible, first because there are other requirements on which metals we may choose (resistivity, for example), and second, because surface states at the metal/semiconductor junction tend to pin the Fermi level deep in the silicon bandgap. This generally produces high ϕ_B contacts to N regions and relatively low ϕ_B contacts to p-type silicon.

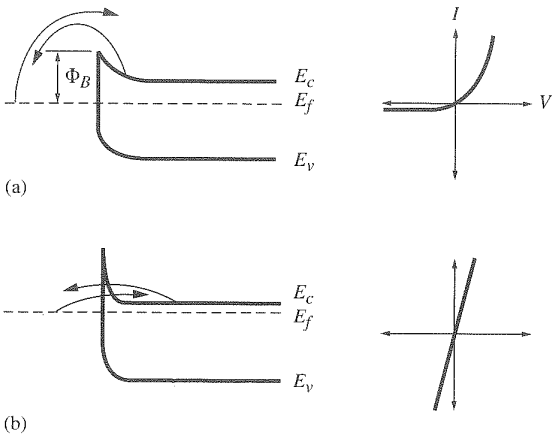


Figure 11–6 Band diagrams and I-V characteristics for metal/semiconductor junctions or contacts, illustrating different modes of current transfer: (a) Schottky or rectifying contact; (b) tunneling or ohmic contact.

Thermionic emission, shown in Figure 11–6(a), in which thermal energy allows the carriers to surmount the barrier, does allow some current to flow over a Schottky barrier. The basic properties of Schottky diodes operating in this mode are similar to the PN diode described in Chapter 1. However, for most metals the barrier is too high to n-type Si, resulting in rectifying behavior with high contact resistance for low to moderately n-doped contacts.

The other way for the carriers to get past the barrier is to tunnel though it [Figure 11–6(b)]. If the silicon is heavily doped, the width of the barrier, defined by the depletion region, x_d, becomes narrow enough that many of the carriers can quantum-mechanically tunnel through the barrier. Ohmic behavior, in which current can flow virtually unimpeded in both directions through a contact, can thus be achieved. This occurs when the doping is greater than about 6×10^{19} cm^{-3}.

The specific contact resistivity, ρ_c, is defined as

$$\rho_c \equiv \left[\frac{\partial V_{MS}}{\partial J}\right]_{V_{MS}=0} \tag{11.7}$$

where J is the current density across the metal/semiconductor interface and V_{MS} is the voltage across it. For a Schottky contact governed by thermionic emission over the barrier, the current density is given by

$$J_S = A^* T^2 \exp\left(\frac{-q\phi_B}{kT}\right)\left(\exp\left(\frac{qV}{kT}\right) - 1\right) \tag{11.8}$$

where A* is Richardson's constant. The specific contact resistivity, in Ω cm^2, is therefore

$$\rho_c = \frac{k}{qA^*T}\exp\left(\frac{q\phi_B}{kT}\right) \tag{11.9}$$

For a tunneling contact, the analysis is somewhat more complicated, resulting in

$$J_s \propto \exp[-2x_d\sqrt{2m^*(q\phi_B - qV)/\hbar^2}] \tag{11.10}$$

where $\hbar$ is Planck's constant, m^* is the effective mass of the tunneling carrier, and x_d is the depletion layer through which the tunneling must occur, which is given by

$$x_d = \sqrt{\frac{2\varepsilon_s\phi_B}{qN_d}} \tag{11.11}$$

where ε_s is the semiconductor permittivity and N_d is the doping concentration in the semiconductor. When x_d is less than about 2.5 nm, tunneling can occur efficiently. Substituting Eqs. (11.10) and (11.11) into (11.7) leads to the contact resistivity, in $\Omega\,\text{cm}^2$, for tunneling contacts:

$$\rho_c = \rho_{co}\exp\left(\frac{2\phi_B\sqrt{m^*\varepsilon_s}}{\hbar\sqrt{N_d}}\right) \tag{11.12}$$

where ρ_{co} is a constant dependent on the metal and semiconductor. The contact resistivity goes down as $\sqrt{N_d}$ goes up because x_d shrinks and tunneling becomes more efficient. For x_d to equal 2.5 nm or less in Si with an Al contact and result in ohmic behavior, N_d must equal $\approx 6 \times 10^{19}\ \text{cm}^{-3}$. The contact resistance, in Ω, is the contact resistivity divided by the contact area A, assuming that J is uniform over the contact area.

Figure 11–7 shows experimental data for specific contact resistivity as a function of N_d, along with the theoretical prediction based on the preceding equations. At low doping, thermionic emission controls the process which is essentially independent of doping, while at high doping, ρ_c is dominated by tunneling which is a very strong function of the doping level. Measured ρ_c values may be higher than indicated in Figure 11–7 if a clean interface is not obtained, due to the presence of unetched material or etch residues in the contact hole, native oxide present at the metal/semiconductor interface, or other contaminants or reaction products. Another common reason for the contact resistivity being higher than expected is that there are fewer dopants in the Si at the surface than desired. This may be due to segregation of dopants into the contact materials above the Si.

Example The expression for specific contact resistivity for high doping levels where tunneling dominates, Eq. (11.12), can be written in the form

$$\rho_c = \rho_{co}\exp\left(\frac{\phi_B C_1}{\sqrt{N_d}}\right)$$

where $C_1 \approx 7.0 \times 10^{10}\ \text{cm}^{-3/2}\,\text{V}^{-1}$ for Si. If $\phi_B = 0.6$ V and $\rho_{CO} = 1.0 \times 10^{-7}\ \Omega \cdot \text{cm}^2$ for a Si-metal contact, how much does the specific contact resistivity decrease when the doping is increased from 1×10^{19} to $1 \times 10^{20}\ \text{cm}^{-3}$?

Answer At $N_D = 1 \times 10^{19}\ \text{cm}^{-3}$,

$$\rho_c = \rho_{co} \exp\left(\frac{\phi_B C_1}{\sqrt{N_d}}\right) = 1.0 \times 10^{-7} \exp\left(\frac{0.6 \cdot 7.0 \times 10^{10}}{\sqrt{1 \times 10^{19}}}\right)$$

$$= 5.9 \times 10^{-2}\ \Omega\ \text{cm}^2$$

At $N_D = 1 \times 10^{20}\ \text{cm}^{-3}$,

$$\rho_c = \rho_{co} \exp\left(\frac{\phi_B C_1}{\sqrt{N_d}}\right) = 1.0 \times 10^{-7} \exp\left(\frac{0.6 \cdot 7.0 \times 10^{10}}{\sqrt{1 \times 10^{20}}}\right)$$

$$= 6.7 \times 10^{-6}\ \Omega\ \text{cm}^2$$

which is about a factor of about 9000 lower. These numbers are in agreement with the values in Figure 11–7.

The general requirements for metal-to-silicon contacts are low contact resistance and good thermal stability. Low contact resistance means a high-dopant concentration

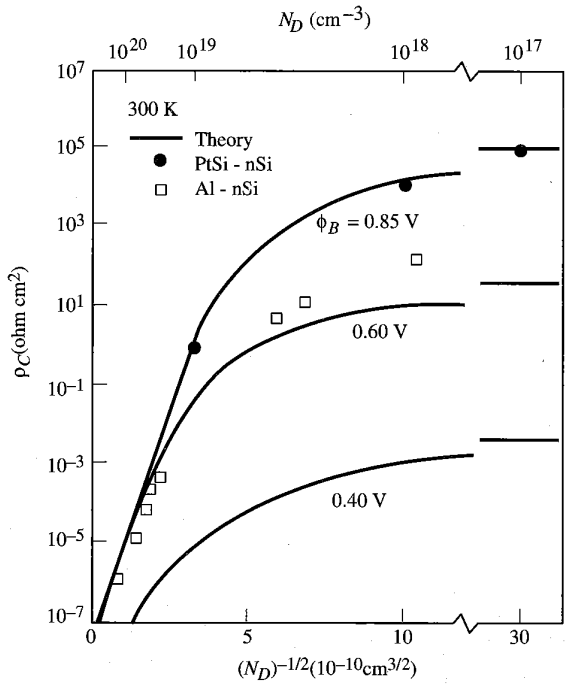


Figure 11–7 Contact resistivity (ρ_C) versus doping for three values of the metal/semiconductor work function. Both theoretical curves, based on Eqs. (11.9) and (11.12), and experimental data are plotted. (After [11.10, 11.11].)

at the Si surface and a good contact between the metal and silicon. Good contact requires an interface free of contaminants and residues. Good thermal stability means that the contact structure does not degrade during subsequent thermal processing, nor does it cause other areas to be adversely affected, such as active areas in the silicon underneath the contact.

A problem with thermal stability resulted in the first major change in the ohmic contact metallurgy. As mentioned previously, early integrated circuits used pure Al as the interconnect material, making direct contact to the Si substrate in heavily doped regions. Al makes a good contact to silicon by reducing the native oxide on the surface, which also removes any other contaminants on the surface. To ensure that the Al reduces the native Si oxide and that the Al is in good physical contact to the Si, a sintering anneal at about 450°C is done. This anneal, in a hydrogen ambient, is also done to anneal out interface traps at the Si-SiO_2 interface in the gates. Subsequent processing at about this same temperature can also occur, such as the deposition of an overlying dielectric or passivation layer.

However, the solubility of Si in Al is significant. At 450°C, the solubility is about 0.5 atomic percent, and at 500°C it is about 1 atomic percent. This means that pure Al in contact with Si will want to absorb the Si from the substrate up to its solubility level for that temperature. The only things limiting this process would be the diffusion of Si in the Al and the amount of Al acting as a sink for the Si. Unfortunately, the diffusivity of Si in polycrystalline Al is quite high, and the amount of Al nearby is large, so a significant amount of Si gets drawn up into the Al. This creates voids in the remaining Si, which are quickly filled by the overlying Al. If the Al penetrates uniformly into the substrate, typical processing conditions and contact/interconnect dimensions would result in about 0.2–0.3 μm of penetration. However, the situation is actually worse than that. That is because when the Al reduces the native oxide, it does so nonuniformly. Therefore the Al penetrates in localized spots. Al "spikes" occur in the Si substrate, penetrating in excess of 1 μm deep. Figure 11–8 illustrates this, showing the Al penetrating in some areas deeper than the diffused junction. This can short the junction. As a practical matter, pure Al contacts to Si cannot be used for junction depths of less than 2–3 μm.

One solution to the Al spiking problem is to use Al films that already have Si in them. Therefore the solubility requirement is already fulfilled. So Al films with approximately 1 atomic percent Si have been commonly used. This is the solubility of Si in Al at about 500°C, the highest temperature the Al sees during processing. The spiking problem is reduced, but another problem arises. When the contact structure is cooled down below

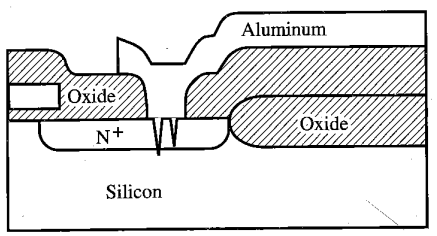


Figure 11–8 Si-Al contact region showing spiking of Al into Si active region. This is due to Si diffusing into the Al to satisfy the solubility requirement, with Al filling the resulting voids in the Si substrate.

500°C, or subsequent thermal steps are done at less than 500°C, the solubility of Si in Al is less. The now-excess Si precipitates out, usually at the interface where heterogeneous nucleation can occur, leaving Si nodules. In fact, there is enough Al dissolved in these nodules to make them p-type (Al is a p-type dopant in silicon), causing higher contact resistance to N^+ regions.

The use of Al(Si) to solve the spiking problem was adequate as long as the contact resistance was not limiting circuit performance. But as contact areas were reduced in size and the contact resistance did become an issue, another solution to Al-Si interaction and spiking was needed for contacts to small junctions. Barrier layers provided the answer. These layers are placed between the Al and Si to prevent or at least slow down any interaction. Such barriers need to be a barrier for chemical interdiffusion between Si and Al at processing temperatures (up to 450–500°C). They also need to be thermally stable. They should have low stress. This usually means that their coefficient of thermal expansion needs to be close to Si ($2.6 \times 10^{-6}\,°C^{-1}$). They need to adhere well to Si and Al, as well as SiO_2. This usually means that there is some interfacial reaction between the barrier and Si and Al. Finally, they should have good electrical conductivity and a low contact resistance to both Si and Al.

Barrier layers have been classified by Nicolet [11.12] into three types: passive barriers, stuffed barriers, and sacrificial barriers. The passive barriers would be chemically inert to both Si and Al and be good diffusion barriers as well. TiN is an example of this type. Some materials are chemically inert, but diffusion along their grain boundaries can be quite significant. However, it has been found that by "stuffing" the grain boundaries with other atoms or molecules, the diffusion of other species can be greatly reduced, making them diffusion barriers as well. The reduction of diffusion by stuffing the grain boundaries is a combination of a physical effect (physically blocking the diffusing species) and a chemical effect (the stuffed species, such as nitrogen or oxygen, chemically bond with the diffusing species, such as Al, stopping their diffusion). A Ti-W alloy with nitrogen stuffing the grain boundaries—achieved by sputtering the Ti-W in an N_2 ambient—is an example of a stuffed barrier. Sacrificial barriers are those in which the barrier material is "sacrificed" in order to prevent reaction between the Al and Si. The material, such as Ti, reacts with either the Al or Si or both, forming aluminides or silicides. But by doing so, it prevents the Al and Si from interacting since the reactions with the barrier material are favored both kinetically and thermodynamically over the reaction between the Al and Si. However, once the sacrificial barrier material is consumed, the barrier is gone, and so these types of barriers are only effective for limited processing temperatures and times.

One of the first barrier layers for contacts was platinum silicide. It is formed by depositing Pt on the wafer, and then annealing at low temperatures (250–350°C). Because the surface Si is consumed during the formation reaction, a new, clean interface is obtained, and very good electrical contact is made. It is also "self-aligned" because the silicide is only formed where the Si is exposed through the oxide contact opening (and the unreacted Pt is chemically removed from other areas). Therefore an extra mask is not needed and no alignment overlap is required. PtSi acts satisfactorily as a barrier between Al and Si, but only if subsequent processing is done at about 400°C or below, and junction depths underneath are greater than about 0.3 μm. That is because PtSi will

react with Al, forming Pt-Al intermetallic compounds and eventually allowing the Al to spike into the Si substrate. Once junction depths got shallower than 0.3 μm and processing temperatures got above 400°C, another barrier was necessary.

Refractory metals by themselves have had limited use as barriers. Ti in particular is attractive since it reduces and breaks through any native oxide and adheres well to both Si and SiO_2. It forms very low resistive contacts, silicon diffusion through it is slow, and therefore it prevents the Si dissolving in the Al. However, it reacts with Al to form $TiAl_3$ at low temperatures. (Above 500°C, Ti will also react with the Si to form $TiSi_2$.) It thus acts as a sacrificial barrier, acting as a barrier only until it is consumed, which happens rather quickly above 400°C. So while Ti is a good contact/adhesion layer, it is not a very good barrier layer. Ti is thus often used in conjunction with a barrier layer on top.

Refractory metal silicides, such as $TiSi_2$, have also been used as barrier layers in contacts. They will also react with Al, but at a somewhat higher temperature (about 500°C for $TiSi_2$) than the noble metal silicides, such as PtSi. However, diffusion along their grain boundaries limits their effectiveness as barriers to below 400°C. Therefore, like Ti, refractory metal silicides are now usually used in contact structures as the adhesion/contact layer, with a better barrier layer on top. An advantage in using refractory metal silicides in contacts is the ability to form self-aligned contact structures. Ti can be deposited over both an oxide layer and into the openings in the contacts without a mask, but during an anneal it will only form a silicide in the contacts where Si is exposed. The unreacted Ti over the oxide can be etched off, leaving the silicide self-aligned to the contact.

In fact, the entire source and drain regions can be contacted by a refractory metal silicide this way, providing for local interconnects. By contacting the entire source and drain regions with silicide, rather than just in the contact holes, the contact resistance is also improved by providing a much larger contact area between the silicide and silicon. This is because the silicon/silicide contact resistance is much larger than silicide/barrier/metal contact resistances, and the net contact resistance thus depends on the silicon/silicide contact area.

As the backend processing temperatures were raised and junction depths decreased in later technology generations, the next commonly used barrier after PtSi was Ti-W. This alloy, generally about 30 atomic percent, or 10 weight percent Ti, acts as a diffusion barrier between the Si and Al, especially when stuffed with nitrogen. While pure W is a good barrier itself, adding Ti helps it adhere to underlying layers and increases its corrosion resistance. Ti-W will react with Al at higher temperatures, about 600°C, while diffusion through the film limits its use to below 500°C. In addition, diffusion of Ti from the film into the Al at 400°C and above has been seen to raise the Al resistivity. Its contact resistance, however, is not as low as PtSi, possibly due to its difficulty in dissolving the native oxide on the Si surface. Thus a bilayer of PtSi and Ti-W has been generally used, taking advantage of the low contact resistance of the PtSi and the barrier properties of Ti-W. One problem with Ti-W is its relatively high stress. It is often very brittle, leading to cracking and peeling on the wafer and to flaking from chamber walls onto the wafer. [11.13] In addition, there is the problem of diffusion of Ti from the film into Al at 400°C and above, raising the resistivity of the Al interconnect line.

Refractory metal compounds, notably TiN, have become very popular barrier layers. These are commonly fine-grained structures, with grain sizes below 10 nm. Because of this, diffusion through them is very slow, making them impermeable to silicon and most other species. Incorporation of oxygen or nitrogen in these films makes them even better barriers. They are also chemically stable and chemically inert with most other layers. There may be some reaction with Al, forming some AlN and $TiAl_3$ at temperatures greater than 550°C. The electrical resistivity of TiN is low enough for use as both a contact material and a local interconnect. The contact resistance to Si is somewhat higher than that of PtSi, Ti, or $TiSi_2$. Therefore, TiN is usually used in a bilayer structure with TiN on top of Ti, with the Ti usually reacted to form $TiSi_2$ for a better contact. The best properties of both layers are utilized: the good adhesion and low contact resistance of the Ti or $TiSi_2$, and the good barrier properties of TiN. This is accomplished without much more processing, since the TiN can easily be deposited immediately after the Ti, just by adding nitrogen in the sputter deposition chamber. Or TiN can be formed on top of $TiSi_2$ by depositing Ti on top of the Si substrate in the contact opening, and annealing in a nitrogen ambient. The Ti reacts with the Si below it to form the silicide while simultaneously reacting with the nitrogen in the ambient to form TiN on top. This was in fact the process described in the CMOS process flow in Chapter 2 (Figures 2–3 to 2–33). The TiN can also be deposited by CVD.

Figure 11–9 shows a multilayer contact structure with $TiSi_2$/TiN contact/barrier layers. The $TiSi_2$ makes good electrical contact to the Si and provides good adhesion. The TiN serves as a barrier to interaction between the Al and both the $TiSi_2$ and Si. Table 11–3 lists commonly used contact barrier layers and compares their maximum processing temperature and modes of failure. It is clear that early contact structures have evolved from simple Al-on-Si structures to barrier-layer structures to multilayer structures. We will later see how W plugs are often used in contacts to improve step coverage and planarization.

11.2.2 Interconnects and Vias

Interconnects have similarly evolved as the technology has progressed and the requirements placed on interconnects have increased. The primary requirements on interconnects are low resistivity, good adhesion to underlying films, stability during processing and operation, and the ability to be deposited and etched.

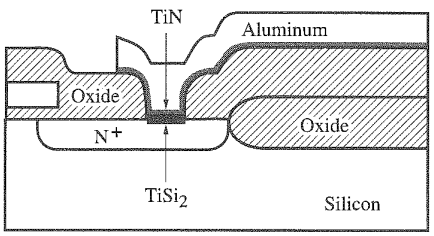


Figure 11–9 Multilayer contact structure with $TiSi_2$/TiN contact/barrier layers between the Si and Al.

Table 11–3 Barrier-layer properties [11.14, 11.15]

Structure	Failure Temperature (°C)	Failure Mechanism (Reaction products)
Al/PtSi/Si	350–400	Compound formation (Al_2Pt, Si)
Al/$TiSi_2$/Si	400	Diffusion ($Al_5Ti_7Si_{12}$, Si at 550°C)
Al/NiSi/Si	400	Compound formation (Al_3Ni, Si)
Al/$CoSi_2$/Si	400	Compound formation Al_9Co_2, Si)
Al/Ti/PtSi/Si	450	Compound formation (Al_3Ti)
Al/$Ti_{30}W_{70}$/PtSi/Si	500	Diffusion (Al_2Pt, $Al_{12}W$ at 500°C)
Al/TiN/$TiSi_2$/Si	550	Compound formation (AlN, Al_3Ti)

Aluminum, aluminum alloys, and aluminum containing multilayers have been the dominant interconnects in silicon technology to this point in time. Al has low resistivity (although not low enough to meet future NTRS specifications), adheres well to Si and SiO_2, can reduce other oxides, and can be etched and deposited using reasonable techniques. Its main shortcoming is its stability during processing and operation, which we will discuss. For early MOS technology, Al was used not only as the primary interconnect (i.e. global) but also as the gate electrode and a local interconnect. However the melting temperature of Al is low (660°C for pure Al), and it cannot withstand high processing temperatures. When self-aligned source and drain regions were needed to reduce gate-source and gate-drain capacitances, a new material was needed for the gate electrode that could withstand the 800–1000°C diffusion steps or implant anneals needed to dope the source and drain regions while the gate electrode material was present. Heavily doped polycrystalline silicon was the solution to this problem. Polysilicon can be utilized as an interconnect as well, although only as a local interconnect because of its resistivity.

The first modification to pure Al interconnects was the addition of Si at about 1 atomic percent. This prevents spiking of Al into the silicon substrate which is caused by silicon's relatively high solubility and diffusivity in Al. Even with the use of barrier layers such as TiN, Si is often still used in Al interconnects. This is because the presence of the Si reduces the driving force for interdiffusion and hence decreases the chance of barrier failure. Si also suppresses the reaction between Al and TiN.

Almost all the rest of the modifications made to the Al interconnect structure have been as a result of aluminum's low melting point and ease of deformation. Two phenomena that occur with aluminum which have caused significant problems and have required modifications in the Al interconnect are hillock growth and electromigration.

Hillock growth occurs when the Al thin film is subjected to high compressive stresses. Because of the polycrystalline structure of Al, coupled with its low melting point and resulting susceptibility to plastic deformation, the Al will deform under such stresses. High stresses can easily be generated due to Al's relatively high coefficient of thermal expansion ($23 \times 10^{-6}\ °C^{-1}$, while Si is $2.6 \times 10^{-6}\ °C^{-1}$ as shown in Table 11–4). If Al is deposited on a Si substrate (or even on another film, such as SiO_2, on a Si substrate) at low

Table 11–4 Mechanical properties of interconnect materials

Material	Thermal expansion coefficient ($°C^{-1}$)	Elastic modulus, $Y/(1-\nu)$ (MPa)	Hardness ($kg\ mm^{-2}$)	Melting point (°C)
Al (111)	23.1×10^{-6}	1.143×10^5	19–22	660
Ti	8.41×10^{-6}	1.699×10^5	81–143	1660
$TiAl_3$	12.3×10^{-6}	—	660—750	1340
Si (100)	2.6×10^{-6}	1.805×10^5	—	1417
Si (111)	2.6×10^{-6}	2.290×10^5	—	1417
SiO_2	0.55×10^{-6}	0.83×10^5	—	~1700

temperatures and the structure is heated, high compressive stresses in the Al can be generated. This is because the Al, which is tightly constrained to the wafer, will want to expand much more than the Si. To relieve the stress, portions of the Al are squeezed up, forming small hills or "hillocks." The movement of the Al occurs primarily along grain boundaries in the Al since diffusion is usually much faster along grain boundaries than through the interior of the grains. Whole grains of Al are often pushed or grow upward, forming hillocks about the size of the grains, 0.5 to 3 μm tall and wide.

Figure 11–10 shows schematically the process of hillock formation. Hillocks can result in electrical shorts between interconnect levels, as well as cause the surface topography to be rough, making lithography and etching more difficult. While the overlying films can help suppress hillock formation by acting as a mechanical barrier, this does not completely eliminate the problem. In fact, so much stress can build up that an overlying SiO_2 layer can crack. When this happens, the Al can protrude from the crack in long whiskers many microns long, increasing the chances of a short.

The addition of elements that have limited solubility in Al, such as Cu, has been found to suppress hillock formation. The excess atoms segregate and precipitate preferentially at the Al grain boundaries. This reduces the Al diffusion along the grain boundaries and suppresses hillock formation.

In a similar manner, voids can form when the Al is subjected to tensile stresses. Upon cooling, the Al will want to shrink more than the Si below it. (Unfortunately any hillocks that have formed during the previous heating do not significantly reverse themselves and disappear during the cooling.) Since the Si is more mechanically rigid than the Al, the stress is relieved by vacancy movement and agglomeration in the Al, with voids forming. This can greatly increase the interconnect electrical resistance and even cause open circuits. The addition of Cu to the Al can suppress void formation in Al as well as suppressing hillock formation.

Perhaps the biggest problem encountered with Al interconnects is that of electromigration. One of the requirements of an interconnect is that it be stable not only during processing, but also during circuit operation. Electromigration is a phenomenon that occurs during circuit operation and has very adverse effects. When an electrical current flows through aluminum (on the order of 0.1–0.5 MA cm^{-2}, which is commonly achieved in integrated circuits), the electrons can actually transfer enough momentum

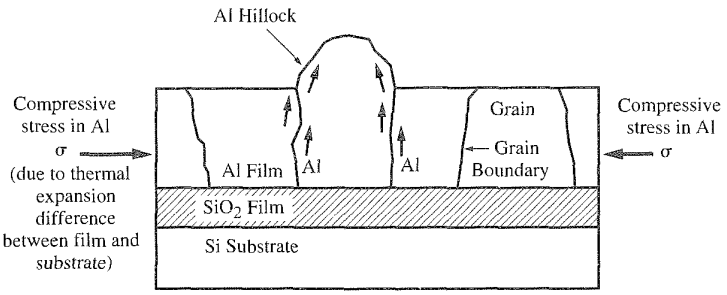


Figure 11–10 Schematic illustration of hillock formation due to compressive stress in an Al film. Al diffusion along grain boundaries is indicated.

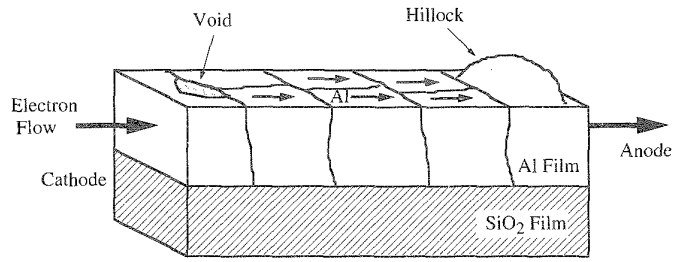


Figure 11–11 Schematic illustration of electromigration with resultant hillock and void formation. Diffusion of Al through grain boundaries is indicated.

to the Al atoms to cause them to diffuse. This diffusion is faster along grain boundaries, and it can cause a buildup of Al in some regions, resulting in hillocks, and can cause a depletion in other regions, resulting in voids. As a result shorts and opens in the interconnects can occur during the normal operation of the circuit. Figure 11–11 shows this process schematically, and Figure 11–12 shows a SEM of an actual circuit where this has occurred. Because most of the Al atoms and vacancies diffuse along the grain boundaries, electromigration is very dependent on the grain structure, including the size and crystallographic orientation of the grains. It also depends on over- and underlying layers, and on the processing history of the interconnects.

Substantial effort has been made to study and remedy this phenomenon. One of the early remedies for electromigration was to add Cu to the Al. As is the case in suppressing hillocks, adding Cu is believed to suppress grain boundary diffusion and hence slow down electromigration. Adding too much Cu, however, causes problems with etching of the interconnect as well as corrosion, and 4 weight percent Cu is the most that is usually added.

Thus, the first modifications to Al interconnects involved switching from pure Al to Al with about 1–2 weight percent Si and 0.5–4 weight percent Cu. Adding these elements to the Al increases the sheet resistivity of the interconnect by as much as 35%,

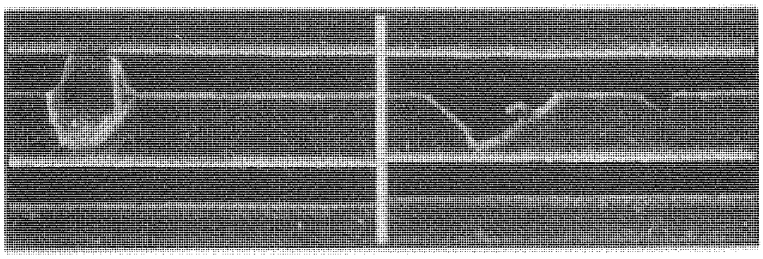


Figure 11–12 SEM top view of hillock and voids that have formed due to electromigration in an Al line [11.16]. Reprinted by permission of Materials Research Society.

but this was a necessary trade-off in order to prevent the failure of circuits due to electromigration and hillock and void formation. Other solutions, similar to those for contacts involving multilayer structures, were later developed and will be discussed later.

The next change to interconnects involved the local interconnects. Polysilicon has been used as the gate electrode since the mid 1970s. But as circuits got faster, the sheet resistivity of heavily doped polysilicon became a concern (Figure 11–4). In addition, the doped silicon in the source and drain regions has a fairly high resistivity (50–100 $\mu\Omega$-cm) which can add appreciable parasitic resistance to MOS transistors. Silicide materials can be used to reduce all these problems.

Silicides have been utilized as, or with local interconnects in three primary ways, as illustrated in Figure 11–13. On top of the polysilicon layer they form a "polycide" structure; on top of the source and drain regions they strap the doped silicon; and finally, they can be used as an independent local interconnect. All of these uses were illustrated in the CMOS process in Chapter 2 (Figure 2–33). In the polycide structure, the silicide decreases the sheet resistance of the polysilicon gate and local interconnect, while retaining the highly reliable polysilicon/SiO_2 interface. On top of the source and drain regions, the silicide reduces the resistance of the diffused layers. The silicide layer also provides a good contact to the silicon in the gate, source, and drain regions, while increasing the contact area. Silicides are also sometimes used as local interconnects by themselves, providing relatively low resistive paths below the lowest metal layers. Silicides are used because they offer low electrical resistivity (10–50 $\mu\Omega$-cm), they are stable at high temperatures, they provide good process compatibility with silicon, they are easy to plasma etch, they provide good contacts to other materials, and they do not exhibit much electromigration. Some of the common silicides used in silicon technology are listed in Table 11–5, along with some of their properties.

As discussed in Chapter 9, the formation of silicides can be accomplished in two ways: by direct deposition of the silicide and by the reaction between the metal and Si on the wafer to form the silicide. The direct deposition method can use sputtering from a composite target, cosputtering from two targets of the metal and Si, co-evaporation of the metal and Si, or Chemical Vapor Deposition (CVD). CVD is currently only done for WSi_2. The reaction method can be accomplished by two different techniques. The first technique is by depositing the metal and amorphous Si, both usually by sputtering, and

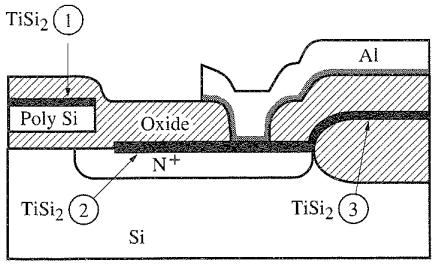


Figure 11–13 Uses of silicides in back-end structures: 1, on top of a polysilicon gate; 2, on top of source and drain regions; and 3, as a local interconnect.

Table 11–5 Properties of silicides (After [11.17])

Silicide	Thin film resistivity (μΩ-cm)	Sintering temp (°C)	Stable on Si up to (°C)	Reaction with Al at (°C)	nm of Si consumed per nm of metal	nm of resulting silicide per nm of metal	Barrier height to n-Si (eV)
PtSi	28–35	250–400	~750	250	1.12	1.97	0.84
$TiSi_2$ (C54)	13–16	700–900	~900	450	2.27	2.51	0.58
$TiSi_2$ (C49)	60–70	500–700			2.27	2.51	
WSi_2	30–70	1000	~1000	500	2.53	2.58	0.67
Co_2Si	~70	300–500			0.91	1.47	
CoSi	100–150	400–600			1.82	2.02	
$CoSi_2$	14–20	600–800	~950	400	3.64	3.52	0.65
NiSi	14–20	400–600	~650		1.83	2.34	
$NiSi_2$	40–50	600–800			3.65	3.63	0.66
$MoSi_2$	40–100	800–1000	~1000	500	2.56	2.59	0.64
$TaSi_2$	35–55	800–1000	~1000	500	2.21	2.41	0.59

patterning and selectively etching the amorphous Si. (The Si may be deposited and patterned either before or after the metal deposition.) The wafer is then annealed so that the metal and amorphous Si react to form the silicide, and the unreacted metal is selectively etched off. The second reaction method is similar but no amorphous Si is deposited. Instead, the metal is deposited on the exposed gate and/or source drain regions, all of which are silicon. The wafer is annealed and the silicide forming reaction occurs wherever the silicon and metal are in contact, that is, on the gate and source/drain regions. The unreacted metal is then selectively etched away. The first of these reaction techniques produces local interconnects independent of either gate or source/drain regions, while the second reaction method produces self-aligned silicide structures on the gate and/or source drains. This second technique is often called a salicide, for "self-aligned silicide." Figure 11–14 illustrates the salicide process.

A high-temperature anneal is required for either the direct deposition or reaction methods of silicide formation. In the reaction method, a high-temperature step is needed to react the metal with the Si to form the silicide and hence to achieve low re-

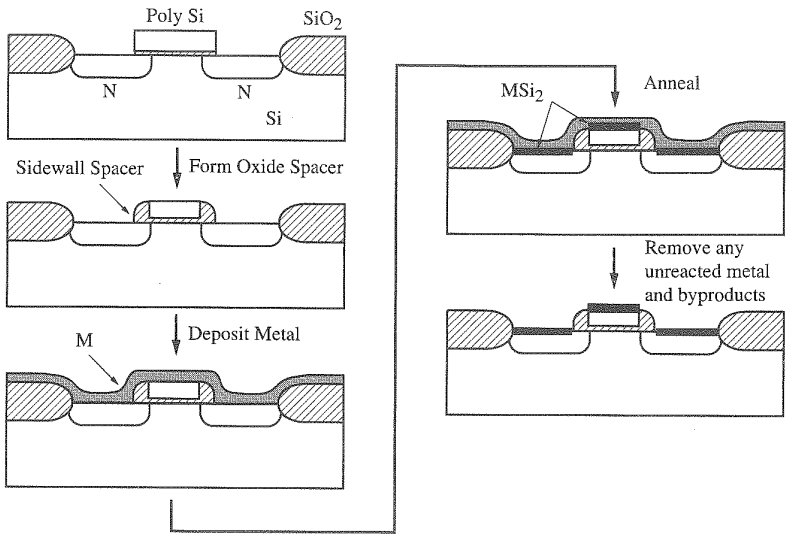


Figure 11–14 Salicide (self-aligned silicide) process. M represents a metal, such as Ti.

sistivity films. A high-temperature step is also required in the direct deposit methods. Depending on the exact deposition technique, the grain structure of the as-deposited film is either very fine, microcrystalline, amorphous, or even an unreacted multilayer structure. In any of these cases, annealing is needed to make the grains grow in size in order to achieve low-resistivity films. This is because in small-grain films there are more grain boundaries which scatter the carriers, reducing the electrical mobility. The final, or minimum, resistivity for various silicides is given in Table 11–5.

There have been several silicides used in silicon technology, including WSi_2, $TaSi_2$, $MoSi_2$, $TiSi_2$, and $CoSi_2$. Tungsten silicide has been used because a CVD deposition process is available for it, providing good step coverage and good process control. In addition, it can be oxidized rather easily, as long as a source of silicon is maintained below it. Therefore, a good thermal oxide can be formed over it rather than a deposited oxide. However, high stresses in WSi_2 often cause stability and adhesion problems. WSi_2, as well as $TaSi_2$ and $MoSi_2$, are not as easy to form through reaction with Si as $TiSi_2$ and $CoSi_2$ are, and therefore are not as attractive for salicide applications.

$TiSi_2$ is very popular due to its low resistivity, and its ability to reduce native oxides and to adhere well to most other materials. However, $TiSi_2$ has a high-resistivity phase ("C49") that forms at moderate temperatures. This must be transformed to a low resistivity phase by annealing, and this transformation becomes more difficult for narrow and thin lines. Also, the thin film has a tendency to agglomerate, or roughen up, at higher annealing temperatures.

$CoSi_2$ behaves better in regard to these particular problems. In addition, it generally shows less lateral encroachment during its formation reaction. Lateral encroachment means the silicide forms beyond the source, drain, and gate regions over the oxide (or

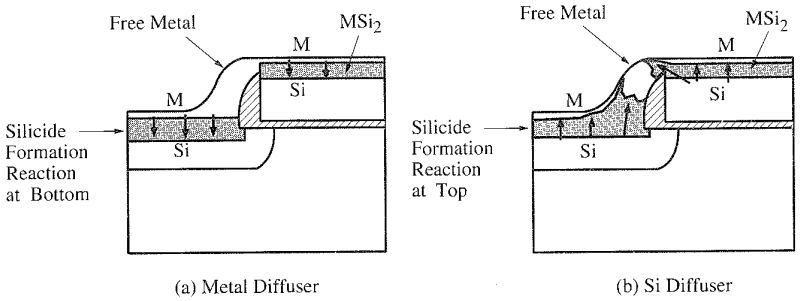


Figure 11–15 Schematic illustration of silicide formation for a silicide that forms by (a) diffusion of metal species and by (b) diffusion of silicon. In the latter case, lateral encroachment of the silicide over the oxide spacer can occur.

nitride) spacer, which can cause an electrical short between those areas. Encroachment is generally less for $CoSi_2$ because of its lower reaction temperatures, and also because the metal is believed to be the dominant diffusing species during silicide formation, unlike $TiSi_2$, where Si is the diffuser. This is shown schematically in Figure 11–15. However, the resistivity of $CoSi_2$ is slightly higher than $TiSi_2$, and it does not have as good adhesion and oxide reducing properties as $TiSi_2$. $CoSi_2$ is more sensitive to surface contaminants and oxides, which can lead to non-uniform silicidation and junction leakage.

TiN, as a byproduct of the $TiSi_2$ salicide process, can also be utilized as a local interconnect as was illustrated in the CMOS process flow in Chapter 2 (Figures 2–31 to 2–33). The reaction between Ti and Si in the salicide process is often done in a nitrogen ambient in a two-step anneal process, as will be discussed in the next section on processing techniques. Portions of the TiN that is formed can be masked and left on the circuit to act as a local interconnect. An advantage of using TiN as a local interconnect rather than polysilicon or a silicide is that TiN is a diffusion barrier for most dopants. TiN is commonly utilized as a local interconnect in SRAM cells, where small cell size is critical, connecting load drains with the input gate of the inverter.

Another material that has been used or considered for a local interconnect is tungsten. Selective deposition of tungsten over Si areas using CVD (see Chapter 9) could be used to form W/polysilicon gate and W/silicon source and drain structures. However, at temperatures above 600°C the W and silicon will react to form a silicide, eliminating the advantage of having a lower resistivity material strapping the gate, source, and drain regions. First-level dielectric processing is commonly done above this temperature, which would be a problem to utilizing W in this manner.

As circuits became larger and more complex, additional interconnect levels have been required (Table 11–1). Adding an additional level of global interconnect seems easy enough. The process could be as simple as opening a hole in the top dielectric level wherever a contact to the lower level is needed, depositing Al (with some Cu in it to inhibit electromigration), forming both the contact and the next level interconnect, fol-

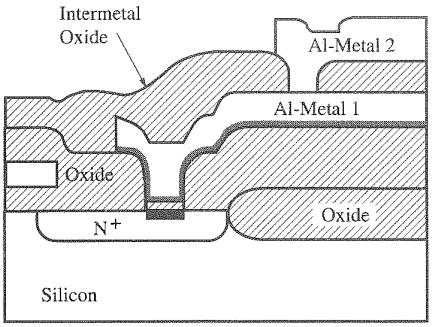


Figure 11–16 Illustration of early two-level metal structure.

lowed by patterning of the aluminum. Then the next level of intermetal dielectric material would be deposited. This structure is illustrated in Figure 11–16. This, in fact, is how the early multilevel interconnect structures were made. Looking at the contact, or via in this case, Al is now making contact to Al, so spiking into Si is not a problem. Good electrical contact between two layers of Al can be achieved if the interface is clean.

However, the topography gets less planar as more levels are added, as Figure 11–16 illustrates. A more planar topography is desired for two main reasons—lithography limitations and step coverage. The lithography limitations arise partly from the limited depth of focus of submicron lithography instruments. The maximum depth of focus is about $\pm$ 0.35 μm for state-of-the-art optical tools (see Chapter 5.). In addition, nonplanar topography results in nonuniform resist thickness, which can result in nonuniform exposure and development. Another lithography issue is that spurious UV light reflections can occur from the sides of metal as it goes over steep steps, again causing lithography problems. There are lithographic techniques that address these problems, such as using multilayer resist structures as discussed in Chapter 5, but these are not always effective or practical solutions.

Nonplanar topography also leads to problems with step coverage and filling of spaces with deposited films. We saw in Chapter 9 how step coverage and filling can be problems, especially for PVD deposited metal (see Figure 9–1, for example). Such problems lead to higher electrical resistance in the line, as well as potential reliability problems. To make matters worse, step coverage problems compound one another. A large step height on one level can lead not only to thinning of the next level on sidewalls, but also cusping and overhangs. This in turn leads to even worse coverage problems on the next level.

Reducing the step heights and achieving more planar topology through processing techniques is called planarization. The Degree Of Planarization (DOP) is defined as

$$\text{DOP} = 1 - \frac{x^f_{\text{step}}}{x^i_{\text{step}}} \tag{11.13}$$

where x^i_{step} is the initial step height in the topography and x^f_{step} is the final step height after some sort of planarization is done. These step heights can refer to hole depths as

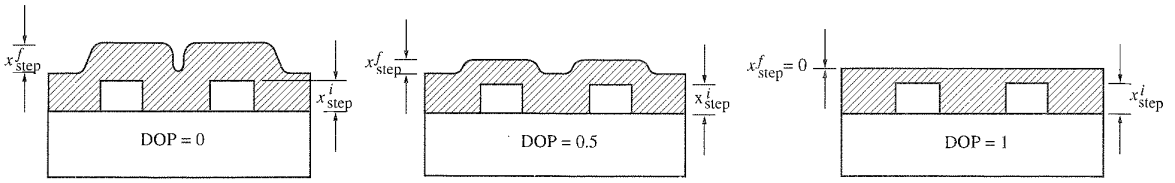


Figure 11–17 Demonstration of Degree Of Planarization (DOP).

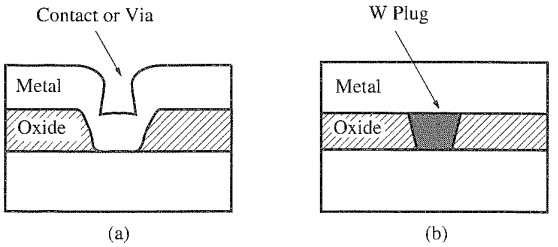


Figure 11–18 (a). Using Al as both the contact and the next level interconnect layer results in nonplanar topography. (b) Using a tungsten plug as the contact results in planar topography.

well. A DOP of 0 means that no planarization is achieved, while a DOP of 1 corresponds to complete planarization—the final step height is zero. Figure 11–17 shows different degrees of planarization. Several methods have been developed to improve the planarity of multilevel interconnect structures. Some involve the metals, which will now be discussed, while others involve the dielectric materials, which will be discussed later.

One way to help achieve planar metal layers is by using a W contact or via fill, also called a W plug. Figure 11–18(a) shows the topography above a contact or via which uses Al as both the contact material and the next level of interconnect. Not only might the step coverage of the Al be a problem, but the resulting topography over the contact can be very non-planar. Using a via fill or plug structure results in the structure in Figure 11–18(b). This can be accomplished in two ways. One method involves a selective W deposition into contact or via holes cut in the dielectric layer, which fills the contacts or vias without depositing any W over the top of the dielectric. The other, more common, method is using a blanket W deposition, which fills the contact or via as well as covering the dielectric. Then an "etchback" is done to remove the W over the top surface of the dielectric. This is illustrated in Figure 11–19. This process of filling holes or trenches in a dielectric level with the metal (usually followed by an etchback), rather than depositing, patterning, and etching the metal on top of a dielectric layer, is called the "damascene" method. Its name comes from the ancient practice of inlaying metal in wood or ceramics for decoration. It can be used for vias or interconnect layers, or both.

Using a W plug produces a much smoother topography over the contact or via, as shown in Figure 11–18(b), resulting in a near-planar contact or via level. The DOP approaches 1 here, where the initial step height corresponds to the height of the contact or via hole. Figure 11–2 illustrated such W contact plugs with an Al metal line on top. With a planar contact or via level, the problem with step coverage of sputtered Al is greatly reduced since there are few, if any, steps for the Al to cover. W is used for the

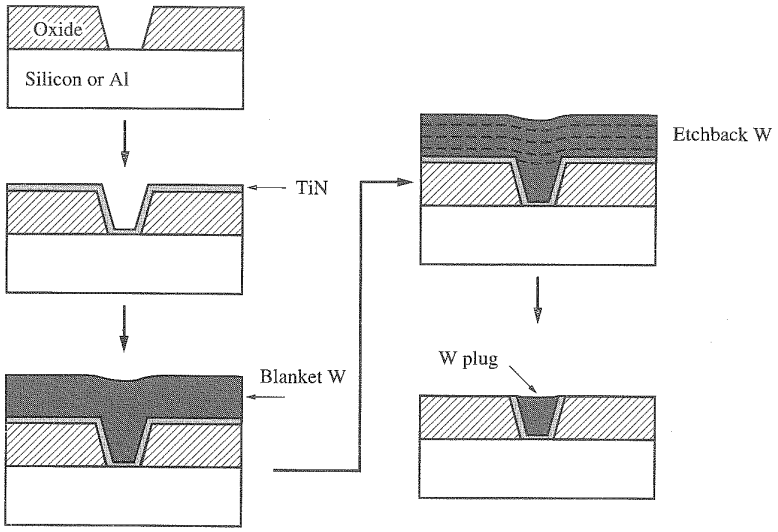


Figure 11–19 Schematic illustration of process to fabricate a W plug for a contact or via. TiN (or a Ti/TiN bilayer) is utilized as an adhesion and barrier layer. An etchback step is done to planarize the W and leave it only in the contact or via hole.

plug rather than Al because the deposition of W using the CVD technique gives much more conformal filling of the contact or via hole. Because of problems with CVD deposition of Al-Cu alloys, as discussed in Chapter 9, CVD Al processes are generally not used in manufacturing.

One planarizing technique under development that does utilize Al as the contact or via metal is called Al reflow. By heating the Al to 450–550°C either during the deposition or immediately afterward, the Al in a contact or via hole will flow and fill the hole, as well as help planarize itself. This is similar to reflow of dielectric materials which are heavily doped with B or P. With this technique, one could deposit the via (or contact) and interconnect levels in one step to obtain good filling of the contact and good step coverage and planarization of the interconnect level.

Just as contacts benefited from the use of multilayer structures, so have interconnects. This is due to a variety of reasons, most of them related to issues we have already discussed. First, electromigration has continued to be a problem that has hampered the reliability of silicon integrated circuits. While the addition of Cu to the Al has helped, it has not completely eliminated the problem. Other methods have been used to inhibit electromigration, or in some cases, negate its effect on the integrity of the interconnect lines. One method uses shunt lines. Another interconnect material is layered below, above, or within the Al layer, as shown in Figure 11–20. If electromigration does cause an opening in the Al line, the current can be shunted through the other material, maintaining the electrical integrity of that interconnect. The material does not have to have as low a resistivity as Al, since current will only flow through it for a short distance. But

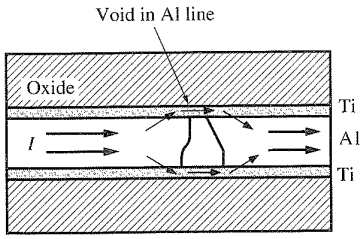


Figure 11–20 Illustration of shunt layers to minimize electromigration problems. Thin layers of metals, such as Ti, are used above and below the Al interconnect line. If a void should form, the current can be shunted around it and through these layers.

it should have good electromigration resistance, and good adhesion or barrier properties. Layers of Ti, Ti-W, TiN, or $TiSi_2$ can be used for this purpose.

These interlayers in interconnects help in electromigration protection for two other reasons. If they have high melting points and strong mechanical properties (see Table 11–4), they can be more rigid than the Al and act as physical or mechanical barriers to hillock and void formation. If Ti is used, it will actually react with Al to form $TiAl_3$, which is even harder mechanically than Ti, and will reduce the effects of electromigration. However, if too much $TiAl_3$ forms, it can increase the resistivity of the interconnect. If TiN is used, it is best to have excess Ti in it which will react with the Al to form $TiAl_3$. In addition, the electromigration of Al is very dependent on the grain texture of Al. The closer all the grains are to a single, usually (111), orientation in the vertical direction, the less electromigration occurs. The grain texture of the Al is very dependent on what underlayer it is deposited on. Al deposited on Ti exhibits much better grain texture (more grains closer to a single orientation) and electromigration behavior than on other films, such as Si, SiO_2, or TiN. If TiN is used between the Ti and Al, as a reaction barrier for example, the Al still has good grain texture. This is because the TiN takes on a pseudogranular structure, passing the Ti structure to the Al interface. Therefore, by using either Ti or Ti/TiN layers underneath the Al layers, electromigration is decreased due to grain texture considerations.

Interconnect and via structures made up of multilayers have been developed for barrier and adhesion reasons, just as with the contact structures. Ti can help with adhesion and electrical contact. TiN can serve not only as a good barrier, but as an adhesion layer as well, particularly in conjunction with W plugs. A TiN or Ti/TiN layer provides for good contact between Al and W and prevents any reaction between them. In addition, when CVD W is deposited in a via, a barrier layer must be used between the W and the underlying Al or else the fluorine in the WF_6 used in the CVD W deposition process will react with the exposed Al. (Similarly, in a contact a barrier is required to prevent the WF_6 from reacting with Ti or $TiSi_2$ layers or with Si.) Since the adhesion of W to dielectric layers is very poor, an adhesion layer is required and TiN (or Ti/TiN) is commonly used for this. A TiN layer can also be used in the interconnect stack to serve as a barrier between the Ti and the Al.

Another reason for using a multilayer interconnect structure is for lithographic purposes. The top Al surface is very reflective, and the reflection of light can adversely and unevenly affect the lithographic process. Therefore, an antireflective material is often deposited on top of the Al to minimize this. TiN again is usually used.

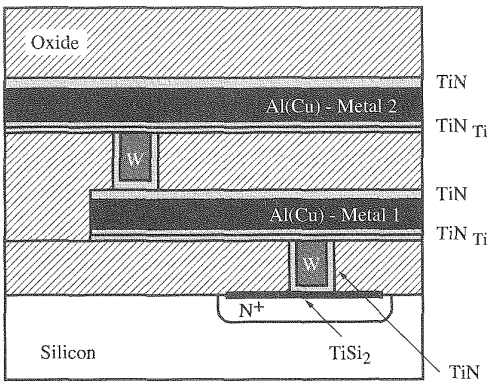


Figure 11–21 A typical metalization scheme used in current IC technology, showing the multilayer structures that have evolved.

For all these reasons, multilayer interconnect structures have been developed. One such structure is illustrated in Figure 11–21 and serves to summarize the issues we have discussed. A W plug is used in both the contact and the vias, resulting in good planarity. At the bottom of the contact, $TiSi_2$ is used to make good electrical contact to the Si, and also serves to strap the sources and drains, providing for lower sheet resistance along those paths. In both the contacts and the vias, TiN surrounds the W, promoting adhesion to the oxide. It also acts as a barrier to prevent attack of the underlying material by WF_6 during W deposition. The global interconnect structure, or "stack," in this example is Ti/TiN/Al(Cu)/TiN. Cu in the Al suppresses electromigration and hillock formation. Ti improves the grain texture of the Al for better electromigration protection and promotes adhesion. TiN is a barrier between the Al and Ti to limit $TiAl_3$ formation. The Ti/TiN and TiN layers on the bottom and top of each interconnect provide an electrical shunt in case of void formation in the Al. They also suppress hillock and void formation by providing mechanical barriers to plastic deformation. TiN on the top also acts as an antireflective layer for lithography. Many variations of this structure are used by different manufacturers, especially in the exact makeup and order of materials in the interconnect stack (the Ti or TiN is often omitted from the bottom of the stack, for example). But the basic idea of multilayer planar structures is very common.

11.2.3 Dielectrics

Dielectric layers are used to separate and electrically isolate conducting layers from each other. Two types of dielectric layers associated with metallization schemes were shown in Figure 11–1. The first-level dielectric is a deposited layer that separates the global (usually Al) interconnects from the diffused substrate areas, polysilicon, and local interconnects. The intermetal dielectrics (IMDs) are deposited dielectrics that

separate the global interconnects. (The other common dielectric layers are the gate oxide and the field oxide, both of which are thermally grown and have been discussed in previous chapters.) Because the first-level dielectric is deposited before any Al is present, while the intermetal dielectrics are deposited after Al is present, they have different processing requirements and restrictions.

There are a number of general requirements for dielectric layers [11.18]. Good electrical isolation is paramount and includes properties such as a low dielectric constant so that interconnect capacitances are low; a high breakdown field strength (>5 MV/cm); and low leakage (bulk and surface resistivity should exceed 10^{15} Ω cm). In addition, dielectric layers should provide good adhesion to silicon, metals, and silicides; low intrinsic stress (50–100 MPa or less); thermal stability (up to 850–950°C for first-level dielectrics and up to 500°C for intermetal dielectrics); contain no moisture or metallic impurities (which can diffuse into the substrate and cause leakage current); low-defect density (pinholes and particulates); high-structural density (low porosity); contain no residual constituents that can outgas; and they should be permeable to hydrogen (so that passivation of the Si/SiO_2 interface states can take place). Table 11–6 lists several dielectric materials used or being considered for use in back-end structures.

Table 11–6 Materials used or considered for interlevel dielectrics (After [11.19])

Material class	Material	Dielectric constant	Deposition technique
Inorganic	SiO_2 (including PSG and BPSG)	3.9–5.0	CVD Thermal oxidation Bias-sputtering High-density plasma
	Spin-on-glass (SiO_2) (including PSG and BPSG)	3.9–5.0	SOD (spin-on-dielectric)
	Modified SiO_2 (e.g., fluorinated SiO_2 or hydrogen silsesquioxane—HSQ)	2.8–3.8	CVD/SOD
	BN (Si)	>2.9	CVD
	Si_3N_4 (only used in multilayer structure)	5.8–6.1	CVD
Organic	Polyimides	2.9–3.9	SOD/CVD
	Fluorinated polyimides	2.3–2.8	SOD/CVD
	Fluoro-polymers	1.8–2.2	SOD/CVD
	F-doped amorphous C	2.0–2.5	CVD
Inorganic/Organic Hybrids	Si-O-C hybrid polymers based on organosilsesquioxanes (e.g., MSQ)	2.0–3.8	SOD
Aerogels (Microporous)	Porous SiO_2 (with tiny free space regions)	1.2–1.8	SOD
Air bridge		1.0–1.2	

CVD silicon dioxide is the dominant dielectric material used in back-end structures. In the 1960s, it was found that if the CVD SiO_2 was doped with phosphorus, forming phosphosilicate glass, better properties for MOS technology were obtained (see Chapter 4, Section 4.2.3). Phosphosilicate glass (PSG) is made up of P_2O_5 and SiO_2. The phosphorus in the films allows the film to getter, or tie up, sodium ions, resulting in more stable MOS characteristics. It also reduces the built-in tensile stress often present in deposited SiO_2 films. In addition, the presence of phosphorus allows one to "reflow" the layer. By annealing the PSG in steam at a temperature from 950–1100°C, the dielectric layer flows in order to reduce surface energy, and any sharp corners are smoothed out. Undoped oxide does not flow nearly as much because of its higher viscosity. Reflowing also helps in filling spaces between closely spaced lines. Reflowing the dielectric results in better step coverage of subsequent layers and was one of the first planarization techniques. The amount of phosphorus incorporated in the films ranges from 4 to 8 wt. percent, with more phosphorus leading to more reflow. However, more phosphorus also enhances the film's absorption of water. Atmospheric moisture absorbed in the film can react with the phosphorus to form phosphoric acid, which can lead to corrosion of aluminum layers on top of and below the film. Absorbed water can also lower the dielectric constant of the film, and can outgas during heating, causing metal deposition problems. A sandwich structure is often used with layers of undoped SiO_2 below and above the PSG to help protect against these problems.

PSG has commonly been used as the first-level dielectric. Since Al is not present at this point in the processing, high-temperature densification and reflow can be used. If silicides are present, temperatures below 950°C are required for first dielectric-level processing since most silicides start to degrade at about 900°C. If boron is added to PSG, to form BPSG (borophosphosilicate glass), the reflow temperature is reduced to about 800°C. Reflowed BPSG is illustrated in Figure 11–22. BPSG, with 2–6 wt. percent boron and phosphorus, also traps sodium and reduces the stress in the film.

When two metal-level interconnect structures were first introduced, PSG or BPSG were initially used for the intermetal dielectric. But since the structures with Al layers could not be heated above 450°C, a reflow step could not be done, and these films had many problems. Because of the relatively high sticking coefficients (around 0.3 to 0.4) characteristic of silicon dioxide films formed by the oxidation of silane, conformal depositions were usually not achieved. This can lead to problems with step coverage and filling of spaces. These problems get worse as metal spacings shrink with each new

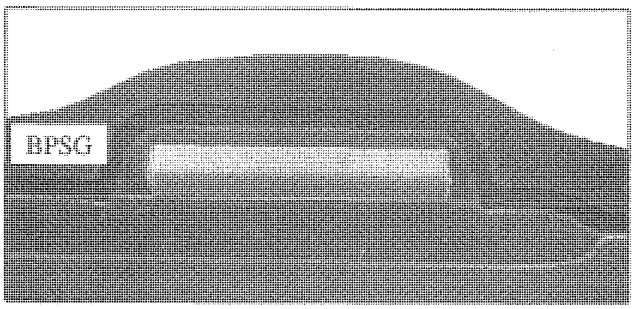


Figure 11–22 SEM image of BPSG oxide layer after 800°C reflow step, showing smooth topography over steps. Photo courtesy of VLSI Technology, Inc.

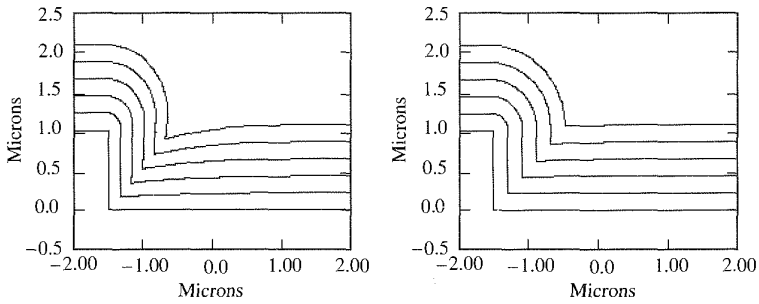


Figure 11–23 SPEEDIE simulations of silicon dioxide depositions over a step for silane deposition ($S_c = 0.4$) on left and TEOS deposition ($S_c = 0.1$) on right, showing less cusping in the latter case.

generation, especially if metal thicknesses do not shrink as fast. Without a reflow step to smooth out the films and fix the coverage and fill problems, failures in the interconnects can result. In addition, low-temperature SiO_2 films that are not annealed at higher temperatures are not as dense and not as good quality. Other deposition and processing techniques, as well as different materials, were needed for intermetal dielectrics.

Plasma-enhanced CVD techniques provide better oxide films for intermetal dielectrics. The energy from the plasma enhances the deposition reactions and provides for denser films and somewhat better coverage and filling at lower temperatures. Silicon dioxide films deposited by decomposing TEOS (tetraethooxysilane—$Si(OC_2H_5)_4$), rather than by reacting oxygen and silane, have also been developed. The reactant species in these depositions tend to have a low sticking coefficients, on the order of 0.1. This results in more conformal films than those produced from silane for both LPCVD and PECVD systems. Figure 11–23 shows SPEEDIE simulations of LPCVD silicon dioxide depositions over a step for silane deposition ($S_c = 0.4$) and TEOS deposition ($S_c = 0.1$). Less cusping occurs in the latter case, resulting in better conformality as well as filling ability. Even better conformality and filling ability at low temperatures is achieved by reacting the TEOS with ozone. Low-temperature deposition processes, using ozone or PECVD, allow the use of TEOS oxides as intermetal dielectrics. The non-plasma, $TEOS/O_3$ films are more porous than silane based or PECVD TEOS oxides and absorb more moisture. Often a thin layer of PECVD TEOS or silane oxide is used below and above $TEOS/O_3$ films. Bias-sputtering or a sequential deposition/etch/deposition process may also be used to help with filling. TEOS films are sometimes used in the first-level dielectric level as well. The TEOS films can be doped with P or B to help with Na gettering and reflow properties.

Planarization is important in multilevel back-end structures and can be divided into local planarization and global planarization. Local planarization occurs over submicron or micron (device) dimensions, while global planarization occurs over circuit or chip dimensions. Figure 11–24 illustrates this. Global planarization is usually desired, but local planarization is easier to achieve. An example of local planarization would be the reflow process, where the areas between closely spaced lines are filled in with the oxide during the reflow anneal, but larger spaces or areas outside are not.

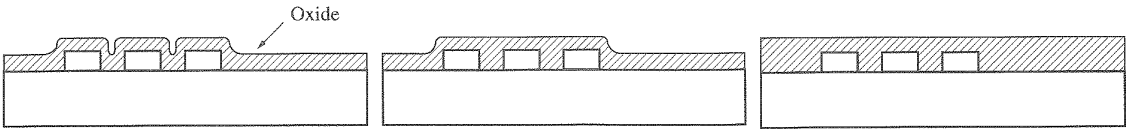


Figure 11–24 Schematic illustration of different types of planarization.

Example In Figure 11–24(b), approximately what degree of planarization has been achieved locally and what degree of planarization has been achieved globally?

Answer By Eq. (11.13), the degree of planarization is

$$\text{DOP} = 1 - \frac{x_{\text{step}}^{f}}{x_{\text{step}}^{i}}$$

The initial step height is the height of the metal line and is some finite number. In regard to local planarization, the final step height corresponds to the step height in the region that was locally planarized, that is, the regions between or within the metal lines. The step height here is close to zero, so that the degree of planarization locally is close to 1. Globally, the final step height would correspond to the maximum step height on the surface. Here it would be close to the initial step height. Therefore, the degree of planarization globally is close to zero.

Another planarization method involves etchback. In this technique the dielectric is deposited, sometimes followed by the deposition of a sacrificial layer. The overall structure is then etched back to planarize it. The earliest such process used photoresist as a sacrificial etchback layer on top of oxide. The photoresist was deposited, or spun, over the oxide layer, filling up the spaces and ending up with a nearly flat top surface. After a hard-bake, an etch is done which etches both the oxide and photoresist, ideally at the same rate. When the bottom of the photoresist is reached, a nearly flat oxide is left behind. This is illustrated in Figure 11–25, and results in local planarization, but only limited global planarization. This can also be done without a sacrificial layer if the dielectric layer is deposited thick enough. This is because the thicker a material is, the smoother the local topography tends to become. Plasma etching, argon sputtering, or chemical-mechanical polishing can be used to etch back the dielectric layer, leaving a more planar topography.

An additional planarizing method is to use a Spin-On-Glass (SOG). This is similar to the photoresist etchback technique, but here the spun-on material is left behind and serves as an intermetal dielectric. SOG materials are initially liquids, containing organic siloxanes or inorganic silicates in an alcohol-based solvent. The liquid SOG is spun onto

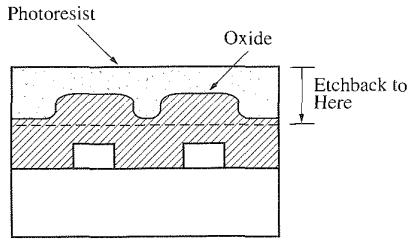


Figure 11–25 Illustration of photoresist etchback process. Photoresist is deposited over rough topography, then the structure is etched back, leaving a smooth top surface on the oxide.

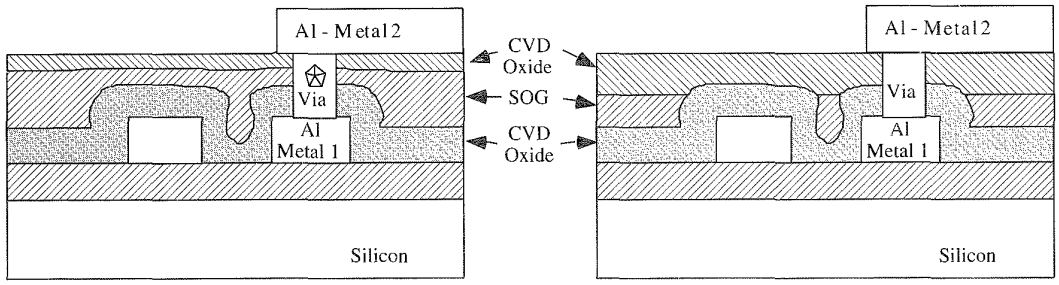


Figure 11–26 Back-end structure (left) without etchback step, showing poisoned via (star-shaped structure) where the via metal is in direct contact with Spin-On Glass (SOG) layer. The back-end structure on the right with an etchback step, does not have poisoned vias since the via metal is not in direct contact with the SOG layer.

the wafer like photoresist, filling the spaces between features. (It can also fill voids between lines as long as the voids are not completely closed off at the top.) Then the wafer is baked and cured, driving off the solvents and polymerizing the silicon-containing groups, resulting in the formation of Si-O-Si bonds. SOG is often sandwiched between layers of CVD SiO_2 in an intermetal dielectric structure, as shown in Figure 11–26. The SOG process is often combined with an etchback. After deposition and curing, the SOG is etchbacked so that no SOG remains over any metal region that will have a via made to it. This is so that when via holes are cut through the intermetal dielectric, they do not pass through any SOG, as also illustrated in Figure 11–26. If they did, residual gases and moisture from the SOG could contaminate the contacts or vias and cause large increases in contact or via resistance, so-called "poisoned vias." SOG films made from organic-based solutions, containing carbon, are more susceptible to this than those made from inorganic solutions. With SOG, local planarization and only very limited global planarization are achieved.

As alternatives to spin-on SiO_2 films, other Spin-On Dielectric materials (SODs) are being developed. These are either inorganic, organic polymer materials, or hybrid organic/inorganic materials (Si-O-C polymers). Table 11–6 lists some common materials. While providing good filling and local and some global planarization like SOG, they have the advantage of having lower dielectric constants than SiO_2. This means the intermetal capacitance is lower, and the *RC* delay of the interconnects is improved. The main problems with these low-*K* materials are thermal stability (i.e., they often decompose at

moderate temperatures) and moisture absorption, both of which tend to degrade their electrical properties and adhesion. Sandwich layers with CVD SiO_2 are often used to mitigate these problems.

A more recently developed oxide deposition technique is called High-Density Plasma (HDP) deposition, and was described in Chapter 9. This high-density plasma results in a dense CVD oxide film at temperatures as low as room temperature. When a separate RF bias is applied on the substrate, the angular dependence of the sputtering yield of the ions is exploited, as in bias-sputtered deposition, resulting in more planar and void-free films. HDP oxides can be doped with phosphorus and boron when used as a first level dielectric, although boron is not usually used since reasonably planar and conformal films are obtained without the need of a high-temperature reflow step. Fluorinated oxides for low-dielectric constant material can also be deposited by HDP. While HDP deposition can planarize any topography if enough film is deposited, it is usually used in conjunction with other planarizing techniques and is mainly utilized for its excellent hole or gap filling ability.

A planarization technique that has recently become part of mainstream processing is Chemical-Mechanical Polishing, or CMP. With this technique, global or near-global planarization can be achieved. As its name implies, CMP is a combination of chemical and mechanical polishing or etching. Wafers are held face down against a spinning polishing pad (made of a roughened polyurethane plate), and a colloidal silica slurry is kept between the pad and the wafer. The chemical composition and pH of the slurry are important and depend on the material being polished. For planarizing an oxide layer, a high-pH alkali-based solution is often used, while a low-pH, oxidizer-based solution is commonly used for metals. In general the chemical reaction at the wafer surface makes the surface material susceptible to mechanical abrasion by the silica in the slurry. For silicon dioxide removal, it is believed that the potassium hydroxide in the slurry reacts with the oxide to form a hydrated silicate layer, which is then removed mechanically by abrasion. For metal removal, the solution oxidizes the surface which is then polished in a similar fashion to the oxide removal.

Figure 11–27 shows a cross section of the CMP pad in contact with the top of a wafer. A semirigid polishing pad is used to allow for any bowing or thickness nonuniformities

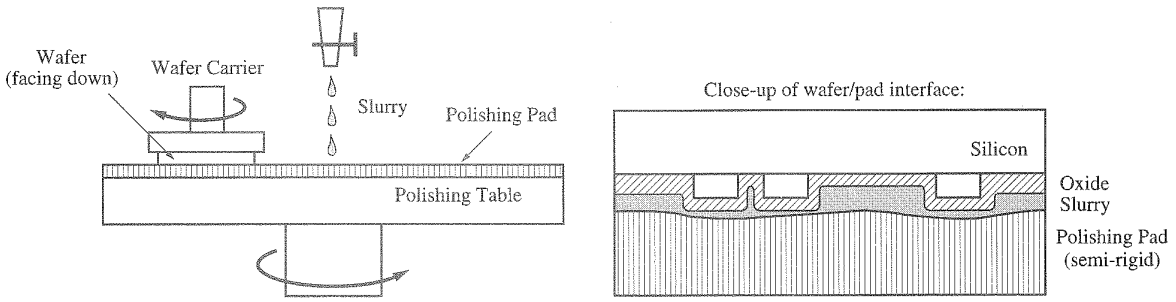


Figure 11–27 Illustration of Chemical-Mechanical Polishing (CMP) system. The right hand figure shows a close-up of wafer/pad interface.

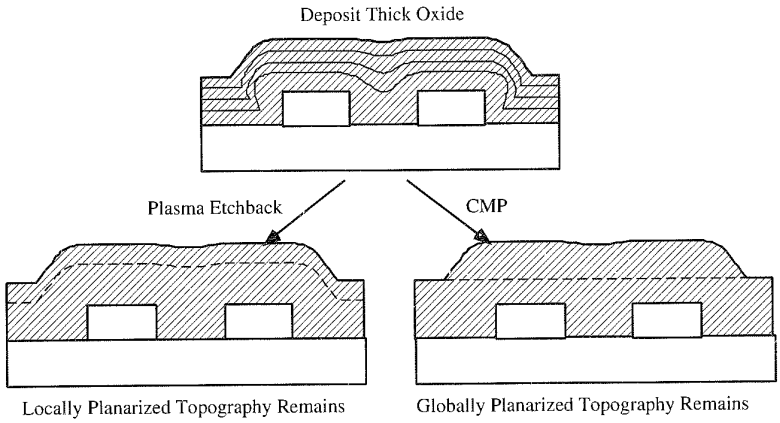


Figure 11–28 Illustrating the differences between using plasma etchback and chemical-mechanical polishing to planarize the dielectric.

in the silicon substrate. These would cause uneven etching if a perfectly rigid pad were used. The CMP process removes elevated features without much thinning of lower areas, thus planarizing the topography. CMP is used commonly for dielectric planarization, and is also being used for the W damascene process for the etchback of W (recall Figure 11–19). While standard plasma etchback processes achieve local planarization, global planarization is obtainable with CMP. Figure 11–28 illustrates this.

It should be pointed out that with all this effort at planarization, there is a drawback, other than just complicating the process. Without planarization, each layer is approximately uniform in thickness. The topography may go up and down, due to the presence or absence locally of features below it, but the vertical thickness of the layer is fairly constant (except just at the edges of features, of course). Therefore a process like etching holes in the layer is relatively straightforward since the depth one has to etch is constant from one area to the next. However, with planarization, local thicknesses of layers can vary significantly. As a result, one must etch different amounts in different locations to get to the level below. Fortunately this problem with planarization usually only occurs for the contact level. After the first dielectric level is planarized, the vias between consecutive interconnects in the upper planarized levels will have the same heights, as shown in Figure 11–1.

As with the contacts, vias, and interconnects, dielectrics have evolved from simple layers into more complex multilayered structures as the processing and circuit demands have increased. Low-temperature processing and planarization have been the two main driving forces in their evolution. Countless numbers of different structures and

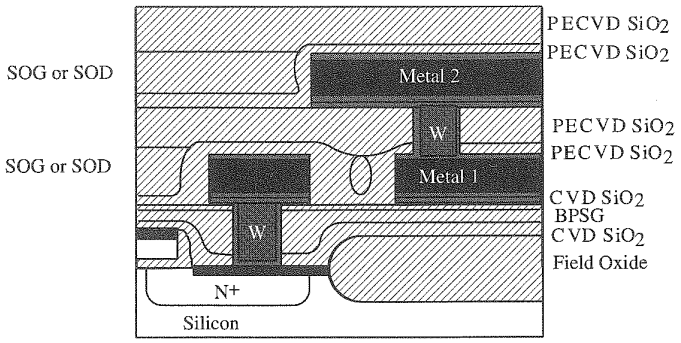


Figure 11–29 Schematic diagram of a back-end structure showing one possible dielectric multistructure scheme. Planarized topography is achieved. Other variations include HDP oxide or the use of CMP, either of which could reduce the number of sublayers —but the general structure would remain the same.

planarization schemes have been developed. Shown in Figure 11–29 is one possible, rather generalized, structure. It utilizes BPSG for the first-level dielectric, a W plug etchback, a Spin-On-Glass (SOG) or Spin-On-Dielectric (SOD) for the intermetal dielectric, and multilayer dielectric structures. The oxide layers could be either silane or TEOS based, using PECVD, bias sputtering, or HDP, and with or without an etchback or CMP. If an HDP oxide is used instead of BPSG or SOG/SOD, a moisture barrier below or above it may not be necessary. The intermetal dielectrics must be deposited and processed at temperatures less than 450°C, while the first-level dielectrics can often be done at temperatures up to 900°C or so.

Figure 11–30 shows SEM images of two current back-end structures. The first with three metal levels and 0.35-μm technology uses SOG and CMP, while the second one, a more advanced technology with five metal levels and 0.25-μm technology, uses HDP oxide and CMP.

11.3 Manufacturing Methods and Equipment

There are countless different structures and fabrication strategies for practical back-end structures, and we cannot hope to cover them all here. However, we will go through some representative processing steps and procedures used to fabricate CMOS integrated circuits. This discussion will provide some of the details that were left out in the CMOS process in Chapter 2. We will refer to that process throughout this section.

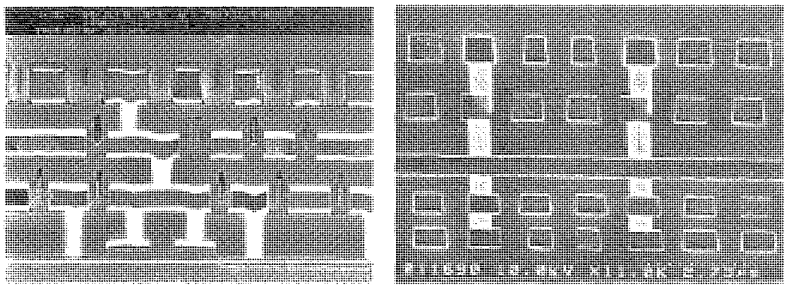


Figure 11–30 Two current back-end structures from VLSI Technology, Inc. The one on the left has three metal levels and uses encapsulated BPSG for the first-level dielectric, and SOG (encapsulated top and bottom with PECVD oxide) and CMP in the intermetal dielectrics. The multilayer metal layers and W plugs are also clearly seen. The SOG regions were deformed during SEM cleaving and erroneously look voidlike. The structure on the right, with five metal levels, uses HDP oxide (with PECVD oxide cap) and CMP in the intermetal dielectrics. Photos courtesy of VLSI Technology, Inc.

11.3.1 Silicided Gates and Source/Drain Regions

We start with a circuit that has been fabricated up to the gate and source/drain formation, with the source/drain N+ and P+ implants and anneals just finished (Figure 2–33 in Chapter 2). The next step is to silicide the source/drain regions and the polysilicon gate electrode to reduce their sheet resistances. $TiSi_2$, a common silicide for this process will be used here. Figures 2–34 to 2–37 show the process, and Figure 11–31 shows the silicide formation step.

After the SiO_2 is removed from the device source, drain, and gate regions, 50–100 nm of Ti is deposited by sputtering over the whole wafer. This is followed by a two-step anneal. The first anneal is at a low temperature, about 600–700°C for a time of 15–60 seconds. A low temperature is used to limit excess lateral growth and to minimize the reaction of the Ti with the SiO_2. Lateral growth refers to the $TiSi_2$ forming laterally out from the silicon contact regions, which could lead, for example, to shorting the gate to the source and drain. The anneal is done in a nitrogen ambient to further limit lateral growth. Where Si is exposed, the Si reacts with the Ti to form $TiSi_2$ at the bottom and N_2 reacts with the Ti to form TiN at the top. Over the SiO_2, TiN forms. This inhibits the formation of $TiSi_2$ over the oxide between the gate and source/drain regions. In addition, nitrogen is believed to stuff the grain boundaries of any unreacted Ti which slows down Si diffusion and helps in retarding lateral growth. The TiN, and any unreacted Ti, can then be selectively etched off using a solution of $NH_4OH:H_2O_2:H_2O$ (1:1:5).

A second anneal is then required to lower the resistivity of the silicide. This is because below about 750°C, a metastable phase of $TiSi_2$ forms, called the "C49" phase, which has a higher resistivity (60–70 μΩ-cm). An 800°C anneal for 30–60 seconds is required to transform it to the equilibrium, "C54," phase, with its lower resistivity of 13–20 μΩ-cm. Going to a higher temperature and longer annealing times for either anneal leads to problems. As mentioned above, a longer time in the first anneal leads to lateral encroachment and reaction with the oxide. For the second anneal, higher temperatures

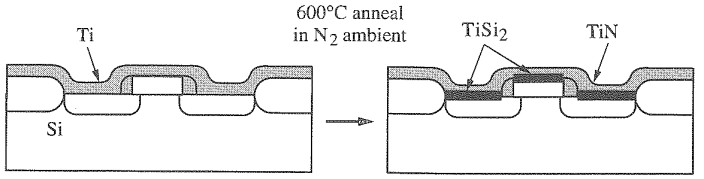


Figure 11–31 Formation of self-aligned $TiSi_2$ regions for gate, source, and drain. TiN also forms. After the formation anneal, the TiN may removed, and a second 800°C anneal is used to decrease the film resistivity.

and longer times lead to the silicide surface roughening up, or "agglomerating." A second problem is that dopants in the contacted Si substrate can diffuse up into the $TiSi_2$ because of their high diffusivities and solubilities in the polycrystalline silicide. So anneals at minimum temperatures and times required to form and transform the silicide are used. This means seconds to minutes for anneal times, necessitating the use of RTA equipment for both anneals.

One option during this process is to mask off some of the TiN area before the selective wet etch and plasma etch, leaving some TiN behind as a local interconnect. This was done in the CMOS process flow in Chapter 2 (Figure 2–37). The second, high-temperature anneal not only reduces the $TiSi_2$ resistance, but also the TiN sheet resistance. The TiN can also be left in the source/drain or gate areas to act as an Al/Si barrier layer for subsequent Al deposition on top. However if a TiN local interconnect, which requires an additional masking step, is not used, the TiN is etched away everywhere. New TiN would then be deposited later over the $TiSi_2$ in the contacts just before the contact metal (W, for example) is deposited.

The thickness of the $TiSi_2$ is dependent on how much Ti is deposited and the time and temperature of the anneals. For a 100-nm layer of Ti, and reaction of all 100 nm of the Ti with the underlying Si, 227 nm of the Si will be consumed, and 250 nm of $TiSi_2$ will be formed (the reaction ratios for $TiSi_2$ and other silicides are given in Table 11–4). This would give a film sheet resistance of approximately $15 \times 10^{-6}\ \Omega$ cm / 2500×10^{-8} cm $= 0.6\ \Omega$/square. But less silicide is produced if the first anneal is done in N_2, since the top part of the Ti reacts with the nitrogen to form TiN instead of the silicide. The actual silicide thickness in this case depends on the relative kinetics of the silicide and nitride formation reactions and will depend on anneal time and temperature. Typical final thicknesses for 100 nm of deposited Ti are about 150 nm of $TiSi_2$ and 75 nm of TiN on top, giving a silicide sheet resistance of about 1.2 Ω/square. The amount of Si consumed would be about 140 nm in this case. The TiN formed on top of SiO_2 and left as a local interconnect would be about 100 nm thick and have a final sheet resistance of about 10 Ω/square. As the diffused junction depth beneath the silicide decreases as the technology progresses, the silicide must also decrease in thickness so that the amount of Si consumed is not greater than the junction depth. In fact, it is typically desirable to leave about 50 nm between the bottom of the silicide and the junction, or else leakage current through the junction may become too large.

While Ti will reduce and break through any thin native oxide on the silicon or polysilicon surfaces, it is still important to clean the wafer surface before the Ti deposition. Polymers, usually C-F compounds, left over from previous dry etching must be removed. In addition, if the native SiO_2 is too thick (greater than about 2 nm), rough $TiSi_2$ surfaces can result. This is similar to the process that occurs in Al spiking into Si. In this case, the Ti will reduce the oxide at weak points first. The Si diffuses through these points and moves rapidly into the adjoining regions before the oxide in other areas is reduced. This leads to faster Si consumption in some areas, and slower consumption in others, resulting in a rough $Si/TiSi_2$ interface and a rough $TiSi_2$ surface. Only if the native oxide is thin will uniform reaction occur, giving smooth silicide films. To minimize the native oxide thickness, a wet HF clean is done immediately prior to loading the wafers into the sputtering equipment. In addition, an RF plasma sputter clean in the sputtering chamber can remove the native oxide and any polymer films. However, too harsh a sputter clean can damage the Si surface. It is also important to minimize any oxygen or water on the wafer surface or in the sputtering chamber since the Ti will react with these species and incorporate oxygen into the silicide film, increasing its resistivity. Minimizing oxygen in the RTA chamber is also important.

11.3.2 First-level Dielectric Processing

After silicide formation, the first-level dielectric layer is deposited. This corresponds to Figure 2–38 in Chapter 2. This physically and electrically separates the local interconnects from the global interconnects, the latter being Al-based metals. Since the topography is usually nonplanar at this stage, and there are many layers that follow which can suffer as a result, it is especially important to planarize the structure here.

The first-level dielectric is usually SiO_2, doped with P and B at 2–6 wt. percent, at a thickness of about 500 nm. This BPSG (borophosphosilicate glass) can be deposited by CVD using the silane reaction [Eq. (9.32)] or the TEOS reaction [Eq. (9.33)], as discussed in Chapter 9. Doping can be done using hydrides for the silane case and organometallic compounds for the TEOS case. The silane deposition can be done using either APCVD or LPCVD at temperatures from 380–500°C, or using PECVD at lower temperatures. The TEOS deposition can be done at 650–800°C, but this temperature can lowered to 400°C or less by using either PECVD or an ozone-enhanced reaction [Eq. (9.34)].

The TEOS films give more conformal coverage than the silane films. As mentioned earlier the $TEOS/O_3$ films, which can be very conformal, can be quite porous and absorb moisture. Absorbed water can have an adverse affect on electrical behavior and reliablity, as well as subsequent processing. Thin underlayers, and sometimes overlayers, of undoped CVD SiO_2 using silane or PECVD films using TEOS are often used as moisture barriers for $TEOS/O_3$ films, or as dopant buffer layers for either silane or TEOS films.

To obtain a smoother and more planar topography, a thermal reflow anneal is done after the PSG or BPSG deposition at 800–1000°C. (The lower temperature is appropriate for BPSG films; the higher temperatures are used for PSG films.) The reflow is often done in a steam environment. Doing so lowers the reflow temperature by about

70°C. It also leaches the dopants from the surface to avoid corrosion problems with the overlying Al, but this is not as important when buffer or sandwich layers of undoped SiO_2 are used.

The as-deposited PSG or BPSG is amorphous and not as dense as thermally grown oxide. A dense film is needed to improve both electrical and physical properties, so a densification anneal is done. This is usually done at temperatures from 600–1000°C, before the reflow anneal, but it can be done simultaneously with the reflow. The thickness of the film decreases as the arrangement of the SiO_4 tetrahedra becomes more regular.

Reflowing the PSG or BPSG is also often done after the contact holes are etched to round off the contact hole corners, resulting in better filling of the holes with the contact metal. This is done in an inert atmosphere so that the bottom of the contact is not oxidized and is done in a separate step from the steam reflow used to planarize the surface.

As an alternative to reflow, or in addition to it, an etchback step is sometimes done after first-level dielectric deposition. This can done either by resist etchback, direct etchback of the dielectric, or with Chemical-Mechanical Polishing (CMP). This was the process illustrated in Figure 11–28 and Figure 2–39 in Chapter 2. The latter provides more global planarization. If an etchback is done instead of a reflow, the film does not need to be doped with boron, but a phosphorus doped layer should be present somewhere in this level for Na gettering.

The recently developed method of High-Density Plasma (HDP) deposition of oxide is being more commonly used for the first-level dielectric as well as intermetal dielectric levels (see Figure 11–30). Silane is preferred since one can achieve planar and void-free films without having to use TEOS, with its inherent disadvantages over silane (carbon residue in films and a more difficult source to use, for example). Since the film is dense and a better moisture barrier than some TEOS-based CVD oxide films and spin-on-dielectrics, it might not need an additional dielectric layer deposited under it as the other planarizing films do. If HDP oxide is used, boron is not needed. Because of the sputtering of the wafer and the high-plasma density, heating of the wafer can occur, and cooling of the substrate via a thermal chuck must be used to cool the wafer and control the temperature. Often CMP is used in conjunction with HDP oxide deposition to achieve global planarization.

11.3.3 Contact Formation

The next step is to open up holes in the first-level dielectric level and form metal contacts to the silicided gate, drain, and source regions, as well as to any local interconnects. This step corresponds to Figure 2–40 in the CMOS process flow in Chapter 2. Anisotropic plasma etching, using CHF_3 for example, is commonly used to open up these areas. It is important to make sure that all of the dielectric material is removed with the etch so that a low-resistivity contact is made. In addition, any residues from the etching process must be removed. These include both organic residues, such as C-F containing polymers from the etch chemistry and photoresist, as well as inorganic residues, such as SiO_2 that have been resputtered from the sidewalls of the contact hole. Usually a cleanup process following the etch and photoresist strip is done which includes both a wet chemical resist clean and an O_2 plasma etch or "ash."

The contact etch is complicated by the differences in dielectric thickness from position to position, as illustrated in Figure 11–1. Because of the planarization techniques used, the thickness of dielectric layer that must be etched through to reach gates, source/drains, or local interconnects can vary by as much as a factor of two. All contact holes must all be etched simultaneously, requiring very selective etching, or else expensive additional mask levels must be used.

Before the contact metal is deposited, the contact/adhesion/barrier structure is formed (see Figure 2–41 in Chapter 2). If the gate and source/drain regions are already silicided, then that layer can act as the contact/adhesion layer. If not, then Ti can be deposited at this point by sputtering. (In many cases, Ti is deposited here in any case to reduce any oxide as well as help with adhesion.) The Ti can then be annealed to form $TiSi_2$ or it may form during a later heat step. An argon *in situ* sputter etch is often done just prior to Ti deposition to clean the surface. This etch must not cause too much damage, so a "soft etch" is done, with a DC bias of only 100–300 V. Usually a few tens of nanometers of Ti are deposited. It is important that the Ti cover the entire bottom of the contact hole, which is not an easy task for today's narrow but deep contact holes. Collimated or ionized sputtering may be used for this purpose.

Next several tens of nanometers of TiN is deposited. For Al contacts this serves as a barrier layer. For W contacts it serves as an adhesion layer between the W and the dielectric as well as a barrier to prevent WF_6 from reacting with the underlying Ti or $TiSi_2$ during the W deposition process. This TiN deposition can be done by sputtering (including ionized PVD), or by CVD. The TiN is needed on the bottom of the contact holes (to act as a barrier), and also on the sides of the contact hole and the top of the dielectric layer to help with W adhesion. CVD deposition usually provides better coverage or filling on the contact bottom and sidewalls than sputtering. After the TiN is deposited, it can be annealed in N_2 or O_2 at 400–500°C for 30 minutes to improve its barrier properties. The N_2 and O_2 stuff the grain boundaries, limiting diffusion by providing a physical barrier in the fast diffusion paths. There can also be a chemical effect, in which the oxygen bonds with any Al diffusing through the TiN, for example.

If a W plug is used as the contact metal, the W is now deposited. As discussed in Chapter 9, W can be deposited either by selective or blanket deposition. The selective process is very difficult to control, and so blanket W is more commonly utilized. This is the process illustrated in Figure 2–41 in Chapter 2. A one-step process, using either the hydrogen reduction reaction [Eq. (9.35)] or the silane reduction reaction [Eq. (9.36)], deposits W in the contact hole as well as on top of the dielectric. A two-step deposition process, rather than a one-step process, is commonly used. A silane reduction reaction is used to deposit the first few tens of nanometers of W and results in a film with good adhesion but with poor step coverage. Next, the hydrogen reduction reaction is used which is slower but has much better step coverage and also leads to less void formation. The disadvantage to the blanket W deposition method is that it deposits W everywhere, including on top of the first-level dielectric oxide where it is not wanted. However, the contact hole can be filled with a W plug by doing a thick blanket deposition and then doing an etchback in the damascene process. This process was illustrated in Figure 11–19 and is also the process illustrated in Figure 2–42 in Chapter 2. The etchback can be

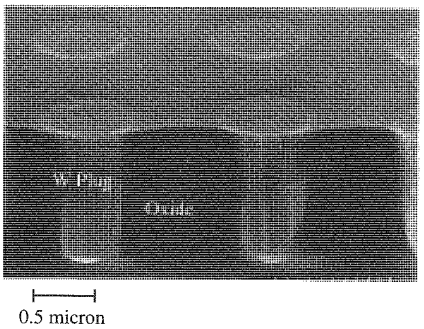


Figure 11–32 SEM image of W plugs after blanket CVD deposition and CMP. Photo courtesy of VLSI Technology, Inc.

done either by dry etching or CMP, although the latter is much more commonly used today. Figure 11–32 shows a SEM image of an array of W via plugs fabricated by blanket W deposition followed by CMP in this case. One can see that very planar topography is achieved.

W contact and via plugs are currently preferred over Al plugs because of more conformal filling by W and several years of experience with this process. There is an effort to develop manufacturable processes using Al with good filling properties, in which case Al could replace W in this step. The Al would be blanket deposited, perhaps with a reflow process during or after the deposition to ensure complete filling. This would be followed by an etchback step, dry etching or CMP, just like the W damascene process. An even more attractive alternative is to deposit the Al for contacts or vias and the interconnect in one step. Methods to do this are discussed in the next two sections.

11.3.4 Global Interconnects

The global interconnects are next deposited. This corresponds to Figure 2–43 in the CMOS process flow in Chapter 2. These are usually made of Al with 0.5–1 wt. percent Cu. Si is still often added, about 1 wt. percent, even though spiking of the Al through the substrate is not a problem if barrier layers are used. (Si may suppress reaction of the Al with Ti containing layers.) The Al layers are usually clad with thin Ti or TiN layers. Typical thicknesses for a Ti/TiN/Al/TiN (bottom to top) structure would be 40 nm/50 nm/500 nm/50 nm. Before deposition, the surface area of the wafer, especially the exposed contact regions, must be clean to ensure good electrical contact between the interconnect and the contact metal. To do this, wet and plasma cleans similar to those for contact deposition are done. Since many low-temperature oxides evolve some amount of water, it is often necessary to outgas the wafers in the sputtering chamber prior to metal deposition. The interconnect stack is then deposited, usually by sputtering. To minimize interface contamination the depositions are done one layer right after the other without opening the chamber.

As mentioned in Chapter 9, the microstructure of the deposited films are very dependent on the deposition conditions. In the case of sputter deposition, the grain size of

the films will be dependent on the substrate temperature, the sputtering conditions (power, pressure, etc.), the vacuum integrity, the surface topography, and the underlying layers. The grain size of deposited Al ranges from 0.5 μm to about 2 μm. During any subsequent annealing, such as during the alloy anneal, the grains can grow (actually the large grains will grow at the expense of the smaller ones). They will grow until they reach a diameter equal to about two to three times the film thickness, so the average grain size of Al interconnects will be about 1 to 2 μm after annealing. The grain structure of deposited Al is normally columnar, with most of the grains spanning the entire thickness of the film with the <111> crystallographic direction oriented upward. The film's texture (crystallographic orientation) also depends on these same deposition conditions. Both grain size and texture influence many other properties of the films, including mechanical, electrical, and reliability properties.

To achieve both filling of deep holes as well as providing good step coverage and better planarization for the interconnect level, one could heat the aluminum, either during the deposition itself or right afterwards. The first method is called "hot aluminum deposition," using the high-temperature, or "hot-," sputter deposition technique. The substrate holder in the sputtering chamber is heated to 450–550°C during the deposition. In the second method, the wafer is heated to this temperature range immediately after the Al deposition, right in the sputtering chamber, and before any capping layer is deposited. In either case, the Al reflows, providing better filling of contact holds and smoother topography. Hot-sputtered Al deposition or deposition followed by reflow can also be used in an Al dual-damascene process where the contact and interconnect levels are again deposited in one step. These hot Al methods are still being developed.

Because reflow is a surface and interface controlled process, it is very dependent on ambient and interface conditions. For example, any oxygen in the chamber will form a thin Al oxide on the surface and inhibit reflow. When using hot Al deposition or doing an Al reflow, it is especially important to have a good barrier and adhesion layer between the Al and underlying layers. Thin Ti/TiN layers are usually used for this. Often a "cold/hot" Al deposition is done for this process. Right after the barrier/adhesion layer is deposited a thin Al layer is deposited at room temperature, often using collimated PVD. This acts as a continous "seed" or nucleation layer. CVD Al may also be used for this. Next the bulk of the Al is deposited by PVD at a higher temperature (400–500°C) for *in situ* reflow, with good adhesion and conformal filling in the contact hole, as well as good planarization and step coverage elsewhere. Topography simulations of reflow used during deposition were shown in Chapter 9 (Figure 9–52). CMP of the Al may still be done after the deposition/reflow but it is much easier and more efficient when the metal is already planarized to a degree.

To pattern the Al-based interconnect structures, a chlorine-containing plasma etch is used, usually in an RIE mode. Because chlorine and water or moisture in the air can corrode the Al, the wafers are passivated immediately after the chlorine etch and before they are removed from the etch chamber and subjected to the atmosphere. The Al surface is passivated by subjecting the wafers to a fluorine-containing plasma, such as CF_4, which replaces the residual Cl on the Al surface with F.

An alternative to etching the global interconnect is to use the dual damascene method to pattern the interconnect, which will be discussed in the next section.

11.3.5 IMD Deposition and Planarization

The intermetal dielectric (IMD) levels are deposited between the global interconnect levels, and some sort of planarization technique is usually used (see Figure 11–1). Currently most intermetal dielectric materials are SiO_2 based, often with multilayers utilizing different deposition techniques to obtain SiO_2 films with different properties. The total thickness of the level is around 0.5 to 1 μm. The SiO_2 films can be deposited at low temperatures (≤400°C) by CVD (including PECVD, LPCVD, HDPCVD, and AP/SACVD) using silane or TEOS, or can be spun on using spin-on-glass. The CVD methods were discussed earlier in Chapter 9.

Spin-on-glass (SOG) provides a way to fill spaces between features with a dielectric and to achieve local planarization. The starting material can contain either inorganic silicates or organic siloxanes. These are dissolved in an alcohol-based solvent, such as ethanol. The liquid is spun on the wafer like photoresist, with the thickness dependent on the spin speed: Faster spinning gives thinner layers. The wafer is then baked and cured, usually in a two-step process. The first anneal is at 150–250°C and the second anneal is at 350–450°C. The solvent is first driven off. Then the film condenses and polymerizes, forming Si-O-Si bonds to produce the SiO_2 layer. Water is produced by the polymerization reaction and must be driven off. Considerable shrinking of the film occurs during these processes, and high-tensile stress can result which can cause cracking of the film. The two types of SOG starting materials lead to different film properties. Some carbon is retained in the films made from siloxane materials and decreases the Si-O-Si polymerization. This decreases the film shrinkage and stress and increases the cracking resistance, but makes the film more porous and more susceptible to chemical attack. With either type of SOG, the final film quality is not as good as CVD SiO_2 due to its porosity, stress, and residual moisture content, and sandwich layers of CVD SiO_2 are usually used in conjunction with these films. The SOG films can be etched back, as was illustrated in Figure 11–26(b), to provide further planarization and to avoid poisoned vias. This is usually done with the organic, siloxane-based films since they are more prone to this problem.

Spin-on-dielectrics (SOD) other than SiO_2 are starting to be used. They not only provide planarization but also lower dielectric constants which reduce interconnect capacitances. Their processing is similar to that of SOG. Other low-K dielectrics, such as fluorinated SiO_2, are deposited by CVD. Some low-K dielectrics can be deposited by either spin-on techniques or CVD.

High-density plasma CVD oxides have gained popularity for this level as well as for the first-level dielectric. As with its use in that application, an additional dielectric layer deposited under it may not be needed, and CMP is still usually done in conjunction with HDP oxide deposition to achieve global planarization.

Other planarization techniques used on the intermetal dielectric level include etchback, Chemical-Mechanical Polishing (CMP), and the dual damascene process. In both the etchback and CMP cases, thick dielectric or photoresist layers are deposited which become more planar locally as the thickness increases. In the case of etchback, the film is then etched uniformly in thickness in a plasma etch system, resulting in a locally planar structure. With CMP, the structure is etched more or less uniformly in space across a horizontal plane, resulting in global planarization, as was illustrated in Figure 11–28.

Even better results can be obtained if "dummy" oxide or metal features are included in less dense areas to make the process more uniform across the entire circuit. Mechanical damage from the polishing as well as chemical and particle contamination from the slurry are concerns. But fairly successful procedures to minimize surface damage and to clean up the chemical residue have been developed. Endpoint detection can be difficult. Some systems detect the change in friction between the wafer and polishing pad when one layer is polished away and the layer below is contacted. Other methods include timed polishing or newly developed *in situ* film thickness measurement techniques. One such method is an *in situ* optical system which uses a window in the polishing pad and measures the interference patterns of reflected light [11.20]. CMP is now the dominant planarization technique in IC fabrication.

We saw earlier how the damascene process can be used for the contact level in which W is used to fill the contact hole. The damascene process could also be used for the interconnect level. In this process the interconnect metal, Al for example, is deposited at high temperature or by collimated or ionized sputtering (all to enhance filling) into interconnect openings in an intermetal dielectric level, then etched back or polished using CMP. This process is especially applicable for copper interconnects, which are beginning to replace Al in some cases. Cu is difficult to etch, and in the damascene process when CMP is used, etching of the metal is not required. In addition good filling of Cu in holes in trenches can be achieved using well-established Cu electroplating techniques.

A more advanced version of the damascene process provides both the via/contact and interconnect levels simultaneously. In this "dual damascene" process, both the openings in the IMD for the metal interconnect and for the contact or vias underneath are opened, one after the other. The metal is then deposited into both layers at once which is then followed by a CMP etchback. The dual damascene process is illustrated in Figure 11–33. An "etchstop" layer, such as a thin Si_3N_4 film, is usually deposited halfway through the thick intermetal dielectric (IMD) layer. When the etching is done for the interconnect trench in the top half of the IMD layer, the etching can easily be stopped when the nitride layer is reached, either due to the etch selectivity or by endpoint detection techniques, or both. This ensures that the interconnect trench is only etched halfway through the IMD.

11.3.6 Via Formation

Via formation is done the same way as the contact, as described above, with TiN used as an adhesion/barrier layer, followed by a W plug using the damascene/etchback method. Ti can be deposited under the TiN to help with adhesion and contact. However, if WF_6 from the W deposition process gets through the TiN, a volatile reaction can occur between it and the Ti. Al could be used as the via metal instead of W, just as was the case for the contacts. This requires good filling ability of the Al, which can be enhanced by using collimated or ionized deposition, hot Al deposition, or Al reflow, as discussed previously.

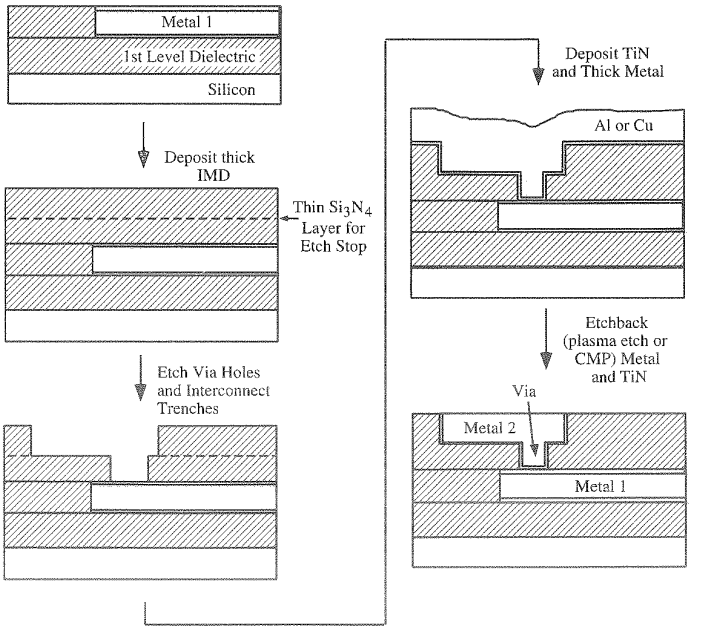


Figure 11-33 Schematic illustration of the dual damascene process.

11.3.7 Final Steps

It is now just a matter of repeating the global interconnect, intermetal dielectric, and via formation steps until the required number of metal layers are fabricated. Generally, the only difference is that as the upper levels of metals are added, the thickness and width of the metal lines often increase. More current can then be carried by the upper levels, with less probability of electromigration. Before bonding and packaging, a final passivation level is added. This would typically be a phosphorus-doped silicon dioxide layer, 500 nm thick, and a 1-μm thick silicon nitride layer. Both are deposited by PECVD at about 400°C. Bond pad openings are etched, and then a final alloy step is done at 400–450°C for 30 minutes in forming gas (10% H_2 in N_2). This alloys the metal contacts in the structure, reduces some of the electrical charges at the Si/SiO_2 interface, and reduces radiation damage in the gate oxide induced during sputtering processes.

11.4 Measurement Methods

A wide assortment of measurements are made on back-end layers and structures. These include morphological, electrical, imaging, chemical, structural, and mechanical measurements. The thickness of each layer is of course important, as are the electrical properties, especially the sheet and contact resistances of the interconnects. Also important

is the reliability of the interconnects, including problems of electromigration and the formation of voids and hillocks. The electrical integrity of the dielectric layers is critical, as are their chemical and structural makeups. Imaging the structures, using electron and ion beam microscopy, is essential to diagnosing and solving problems that occur during their fabrication and use. Imaging techniques generally use standard optical microscopes, SEM or TEM methods (Figures 11–2, 11–22 and 11–30, for example), which have been extensively discussed in earlier chapters. We focus here mostly on measurement methods unique to back-end structures.

11.4.1 Morphological Measurements

The morphology of a material or structure refers to the thickness, surface roughness, and grain size. The thicknesses of dielectric layers are usually measured using the optical methods described in Chapter 6. The interconnect layers are commonly measured either by using the stylus/etch step method described in Chapter 6, or by electrical resistance measurements. The sheet resistance, ρ_S (in Ω/square), of the layer equals the sheet resistivity, ρ (in Ω cm), divided by the layer or film thickness, x_f:

$$\rho_s = \frac{\rho}{x_f} \tag{11.14}$$

Thus a measurement of ρ_s using a four-point probe, as described in Chapter 3, directly gives the layer thickness provided that ρ is assumed to be known as a basic material property. This measurement is generally done on test wafers on which the interconnect material is deposited at the same time as the device wafers. Other, more exotic, metal thickness measurement techniques involve acoustic or sonar measurements, including laser-generated sound wave and echo detection systems. Though more complicated, these measurements can be done nondestructively on metal lines and films without test structures or test wafers.

Surface roughness, or local nonuniformities in thickness, can be measured by either the stylus method, or by a related technique, Atomic Force Microscopy (AFM). Here an extremely sharp tip is raster-scanned over the sample surface, and its vertical deflection is measured very accurately generally using laser techniques or piezoelectric sensors. A 3D mapping of the surface topography can be acquired. The measurement sensitivity is so high that surface grain size can often be determined.

11.4.2 Electrical Measurements

The primary electrical measurements made on back-end structures are sheet resistance and contact resistance of the interconnects and breakdown voltage of the dielectrics. In addition, the reliability of the interconnects is often measured electrically by accelerated electromigration measurements.

The sheet resistances of the interconnects directly affect the electrical performance of the circuit. In addition, they can provide information about the film itself. The sheet

resistances of the interconnects are often measured by four-point probe measurements or using resistor structures, and depend on layer thickness, chemical composition, and structure. The sheet resistivity of Al lines generally increases as the doping with Si and Cu increases. Also, the grain size of polycrystalline films affects the resistivity: the smaller the grain size the higher the resistivity due to more electrical scattering at the grain boundaries. In the case of $TiSi_2$, the crystal structure affects the resistivity, with the C49 phase having a higher resistivity than the C54 phase.

Specific contact resistivity was described earlier in connection with Figure 11–7. What is usually measured experimentally is the contact resistance, R_c, obtained by measuring the current through the contact for a given voltage across the contact interface. R_c is related to the specific contact resistivity by

$$R_c = \frac{\rho_c}{A} \tag{11.15}$$

where A is the contact area and R_c is in Ω. If the current flowing across the metal-semiconductor interface were uniform, then A would just be the actual area of the entire contact. One could measure V and I across a contact to get R_c, then multiply by the contact area to get ρ_c. However in actual contact structures, the current flow can be highly nonuniform.

Figure 11–34 shows the cross section of a simple Al-to-silicon contact. The current in this example is flowing from the left in the semiconductor, up through the contact, and to the right in the interconnect line. Since the electrical resistance in the N^+ silicon is much greater than in the interconnect line, the current preferentially takes the route of lowest resistance, leaving the semiconductor and entering the metal interconnect close to the left edge of the contact. Thus the current density is higher at the left edge of the contact as it flows up into the metal interconnect. This is called "current crowding." The effective area of the contact is thus decreased from the actual area.

A transfer length, ℓ_t, is defined by the equation

$$\ell_t = \sqrt{\frac{\rho_c}{\rho_s}} \tag{11.16}$$

where ρ_s is the sheet resistance of the N^+ doped silicon region in Ω/square. The transfer length is the length within which most of the current flows through the contact interface, as measured from the leading edge of the contact. (More precisely, it is the

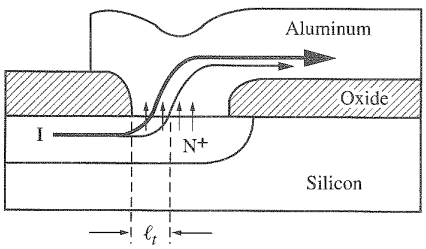


Figure 11–34 Schematic cross section of contact region, showing current crowding toward one end of the contact due to the higher resistivity of the silicon compared to the Al. The transfer length, ℓ_t, defined by Eq. (11.16) and within which most of the current flows, is also indicated in the figure.

distance from the edge of the contact at which the current density drops to $1/e$ of the density at the leading edge.) For ρ_c equal to $10^{-7}\ \Omega\ \text{cm}^2$ and ρ_s equal to 100 Ω/square, the transfer length, ℓ_t, is 0.3 μm. The voltage also varies across the contact since the current varies. So depending on how the actual measurement is made and how the contact structure is laid out, different values of R_c may result.

To take the nonuniform current flow into account, several test structures have been developed to measure contact resistance and resistivity. An excellent reference on this topic is Schroder's book [11.21]. In these structures, a measurement of voltage drop across the contact with current flowing through the contact is made. Then 1D or quasi-2D electrical models are used to relate those values to the actual contact resistivity. In each of these methods, it is assumed that the contact width—that is, the dimension of the contact into the page in Figure 11–34—equals the diffusion (N^+ region) width. This is not always a good approximation as we will discuss later. It is also assumed that the metal resistance is negligible compared to the N^+ region resistance, which is generally a reasonable assumption. Even if silicide layers, TiN barriers, and W plugs are used between the silicon and the metal in the contact hole as shown in Figure 11–21, the N^+ region will dominate the resistance in the lateral direction. Thus accurate measurement of the contact resistivity between the silicon and different contacting layers or multilayer structures can be made with this method.

While a number of test structures are utilized for measuring contact resistance [11.21, 11.22], the simplest and most common is the "cross-bridge Kelvin structure," shown in Figure 11–35. In this structure, the voltage is measured at right angles to the current flow, so that an "average" voltage is measured. In this case, Eq. (11.15) can be used directly, with R_c (called R_k in this case) equaling the measured V divided by the measured I. Thus the contact resistance is simply

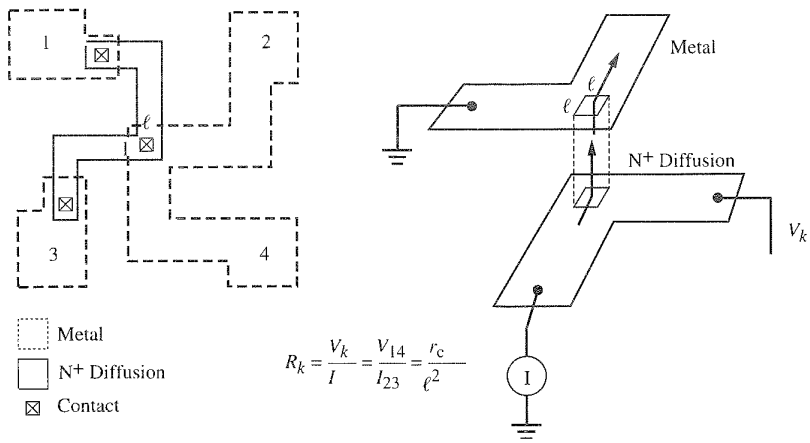


Figure 11–35 Cross-bridge Kelvin structure used to measure an average contact resistance, called R_k in the figure. (After [11.22].)

$$R_k = \frac{\rho_c}{\ell^2} \tag{11.17}$$

where ℓ is the size of the contact. Other more complex structures are described in detail in [11.21].

Example Using the cross-bridge Kelvin structure with a 1-μm-by-1-μm contact, it was found for that for a current of 10 μA through the contact (I_{23} in the structure), the voltage difference across the contact (V_{14}) was measured to be 320 μV. What is the specific contact resistivity for this contact?

Answer For the cross-bridge Kelvin structure, the average contact resistance, R_c (or R_k) is simply V_{14}/I_{23}. Then the specific contact resistivity, ρ_C, can be calculated from Eq. (11.17).

$$\frac{V_{14}}{I_{23}} = R_k = \frac{\rho_c}{\ell^2}$$

$$\frac{320\ \mu\text{V}}{10\ \mu\text{A}} = 32\ \Omega = \frac{\rho_c}{\ell^2}$$

$$\rho_c = 32\ \Omega(1 \times 10^{-4}\text{cm})^2 = 0.32\ \mu\Omega\text{cm}^2$$

Unfortunately, the current through the contacts may not only be nonuniform in the main direction of the current flow (shown in Figure 11–34), but can also be nonuniform and non-ideal in the lateral dimension, around the contact. This is because the contact width does not necessarily equal the N^+ diffusion region width, and two-dimensional current crowding and flowing around the periphery of the contact can occur, as shown in Figure 11–36. This results in an overestimation of ρ_c in experimental test structures which gets worse as the contact area gets smaller. Loh [11.22, 11.23] did extensive 2D computer simulations to take this into account. He plotted his results in a generalized, universal format which can be utilized to convert measured values of R_c to correct values of ρ_c.

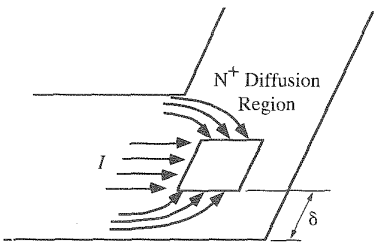


Figure 11–36 Current flow in lateral direction, showing current crowding and flowing around the periphery as it enters the narrower contact region. (After [11.23].)

It is important to note that in an actual circuit structure, the entire N^+ diffusion region is often strapped with silicide, as illustrated in Figure 11–1. This reduces the contact resistivity appreciably because the silicon-to-silicide contact area is significantly increased, with the contact area now defined by the silicide area and not the contact hole area. Using the test structure in Figure 11–35 with the silicide strapping the entire diffusion region would only measure the metal-to-silicide contact resistivity and not the silicide-to-silicon contact resistivity which usually dominates the contact resistance. One could use the silicide as part of a metal stack, forming it only in the contact hole, and use the Kelvin test structure to measure the silicide-to-silicon resistivity. However, this would involve structures and processing that would be different than those actually used in the circuit fabrication. Other test structures have been developed to measure self-aligned silicide contacts in which the silicide straps, or covers, the entire S/D regions [11.24]. These structures are more representative of the actual processes that are used in modern technologies.

As a process monitor, contact resistance is also measured in long contact chains. These are hundreds, sometimes thousands, of contacts that are in series with each other, located in the scribe streets or in test chips on the wafers. These contact chains pass over various topographies. A current is forced through each of these long chains, and the measured voltage is a measure of the average contact resistance. Via resistance is also monitored in this manner. These structures are used to monitor the contacts and vias as a function of processing conditions and structures, and to measure lot-to-lot variability. A high value of resistance in these structures could indicate a problem with incomplete etching, etch residue, poor metal deposition, stress-induced voids in contact regions, or other problems.

The quality of a dielectric layer is usually defined in terms of its breakdown voltage, or field strength, expressed in units of $V\ cm^{-1}$. In this measurement, a voltage is applied across the dielectric layer and is ramped up at a constant rate (see Figure 4–15). The voltage is monitored, and at a certain point, the voltage drops abruptly. This occurs when the dielectric "breaks down" and excessive current flows through it, discharging the applied charge. The physical mechanism for oxide breakdown may involve bond-breaking or hole trapping at defect sites, leading to changes in the structural and electronic properties of the material. These defects can include weak Si-O bonds, oxygen vacancies, surface roughness, local thinning, and sodium or metal contaminants in the film. An alternate test is the constant current test (as opposed to the voltage-ramp test). Here a current is passed through the dielectric at a constant level until the oxide is broken down. The charge-to-breakdown is measured, and is on the order of $10\ C\ cm^{-2}$ for a good oxide. Other breakdown measurements include time-to-breakdown under constant voltage, and charge-to-breakdown under ramped current. Chapter 6 discusses some of these measurements in connection with thermal oxides.

The reliability of interconnects is measured by electrical tests. Common causes of failure of interconnect lines are void and hillock formation, either due to stresses on the lines or due to electromigration. Stress-induced voids can be detected simply by measuring the electrical resistance of the line. An open circuit would indicate a line-spanning void. When shunt layers such as TiN are used, a void which spans the entire Al layer would not cause an open circuit since current could still flow through the shunt

line. In this case, a failure would be defined for a certain percentage increase in the line resistance, perhaps 20%. Stress-induced hillocks which crack overlaying dielectric layers and contact adjacent interconnect layers can be detected by short circuits. Failure due to electromigration—current-induced void and hillock formation—is measured using accelerated lifetime tests.

It would take too long to monitor the interconnects for electromigration under normal circuit operating conditions. Therefore, tests are done at elevated temperatures and higher than normal current levels, so-called accelerated lifetime tests. In these tests, current at elevated densities (about 1–2 MA cm^{-2}) and temperatures (200–225°C) is forced through an interconnect line. The voltage is monitored through taps to this line, and the resistance is measured. An open circuit, or even a partially open circuit, indicating void formation due to electromigration, is thus detected by an increase in the resistance. A second line adjacent to the line under test is monitored as well. A short circuit to this line indicates hillock formation and extrusion. Again, a certain percentage change in resistance (20% or so) defines a line failure. The time to failure for a number of lines is determined and statistically analyzed. This is done by plotting the cumulative percent that have failed versus time on a lognormal plot. The median time to failure, or t_{50}, is found by finding the time at which 50% cumulative failure occurs. The Median Time To Failure (MTTF) has been found to follow the general relationship:

$$\text{MTTF} = AJ^{-n} \exp\left(\frac{E_A}{kT}\right) \tag{11.18}$$

where J is the current density, n is typically close to 2, although values ranging from 1–3 have been reported, and A is a constant which depends on film structure (grain size, etc.) and processing. E_A is the activation energy for electromigration and is often associated with Al grain boundary diffusion. Its value ranges from 0.5 to 0.8 eV. Once the exact values for this expression have been determined experimentally at the elevated temperatures and currents for the particular structure and process, one can extrapolate down to the normal operating conditions to determine how long the interconnects will operate before failure occurs. This of course assumes that the failure occurs by the same mechanisms at the lower temperatures and current densities, which may not always be the case.

Example Under accelerated testing at 225°C and a current density of 1 MA cm^{-2}, the MTTF of an Al(Cu) interconnect line is found to be 200 hours. If n equals 2.0 and E_A equals 0.7 eV, what is the MTTF at operating conditions of 70°C and 0.1 MA cm^{-2}?

Answer Using Eq. (11.18), one can determine A, the MTTF constant for this case:

$$\text{MTTF} = AJ^{-n} \exp\left(\frac{E_A}{kT}\right)$$

$$200\ \text{hr} = A(1 \times 10^6 \text{Acm}^{-2})^{-2.0} \exp\left(\frac{0.7\ \text{eV}}{8.62 \times 10^{-5}\text{eVK}^{-1}498\text{K}}\right)$$

Solving for A gives $A = 1.66 \times 10^7\ \text{hr A}^{-1}\ \text{cm}^2$. Solving now for 70°C and 0.1 MA cm^{-2} gives

$$\text{MTTF} = 1.66 \times 10^7\ \text{hr}\ A^{-1}\text{cm}^2(0.1 \times 10^6 \text{Acm}^{-2})^{-2.0} \exp\left(\frac{0.7\ \text{eV}}{8.62 \times 10^{-5}\text{eVK}^{-1}343\text{K}}\right)$$

$$= 3.18 \times 10^7\ \text{hrs} = 3620\ \text{years}$$

This may seem like a long time (and it is), but this time is the median time (t_{50}) to failure. Generally the goal is to have the earliest failures (where the first 0.1% or so lines fail) at not less than 10–20 years of service.

11.4.3 Chemical and Structural Measurements

Many techniques are utilized to obtain chemical and structural information from metallization structures. The chemical information relates to the chemical or compositional makeup of the films, including impurity concentrations and distributions. Structural information relates to microstructural, as opposed to macrostructural or morphological, attributes. This includes lattice spacings, material density, and crystallographic information. Most of the techniques to measure the chemical and structural characteristics have been discussed in earlier chapters, and we will only briefly discuss them here.

For compositional analysis of thin films, many techniques are available. In Auger Electron Spectroscopy (AES, see Chapter 4), elemental identification and mapping are done by generating Auger electrons by a focused electron beam and then analyzing the energy of the Auger electrons. Combining this with simultaneous sputtering of the surface leads to Auger depth profiling. A typical application of Auger depth profiling would be to investigate interdiffusion of elements between different interconnect layers.

Rutherford Backscattering Spectrometry (RBS, see Chapter 4), in which the yield and energy of backscattered helium ions are measured, gives compositional information without the need of standards. Forward scattering can be used for the measurement of hydrogen in thin films as a function of depth, which other techniques have difficulty measuring. In Fourier Transform Infrared Spectroscopy (FTIR, see Chapter 3), films are radiated with infrared radiation, exciting the molecular bonds. Measurement of the absorption spectrum leads to chemical bond identification in thin films. For example, one can measure Si-OH bonds in dielectric layers, determine relative water concentrations in SOG films, or measure B and P levels in BPSG. Electron Spectroscopy for Chemical Analysis (ESCA), also known as X-ray Photoelectron Spectroscopy (XPS,

see Chapter 4), is used to identify surface constituents and to measure bond energies by irradiating a surface with a monoenergetic X-ray beam and analyzing the photoemitted electrons. Total reflection X-Ray Fluorescence (TXRF, see Chapter 4) is used to measure contaminant levels on surfaces. This technique utilizes a low-divergence X-ray beam impinging at a glancing angle on a surface. Fluorescence X-rays are emitted by atoms on the surface and detected. Each element emits X-ray photons of a characteristic energy, allowing elemental identification of impurities. An example of its use is the measurement of contaminants left on the surface after a via etch. The most common technique to measure impurity atoms in thin films, as well as in the silicon substrate, is by Secondary Ion Mass Spectrometry (SIMS, see Chapter 4). By mass analysis of sputtered (secondary) ions as a result of bombardment with a focused beam of oxygen or cesium primary ions, very accurate measurement of low levels of impurity atoms can be obtained as a function of depth. Measuring the redistribution of dopant atoms into and out of silicides is an example of SIMS use in interconnect structures.

X-Ray Diffraction (XRD) is often used for structural measurements. By analyzing the constructive and destructive interference of X-rays scattered from thin films, one can measure lattice parameters and make phase identifications. One can also measure stress (as discussed in the next section), and measure preferred orientation, or texture. Crystallographic information can also be determined by utilizing the diffraction of backscattered electrons, rather than X-rays. In Electron Back-Scattering Diffraction (EBSD), a small electron beam is incident on the surface at a glancing angle of less than 45°. A phosphor screen is located near the specimen to collect the backscattered electrons. Diffraction patterns are generated on the screen, allowing one to determine the crystallographic orientation of the region where the beam is incident. In this manner, the crystallographic orientation of individual grains of an Al interconnect line can be determined. By using computer-automated diffraction pattern analysis techniques, one can map out the crystallography of entire interconnect lines, leading to analysis and understanding of reliability problems such as electromigration.

The dielectric properties, particularly the dielectric constants of films, are measured using optical techniques. Ellipsometry, discussed in Chapter 6, is primarily used for this purpose. The dielectric constant of a layer is not only useful to determine the electrical properties of a dielectric layer in an interconnect structure, but it also gives information regarding its structure and density. However other properties, such as composition and stress, also affect these parameters, and it is often difficult to separate them all out.

A rather crude, yet simple, measure of a film's density can be obtained by measuring the etch rate of the film. Buffered HF can be used to measure the relative etch rate of SiO_2 films, for example. Films that are less dense have higher etch rates. However, the film's stoichiometry, structure, and other impurities such as hydrogen can also affect the etch rate so one cannot unequivocally equate etch rate with film density. But such data still can give an indication of differences in films that have different deposition conditions, are processed differently, or are in different locations on the surface of the wafer. Measuring the etch rate of PSG films can also be used for a quick determination of the phosphorus concentration since there is a good correspondence between the two properties.

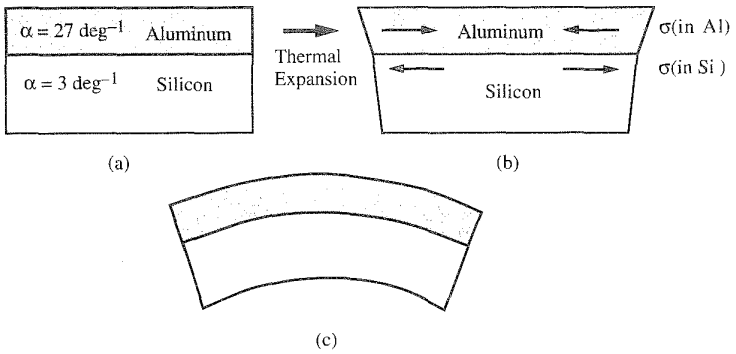


Figure 11–37 Thermal expansion differences between Al and Si, shown in (a), lead to stresses in the Al thin film and Si substrate, shown in (b). In this case, tensile stress develops on top of the Si substrate and compressive stress develops in the Al thin film. σ designates the normal, or direct, stress in each case. (c) shows the bending of the structure (exaggerated) that occurs due to these stresses.

11.4.4 Mechanical Measurements

Stresses that develop in interconnect layers and structures can greatly affect their physical integrity, leading to failures such as cracking and loss of adhesion, as well as to reliability problems such as void and hillock formation. A common source of stresses in thin films is differences in thermal expansion. If two films (or a film on a substrate) with different coefficients of thermal expansion, α, are in contact with each other, then upon heating or cooling one film will want to shrink or expand more than the other. If good adhesion between the films is maintained, then stresses (normal, or direct, stresses) will develop in the two films. Figure 11–37 illustrates this for an Al film on a Si substrate. Al has a much higher coefficient of thermal expansion than Si, as seen in Table 11–4. Upon heating, the Al film will want to expand more than the Si. If good interface contact is maintained, then the Al will pull out on the Si, inducing a tensile stress on the top of the Si. Likewise, the Si will pull in on the Al, causing a compressive stress on the Al film. (If another thin film such as SiO_2 is between an Al film and a thick Si substrate, it is the difference in thermal expansion coefficients between the Al and the much thicker Si that primarily determines the thermal stress in the Al.) To relieve this stress, plastic flow of the Al can occur. Hillock formation in Al thin films, illustrated earlier in Figure 11–11, can result from this.

Another type of stress in thin films is called "intrinsic" stress. This is the stress present after deposition, at the deposition temperature. This can be due to differences in lattice constants between the film and the layer it is deposited on, energetic ion bombardment during deposition, or gas or precipitation inclusion in the film. If the deposition is done at an elevated temperature, the stress measured afterwards at room temperature would include contributions from both the intrinsic stress and the stress due to differential thermal expansion.

A common way to measure stress in a thin film is by measuring how much the substrate bends after the film is deposited or processed. In the absence of any movement restraints, the structure in Figure 11–37(b) will in fact bend in order to balance the force moments caused by the stresses. This is shown in Figure 11–37(c). For the top film in compressive stress, as is the case in this example, the entire structure will bend with the edges downward relative to the middle. The interface stays constant in length, but the top film wants to expand and the bottom film wants to shrink, leading to the concave downward bending. The amount of bending is dependent on the internal mechanical properties of the film and substrate (specifically Young's modulus and Poisson's ratio), the thicknesses, and the stresses in the film and substrate. By measuring the amount of bending, and knowing the film thickness and mechanical constants, one can extract the stress in the film. If we assume that the film thickness is much smaller than the thickness of the substrate and that the film is in an isotropic biaxial stress state (i.e., $\sigma_z = 0, \sigma_x = \sigma_y$), then the biaxial stress in the film, $\sigma_f (= \sigma_x = \sigma_y)$, approximately equals

$$\sigma_f = \frac{1}{6R} \frac{Y_s x_s^2}{(1 - \upsilon_S) x_f} \tag{11.19}$$

Here R is the radius of curvature (i.e., a measure of the bending), Y_s is Young's modulus of the substrate, υ_s is Poisson's ratio for the substrate, x_s is the substrate thickness, and x_f is the film thickness. In practice the curvature is measured before and after the film is deposited (or before and after the film is processed in some way), since the wafer might not be flat to start with. Thus $1/R$ would be given by $(1/R_{\text{final}} - 1/R_{\text{initial}})$. The difference in curvature is then related to the stress induced by depositing or processing the film.

Example What is the stress in a 0.5-μm Al film on a 600-μm-thick (100) Si wafer if it induces a radius of curvature of 2.7×10^4 cm? (Assume $1/R_{\text{initial}} = 0$.)

Answer $Y_s/(1-\upsilon_s)$ is the bulk modulus of the substrate and given in Table 11–4 as 1.805×10^5 MPa, which equals $1.805 \times 10^5 \times 10^7$ dynes cm^{-2}. Substituting numbers into Eq. (11.19), we have

$$\sigma_f = \frac{1}{6R} \frac{Y_s x_s^2}{(1 - \upsilon_S) x_f}$$

$$= \frac{1}{6(2.7 \times 10^4 \text{cm})} \frac{1.805 \times 10^{12} \text{dynes cm}^{-2} (600 \times 10^{-4} \text{ cm})^2}{0.5 \times 10^{-4} \text{cm}}$$

$$= 8.0 \times 10^8 \text{ dynes cm}^2$$

Another technique used to measure stress in a thin film is with X-Ray Diffraction (XRD). The stresses in the film induce a change in the lattice parameter of the film, which can be measured by XRD techniques. For a crystalline or polycrystalline film with an isotropic biaxial stress, the film stress σ_f is

$$\sigma_f = -\frac{Y_f}{2\upsilon_f}\left(\frac{a_o - a_s}{a_S}\right) \tag{11.20}$$

where a_s is the lattice constant in the stressed film, a_o is the lattice constant in the unstressed film, Y_f is Young's modulus in the film, and υ_f is Poisson's ratio for the film. Conventional XRD measures the average stress over large areas. Blanket, unpatterned films are used in this case. However, recent advances in high brightness XRD, using synchotron X-ray sources, have allowed "microdiffraction stress analysis." With this technique the local stresses in patterned films can be measured.

X-ray diffraction methods are not well suited for amorphous films since they do not have well-defined lattice spacings. Instead, in addition to using wafer bending techniques, stresses in amorphous films are often measured by FTIR. In this technique bond angles are determined from the light absorption measured at different wave numbers. The bond angles are assumed to be stress dependent, and stress levels can thus be determined from the change in bond angles.

Figure 11–38 shows a measured stress-temperature curve, using X-ray diffraction, for an Al film on a SiO_2 coated Si substrate. The stress in the Al starts out tensile (by convention, a tensile stress is positive, a compressive stress is negative) at about 50 MPa. As the sample is heated, the difference in thermal expansion results in the stress in the Al becoming more compressive. The stress in the film due to differences in thermal expansion coefficients is

$$\sigma_f = (\alpha_S - \alpha_f)\Delta T \frac{Y_f}{1 - \upsilon_f} \tag{11.21}$$

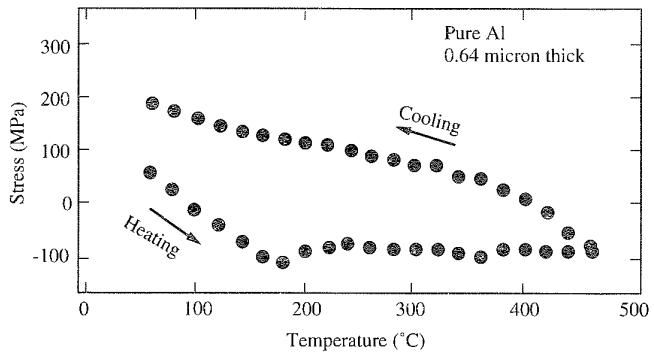


Figure 11–38 Stress-temperature curve of 0.64-μm-thick Al film on SiO_2. The lower points are for heating, while the upper points are for cooling. (After [11.25].)

where ΔT is positive upon heating and negative upon cooling. Thus the slope of the stress-temperature curve initially is approximately $\Delta\alpha\ Y_f/(1-\upsilon_f)$. At about 200°C, the elastic behavior turns to plastic deformation as processes occur that relax the stress. This might involve grain boundary or dislocation movement as well as hillock formation. Instead of following the elastic curve, the stress remains fairly constant up to 450°C. Upon cooling, elastic deformation again occurs, with about the same slope in the curve as at the beginning. Then plastic deformation again occurs when the temperature is lowered below about 350°C. Film hardening is evident as the sample is cooled from about 200°C back to room temperature. Note that the final stress level in this example is significantly higher than the starting stress level because of the plastic deformation that takes place.

Another important thin film property is adhesion. Measurement of this property in a quantitative way is difficult. A simple, very crude technique to measure film adhesion is the "scotch tape method." A thin film is deposited on a substrate or on another thin film. Adhesive tape is attached to the top and then pulled off. The amount of film stripped off is a qualitative measure of the film's adhesion to the layer below. More quantitative tests of this type involve attaching a metal pin to the film with epoxy, then using a calibrated mechanical force to pull the pin at a specific rate and/or force. Another technique is to deposit an additional thin film on the top of the thin film/substrate couple. With prior knowledge of the top film's stress as a function of thickness, one can increase the stress applied to the lower thin film by increasing the thickness of the top film until the lower films loses adhesion to the substrate. Using stress simulation methods, one can determine the stress, or force, needed for the lower film to lose adhesion, thus quantifying the adhesiveness to some degree. This technique is often close in practice to the scenario in which thin films do lose their adhesion (i.e., when other stressed films are put on top).

11.5 Models and Simulation

Back-end processing involves the deposition and patterning of insulators and conductors. Because of this, the simulation tools described in Chapters 9 and 10 for deposition and etching are very useful in simulating back-end processes. In fact, many of the examples given in those chapters involved back-end processing issues. Thus tools like SPEEDIE [11.26], ATHENA [11.27], and TAURUS [11.28] are widely used for understanding and simulating back-end structures and process flows. Since deposition and etching were covered in those earlier chapters, we will focus here on a few additional simulation tools and simulation problems that are particularly useful in back-end structures. The first section deals with silicide formation which, as we have seen, is widely used to strap junctions and polysilicon lines to reduce their sheet resistances. The second section discusses models for CMP which is now a key technology for back-end structures. The next two sections deal with polycrystalline materials and describe modeling approaches for grain growth and diffusion in these materials. Such models are useful in polysilicon, but also more generally in other back-end materials used for interconnects, almost all of which are polycrystalline. We will find that the single-crystal

models developed in earlier chapters for diffusion can be extended to polycrystalline materials, although because the properties of these materials vary with processing conditions and even with position on a given chip, the models are not as predictive as they are in the single-crystal case. Finally, we cover a problem unique to back-end structures. This is the electromigration failure mode which can cause catastrophic failure of interconnects, particularly aluminum lines. Sophisticated models have been developed to predict these types of failures and some simulation capability exists.

11.5.1 Silicide Formation

In Chapter 6 we saw how the thermal oxidation of silicon is modeled using the linear-parabolic Deal-Grove model. Oxygen diffuses through the growing SiO_2 layer and reacts with Si at the Si-SiO_2 interface. For short times and thin oxides, the reaction between Si and O limits the oxidation process, and the oxidation rate is linear with time. For long times and thick oxides, the diffusion of oxygen through the oxide limits the oxidation process, and the overall oxidation rate is parabolic in time (that is, proportional to $\sqrt{t}$). In small structures where 2D and 3D effects are important, it is necessary to include stress effects on the oxidation process due to volume expansion in constrained regions. Stress-dependent oxidation parameters and appropriate visco-elastic flow models are needed for these cases. This same linear-parabolic model was also used for silicide oxidation in Chapter 6. Silicide formation, the reaction of a metal with the underlying Si to form the silicide, can be modeled in a similar fashion. In fact, Cea and Law [11.29, 11.30] and others have used the same models used in SUPREM-IV for the oxidation of silicon to simulate the formation of $TiSi_2$. We will describe this approach here.

The formation of $TiSi_2$ occurs by the diffusion of Si through the forming $TiSi_2$ and the reaction between Si and Ti at the $TiSi_2$-Ti interface to form new $TiSi_2$. Figure 11–39(a) illustrates this. The similarity between this process and the silicon oxidation process is obvious. Either the diffusion of the Si or the reaction at the interface can dominate the

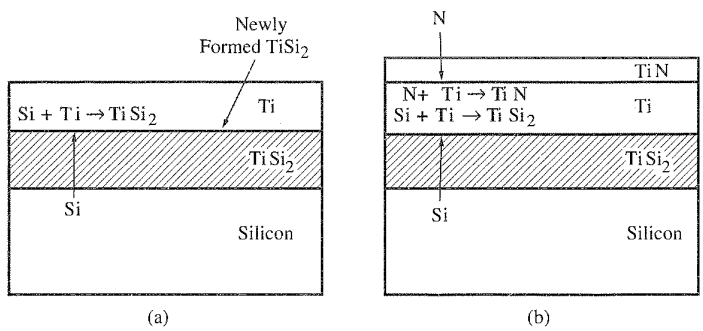


Figure 11–39 Illustration of $TiSi_2$ salicide formation: (a) $TiSi_2$ formation only, showing Si diffusion through silicide forming new silicide at the $TiSi_2$/Ti interface. (b) $TiSi_2$ formation in nitrogen ambient with simultaneous TiN formation on top. Nitrogen diffuses through the TiN layer, forming new TiN at the TiN/Ti interface.

overall formation kinetics depending on the conditions, leading to linear-parabolic behavior. For oxidation of silicon, there is a net increase in material volume, leading to a compressive stress. For $TiSi_2$ formation, one might expect the opposite to be true. This is because for every 1 nm of Ti consumed, 2.27 nm of Si are consumed producing 2.51 nm of silicide. Therefore for every 1 + 2.27 = 3.27 nm of reactants, only 2.51 nm of product is formed, leading to a net volume reduction of about 23%. Hence tensile stresses should result in the $TiSi_2$ film. However, experimentally it is found that either compressive or tensile stresses can develop in silicide films depending on the conditions of formation. For the formation from codeposited thin multilayers of Ti and Si, tensile stresses are usually produced. But for a single, thicker layer of Ti deposited on a substrate of Si, compressive stresses often arise. Other factors can affect the stress development, such as differences in thermal expansion, relaxation processes, localized compressive stress at the $Ti/TiSi_2$ interface where the new $TiSi_2$ is forming, and incorporation of argon, oxygen, or excess silicon. These may dominate over simple, net volume reduction.

The linear parabolic model for SiO_2 growth [Eq. (6.27)] is also useful in describing silicide growth since the physical processes are similar. Thus

$$\frac{x_s^2}{B} + \frac{x_s}{B/A} = t + \tau \quad \text{or} \quad x_s = \frac{A}{2}\left\{\sqrt{1 + \frac{t+\tau}{A^2/4B}} - 1\right\} \tag{11.22}$$

where x_s is the silicide thickness, t is the time, τ is the effective initial time [given by Eq. (6.26)], and B and B/A are the parabolic and linear rate constants, respectively. When the parabolic term dominates, these expressions reduce simply to

$$x_s^2 \cong B(t + \tau) \tag{11.23}$$

Physically this corresponds to the diffusion process (Si in the $TiSi_2$) being the rate-limiting step. B is primarily determined by the Si diffusivity in $TiSi_2$. When the linear term in Eq. (11.28) dominates, the growth model reduces to

$$x_s \cong \frac{B}{A}(t + \tau) \tag{11.24}$$

Physically this corresponds to the interface reaction forming new $TiSi_2$ being the rate-limiting step. B/A is largely determined by the rate of this reaction.

Cea and Law [11.29, 11.30, 11.31] have simulated the formation of $TiSi_2$ in 2D using the FLOOPS process simulator [11.32]. FLOOPS simulates the silicidation process in the same way that SUPREM models silicon oxidation. To do this in 2D, and to be able to include stress effects which are functions of position, the various mechanisms in the silicidation process are modeled step by step. The diffusion of Si from its source (either the Si substrate or polysilicon layers) through $TiSi_2$ is used to calculate the concentration of free Si at the $TiSi_2/Ti$ interface. The velocity of the $Si/TiSi_2$ interface moving downward is determined by how many Si atoms leave the interface. The reaction of Si and Ti at the $TiSi_2/Ti$ interface is then modeled. The amount of Ti consumed and $TiSi_2$ produced is a linear function of the free Si concentration and B/A, the linear rate constant. The formation of $TiSi_2$ thus occurs only when Ti is on Si, or on $TiSi_2$ which is on Si.

Table 11–7 *B* and *B/A* values for $TiSi_2$ formation [11.29]

	B ($\mu m^2\ min^{-1}$)	B/A ($\mu m\ min^{-1}$)
(100) Si	$1.85 \times 10^{11} \exp\left(\frac{-2.5\ eV}{kT}\right)$	"large"
(111) Si	"large"	$8.9 \times 10^{11} \exp\left(\frac{-2.5\ eV}{kT}\right)$

For these simulations Cea used the experimental silicide thickness versus time data of Pico and Lagally [11.33] to extract values for B and B/A for $TiSi_2$. As discussed in Chapter 6 (Figure 6–17), plotting x_s versus $(t + \tau)/x_s$ gives $-A$ as the y-intercept and B as the slope. By this method, B, the parabolic rate constant, and B/A, the linear rate constant, can be easily extracted from experimental thickness-time data. Cea's values for B and B/A are given in Table 11–7. For (100) Si, B/A is "large" meaning that the Si/Ti reaction is very fast and the formation is limited by the diffusion of Si, as governed by the parabolic rate constant, B. Therefore, any sufficiently large value of B/A will work (and will have no effect on the outcome). Alternately, for (111) Si the formation is limited by the Si/Ti reaction, and the parabolic rate constant is "large."

Example $TiSi_2$ is formed by depositing 55 nm of Ti on a (100) Si substrate and annealing in an argon atmosphere. If the anneal is done long enough so that all the Ti is reacted, how thick is the $TiSi_2$ that is formed, and how much Si is consumed? Repeat the problem if the anneal is done at 700°C for 30 sec using Cea's values in Table 11–7 for B and B/A.

Answer When the annealing is long enough so that all the Ti is reacted, 2.51 nm of $TiSi_2$ is formed for each 1.00 nm of Ti, consuming 2.27 nm of Si. (Since the anneal is done in an inert, argon ambient, no TiN is formed.) Therefore

$$\text{Amount of } TiSi_2 \text{ formed} = 55\ nm \times 2.51 = 138\ nm$$
$$\text{Amount of Si consumed} = 55\ nm \times 2.27 = 125\ nm$$

For a 700°C, 30-sec anneal, not all of the Ti may have reacted. We must therefore use the Deal-Grove-like equations. For (100) Si, Table 11–7 shows that B/A = "large" and $B = 1.85 \times 10^{11} \exp(-2.5/kT)\ \mu m^2 min^{-1}$. The "large" B/A value means that the growth is not limited by the Si surface reaction. Therefore using Eq. (11.23), and letting τ be zero, the silicide thickness x_o is given by

$$x_s = \sqrt{Bt}$$

At 700°C,

$$B = 1.85 \times 10^{11} \exp\left(- \frac{2.5 \text{ eV}}{8.63 \times 10^{-5} \text{ eVK}^{-1} 973\text{K}} \right) \mu\text{m}^2 \text{ min}^{-1} = 0.022 \ \mu\text{m}^2 \text{ min}^{-1}$$

Therefore x_s, the amount of $TiSi_2$ formed, is

$$x_s = \sqrt{0.022 \ \mu\text{m}^2 \text{ min}^{-1} * 0.5 \text{ min}} = 0.105 \ \mu\text{m} = 105 \text{ nm}$$

The amount of Ti needed to form this = 105 nm $* \frac{1}{2.51}$ = 42 nm, and, thus

Amount of Si consumed = 42 nm $\times$ 2.27 = 95.3 nm.

As with the oxidation models presented in Chapter 6, the B and B/A parameters may be affected by stress. There is some experimental evidence [11.30] that Si diffusion in $TiSi_2$ is in fact retarded by high-compressive stress levels, and this is incorporated in Cea's simulations by including a value of 1.3 nm^3 for V_D, the activation volume for the stress-dependent Si diffusivity in $TiSi_2$. This is utilized in Eq. (6.44), repeated here

$$D(\text{stress}) = D\exp\left(- \frac{PV_D}{kT} \right) \tag{11.25}$$

where P is the hydrostatic pressure (positive for compression) in the growing silicide. This will locally affect the value of B since $B = 2DC^*/N_1$ in the Deal-Grove model. In regions of high stress, or hydrostatic pressure, such as at corners, the diffusivity of the silicon through the silicide will be reduced and the silicide growth rate will be reduced.

Figure 11–40 shows a simulation in which 40 nm of Ti is deposited on top of a 2.0-μm-wide polysilicon gate structure. The left side of Figure 11–40 shows the initial structure. The structure is then annealed in an RTA for 30 sec at 650°C in an argon

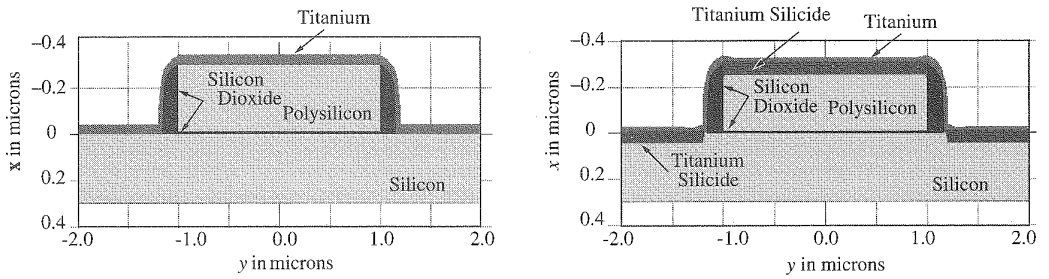


Figure 11–40 Simulation of $TiSi_2$ formation using FLOOPS [11.32] on a 2-μm-wide gate structure: left, before formation anneal step; right, after formation anneal step, 30 sec at 650°C in an argon atmosphere. $TiSi_2$ is formed everywhere the Ti is in contact or close to Si [11.31].

atmosphere, with the resulting structure on the right of Figure 11–40. The linear and parabolic rate constants used are those in Table 11–7 for (100) Si. As the $TiSi_2$ grows volumetric changes can cause stresses in the films. These are governed by the mechanical properties of each material, in particular the viscosity, Young's modulus, and Poisson's ratio. Values for the latter two were taken from the literature for the various materials (Si, oxide, polysilicon, Ti, and $TiSi_2$) in the structure. The viscosities for all the materials except $TiSi_2$ were taken to be high so that they behave elastically. The viscosity of $TiSi_2$ was set lower based on recent experimental measurements showing significant stress relaxation during the formation anneal [11.34]. As can be seen in the simulation, about 20 nm of the Ti remains, so that 20 nm reacted. Approximately 48 nm of $TiSi_2$ is formed, consuming about 43 nm of Si. This agrees reasonably well with the ratios given earlier for $TiSi_2$ formation (reaction of 1 nm of Ti consumes 2.27 nm of Si, forming 2.51 nm of $TiSi_2$). As one can see in the figure, the $TiSi_2$ forms only near where Ti is initially in contact with Si—in this case over the polysilicon gate and on the source/drain regions. The sidewall spacer prevents $TiSi_2$ formation from shorting the gate to the source and drain.

The formation of $TiSi_2$ is often done in a nitrogen ambient. As discussed previously, this causes TiN to form on top of the silicide and on top of the oxide spacer, inhibiting lateral encroachment of the $TiSi_2$, and perhaps serving as a local interconnect. The TiN forms at the expense of $TiSi_2$ and must be included in any simulation when the silicidation is done in a nitrogen ambient. The model for TiN formation is similar to $TiSi_2$ formation. In this case, nitrogen diffuses through the forming TiN, reacting with the Ti at the Ti/TiN interface, as illustrated in Figure 11–39(b). This reaction can also be modeled with the Deal-Grove linear-parabolic model. This requires simulating two reacting and moving interfaces simultaneously.

An example of $TiSi_2$ formation in a nitrogen ambient is shown in Figure 11–41. Here 40 nm of Ti is deposited on a 0.35 μm polysilicon gate structure, with adjoining source/drain regions. The structure is annealed for 30 seconds at 650°C in a nitrogen ambient. The model parameters are the same as in the previous simulation, except that

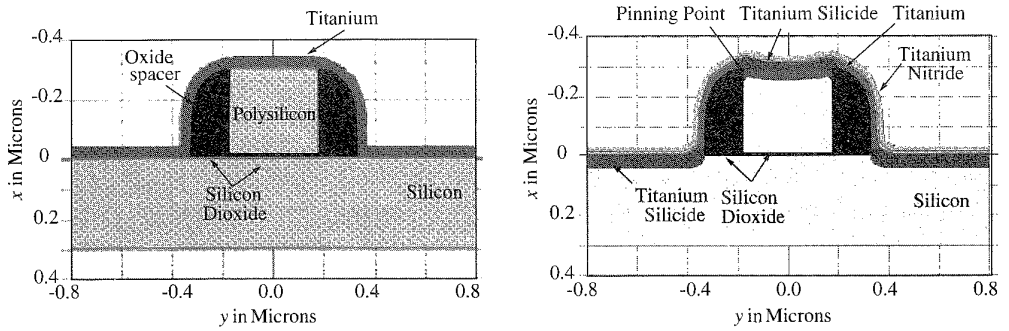


Figure 11–41 Simulation of $TiSi_2$ formation using FLOOPS [11.32] on a 0.35-μm-wide gate structure: left, before formation anneal step; right, after formation anneal step, 30 sec at 650°C in a nitrogen atmosphere. $TiSi_2$ is formed everywhere the Ti is in contact or close to Si. TiN is formed on top of the Ti layer. Bowing in the silicide on top of gate and the lateral encroachment are evident [11.31].

parameters for TiN have been added. Based on limited experimental data, the linear and parabolic rate constants for TiN growth were set to be 20% of the values for $TiSi_2$. The simulated structure after the RTA silicidation anneal in a nitrogen ambient is shown on the right in Figure 11–41. $TiSi_2$ forms as before over the polysilicon gate and Si source/drain regions, but in addition TiN forms over the entire structure. Not all of the Ti over the silicon is reacted, with about 7.5 nm remaining. Approximately 32.5 nm of Ti and 44 nm of Si are consumed, producing approximately 50 nm of $TiSi_2$ and 27 nm of TiN over the silicide. In the simulated structure, some lateral encroachment of the silicide over the oxide spacers is evident, as well as downward bowing of the silicide over the gate. Both of these features are characteristic of experimental $TiSi_2$ formation and are more pronounced in small-scale structures. The lateral encroachment is due to the Si diffusion through the silicide and over the oxide to react with the Ti there. The continued formation of TiN over the oxide spacer would block the lateral growth of $TiSi_2$ and prevent it from encroaching further. In addition, nitrogen is believed to stuff the grain boundaries of any unreacted Ti, slowing Si diffusion and helping to retard lateral growth—however this is not explicitly modeled here.

The bowing is believed to be due to the mechanical pinning of the bottom of the silicide at each end to the top of each oxide spacer [11.35, 11.36]. This pinning point is shown in Figure 11–41. As the $TiSi_2$ is formed, Si leaves the polysilicon at the $TiSi_2$/polysilicon interface, diffusing up to the $TiSi_2$/Ti interface where it forms more $TiSi_2$. The $TiSi_2$ then moves down to fill the space left from the vacant Si. If pinning of the silicide occurs at the top of the oxide spacers, consumption of Si to form the silicide can only occur in the middle of the gate and not the ends. All this leads to a bowed downward structure.

A final simulation example is shown in Figure 11–42. In this case the TSUPREM-IV simulation tool was used to simulate the formation of $TiSi_2$. 25 nm of Ti was deposited on a MOSFET structure similar to those in the previous examples. The oxide spacer was overetched prior to Ti deposition in this case. A 1.3 min anneal at 650°C formed 62.5 nm of $TiSi_2$. Finally, the remaining Ti on the sidewall spacers was etched off, resulting in the structure in Figure 11–42. The basic physical models are similar to those described above, producing the expected amount of silicide over the exposed silicon regions. .

These models and simulations should be applicable to other silicide formation processes, as well as to any reaction between adjacent layers which produces a new

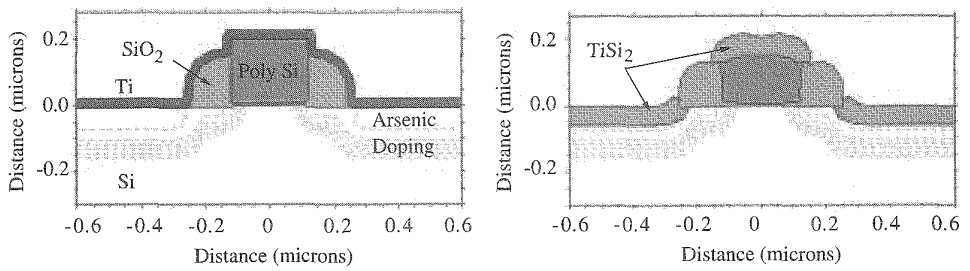


Figure 11–42 TSUPREM-IV [11.28] simulation of the formation of $TiSi_2$ on the gate and source-drain regions of a MOSFET: left, before silicidation anneal; right, after silicidation anneal and removal of unreacted Ti.

phase between them. These models can also be used for the silicides which form by the metal, rather than the Si, diffusing through the silicide.

11.5.2 Chemical-Mechanical Polishing

We have seen many examples throughout this chapter of how Chemical-Mechanical Polishing (CMP) is used to planarize films, utilizing a combination of chemical and mechanical polishing or etching. Models and simulation tools have been developed for this complex process. They are, however, mostly phenomenological in nature. The starting point for modeling polishing is the Preston equation [11.37]:

$$P = K_p p v \tag{11.26}$$

where P is the polish or etch rate. K_p is a constant (with units of pressure) and is a function of the mechanical and chemical properties of the film and pads and the composition and properties of the polishing slurry, p is the applied pressure between the wafer and the pad, and v is the relative velocity between the pad and wafer. The applied pressure, p, depends on the geometry or topography of the surface. Since the pads used are not perfectly rigid, they conform to the surface to a limited degree, but of course not exactly. As a result the pressure at each point varies, with protruding points receiving high pressure, and sunken or shadowed points receiving little or no pressure. Because of the semi-rigidity of the pads, the resulting profiles depend on both the stiffness of the pad as well as on the density and spacing of elevated features on the surface.

Most models for CMP attempt to determine the relative pressure at each point and then calculate the relative removal rate at each point assuming that it is linearly proportional to the local pressure, as given by the Preston equation. One such model was developed by Warnock [11.38], and is utilized in the commercially available process simulator ATHENA [11.27]. Here the local polish rate, P_i, is given by

$$P_i \propto \frac{K_i A_i}{S_i} \tag{11.27}$$

The three terms in this expression take into account the increases or decreases in pressure depending on the local topography as well as the rigidity of the polishing pad and wafer. K_i is called the kinetic factor at point i and is a geometrical factor that takes into account sloped surfaces by calculating an effective vertical component of the horizontal polish rate. A_i is an accelerating factor for points which protrude above neighboring points, while S_i is a shading factor for points that are lower than neighboring points so that the pressure, and hence polishing rate, is decreased. The three factors are calculated at each point and depend on the geometrical relationship of the point to its neighboring points through various mathematical-geometrical expressions. They are also dependent on the rigidity of the polishing pad. Most of the constants in these expressions are empirically determined.

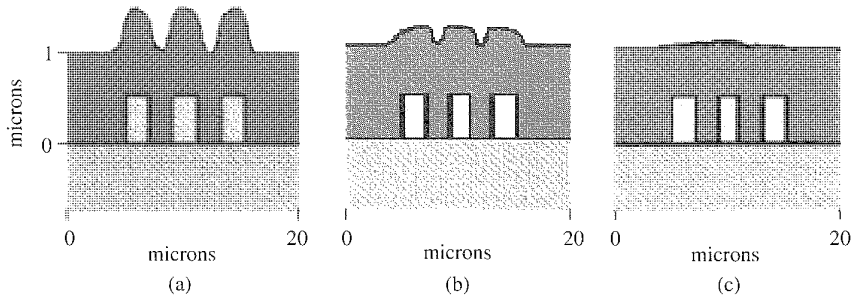


Figure 11–43 ATHENA [11.27] simulation of chemical-mechanical polishing of SiO_2 over Al lines: (a) before polishing; (b) after three minutes of polishing; (c) after six minutes of polishing. Nominal model parameter values are used, including *soft* = 25, *height.fac* = 0.02, *length.fac* = 4.2, and *kinetic.fac* = 10.

In Figure 11–43 an ATHENA simulation is illustrated. SiO_2 is deposited over three closely spaced Al lines, as shown in Figure 11–43(a). CMP is then simulated for three minutes and six minutes polishing time. The parameters used are given in the figure caption.

The *soft* parameter is the normalized polishing rate, equivalent to the proportionality constant implicit in Eq. (11.27). The *height.fac* parameter is the vertical deformation scale and measures how much the polishing pad will deform with respect to the height of the feature. The *length.fac* parameter is the horizontal deformation scale and describes the distance at which shadowing will be felt by a tall feature. Both are measures of the pad's flexibility and are used in calculating A_i and S_i at each point. The *kinetic.fac* parameter is just the proportionality factor for K_i. In Figure 11–43 we have used a nominal value for the height factor of 0.02 μm, where parts (b) and (c) show the profiles after three and six minutes, respectively. One can see that very good global planarization is achieved. In Figure 11–44, we repeat the simulations but use a smaller value for the height factor of 0.005 μm. This corresponds to a more rigid pad. One can see that even less rounding occurs, resulting in virtually perfect planarization here.

When the polishing rate is somewhat different for two different materials and the pad is not perfectly rigid, a phenomenon called "dishing" can occur. This would happen, for example, when CMP is done on metal for via formation in an oxide dielectric layer (see Figure 11–19). An overpolish, like an overetch, is done to ensure that all the metal is removed from on top of the oxide layer. But metals generally polish faster than oxide in the metal polishing slurry. (This would be worse for Al or Cu compared to W, since the latter is harder and etches slower.) Due to the semi-rigidity of the pad and the difference in polishing rates, this causes the metal in some places to be removed below the top of the oxide layer. The shape of the resulting metal surface is curved, or "dished." This is demonstrated in a simulation, shown in Figure 11–45. Having less selectivity between materials would reduce this, but would make it harder to stop the polishing without removing too much oxide. Having a more rigid pad would also help but could cause unequal polishing due to any bowing or thickness nonuniformities in the

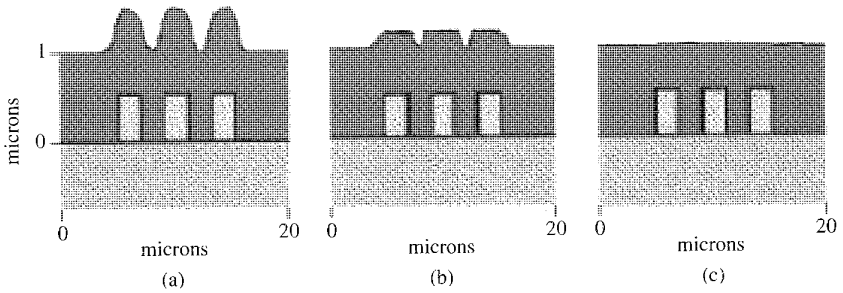


Figure 11–44 ATHENA [11.27] simulation of chemical-mechanical polishing of SiO_2 over Al lines: (a) before polishing; (b) after three minutes of polishing; (c) after six minutes of polishing. This time the *height.factor* = 0.005 μm, corresponding to a more rigid pad than in Figure 11–43.

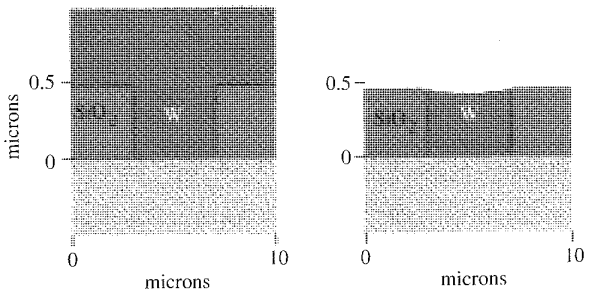


Figure 11–45 ATHENA [11.27] simulation of chemical-mechanical polishing of a tungsten via structure: Left, before polishing; right, after polishing. Due to the faster polishing of tungsten compared to silicon dioxide and the semirigid pad, "dishing" of the tungsten plug can result.

silicon substrate—the main reason that a nonrigid pad is used. Dishing in general is worse for wider structures, such as contact pads, but can still affect smaller features as well. Dishing effects can also result when an initially non-planar surface is polished and some "residue" of the original surface remains after polishing. Again this is a result of using semi-rigid pads.

One can also see in Figure 11–45 that the oxide layer is also slightly polished, especially at the edges. This is called "erosion" and refers to any excess removal of material during the over-polishing.

11.5.3 Reflow

Reflow is the smoothing of the topography of a layer during a thermal treatment. It is usually done intentionally, with the flow of the material causing sharp features to be flattened or smoothed, and valleys and trenches to be filled in. Reflow of PSG or BPSG

layers is commonly done to smooth and planarize the first-level dielectric level. Reflow of metal layers such as Al, though not commonly done today, could be used to planarize and fill contacts and vias and is being investigated for future use. We also saw an example of photoresist reflow in Chapter 5 (Figure 5–50), where the purpose was similar. In that case reflow was used to minimize sharp edge features in the resist pattern associated with standing wave effects during the exposure process.

The reason reflow occurs is because it minimizes the total energy of the system. In this case, the surface energy of the structure is reduced. This occurs by minimizing the curvature of surface features. Surface energy arises because atoms at the surface are not completely surrounded by other atoms as they are below the surface, leading to either unsatisfied bonds with unpaired electrons (essentially broken bonds) or distorted bonds at the surface. (The unsatisfied bonds will in fact often stretch or distort in order to form some kind of bond to the neighboring atoms, but the energy is still higher than in the bulk due to the nonequilibrium interatomic spacings and angles.) A topography that is smooth has less surface area, and hence fewer unsatisfied or distorted bonds, than a topography that is rough and has steps, valleys, or sharp features. If the structure is heated and atoms have a chance to move around, they will preferentially move so that the surface becomes smoother and the surface energy is reduced. Another way to say this is that the atoms will move to regions of lower chemical potential, decreasing the total energy of the system.

W.W. Mullins [11.39, 11.40], who did much of the pioneering work on modeling reflow (and whose derivations we follow below), has mentioned four mechanisms by which atoms move at high temperatures to reduce the surface energy. These are surface diffusion, volume diffusion, evaporation/condensation, and viscous flow. Figure 11–46 illustrates these mechanisms. In the surface diffusion mechanism, atoms diffuse along the surface of the material to regions of lower chemical potential. New atoms are exposed, which can then diffuse along the surface. In the volume diffusion mechanism, the atoms move to reduce the surface energy by diffusing through the volume of the material. In the evaporation/condensation mechanism, the atoms preferentially evaporate from areas of high energy and preferentially condense in areas of low energy. For amorphous materials, viscous flow can also occur. This mechanism occurs when groups of atoms or

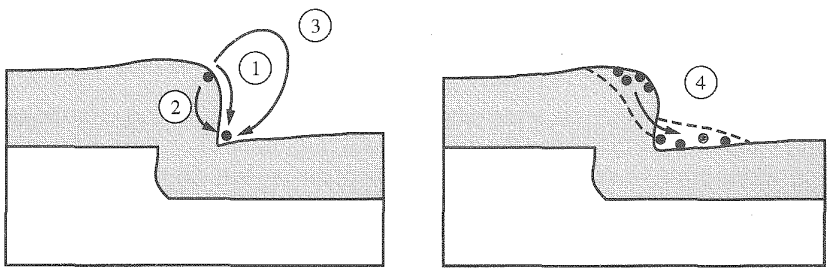


Figure 11–46 Illustration of high-temperature reflow mechanisms: (1) surface diffusion; (2) volume diffusion; (3) evaporation/condensation; and (4) viscous flow.

molecules in the solid slide past one another, breaking and reforming interatomic bonds, similar to motion in a liquid.

In all four cases the driving force is to reduce the total surface energy, and the forces involved in each are all proportional to γ_s, the surface energy per unit area (or surface tension). Which mechanism dominates for a given situation depends on the material, the temperature, and the topography. Mullins points out that for metals often the surface diffusion mechanism dominates near a metal's melting point for small feature sizes. Brain [11.41, 11.42], whose simulations we present below, argues that for thin film Cu reflow in integrated circuit structures, surface diffusion dominates over volume diffusion for normal reflow temperatures. For silicon dioxide reflow, viscous models are often used. Here we will concentrate mostly on surface diffusion models and show the results of simulating Cu reflow. At the end of this section we will show a simulation example of oxide reflow via viscous flow.

As with most models describing the motion of atoms or particles, we start with

$$v = M \cdot \text{Force} \tag{11.28}$$

in which v, the average velocity of a particle, equals M, the mobility of the particle, times the force acting on the particle. For the case of surface diffusion, v is the average velocity of a surface atom's diffusion along the surface. The force is determined from

$$\text{Force} = -\frac{\partial \mu}{\partial s} \tag{11.29}$$

where μ is the chemical potential, in this case the increase in chemical potential per atom that is transferred from a point of zero curvature to a point of curvature K on the surface. The unit length that it travels along the surface is s. The chemical potential is related to the curvature, K, and the per-area surface energy, γ_s, by

$$\mu(K) = K\gamma_s\Omega \tag{11.30}$$

where Ω is the atomic volume of the atom. Therefore

$$\text{Force} = -\frac{\partial \mu}{\partial s} = -\gamma_s \Omega \frac{\partial K}{\partial s} \tag{11.31}$$

This means that the higher that γ_s is for the material, the larger is the force acting on the atom. Also, the larger the gradient in K, the more force there is to move an atom from one point to another. The negative sign means that the force pushes the atom from an area of high K to an area of low K. As mentioned above, the curvature, K, is a measure of how "unsmooth" or how rounded the surface is at a point. In two dimensions, K is equal to the inverse of the radius of curvature, R, at that point

$$K = \frac{1}{R} \tag{11.32}$$

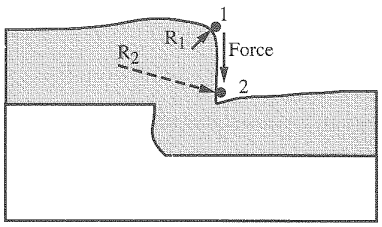


Figure 11–47 Illustration of differences in radii of curvature, R_1 and R_2 at points 1 and 2, respectively, causing a movement of an atom to a point of lower chemical potential (lower surface energy).

In three dimensions, K is given by

$$K = \frac{1}{R_x} + \frac{1}{R_y} \tag{11.33}$$

where R_x and R_y are radii of curvature in any two orthogonal directions. Figure 11–47 shows how the force acting on an atom due to surface energy minimization is in the direction away from a point of higher curvature to a point of lower curvature. R_1 is less than R_2 (so $K_1 > K_2$, and $\mu_1 > \mu_2$), so that the atom, given enough thermal energy, would move from point 1 toward point 2, causing a flattening of the topography. (In fact, right at the cusp the curvature is negative, as is the case for any concave surface, and the driving force to that point is even higher).

On an atomistic level, one can imagine that an atom sitting on the surface at a point of high curvature, such as point 1 in Figure 11–47, will be surrounded by and bonded to fewer atoms at equilibrium interatomic spacings than one at a point of low curvature, such as point 2 or even right in the cusp. Thus the atom will lower the total energy of the system by being at point 2. We will see a similar effect when we look at grain growth and see why grains grow "against" the curvature of the grain boundary to maximize the number of bonds and lower the total energy.

Substituting Eq. (11.31) into (11.28), and given that the mobility, M, is D_s/kT, where D_s is the surface diffusivity of the atoms, yields

$$v = -\frac{D_s}{kT}\gamma_s \Omega \frac{\partial K}{\partial s} \tag{11.34}$$

The surface current or surface flux of atoms, F_s (in atoms cm^{-1} sec^{-1}), equals

$$F_s = v \cdot \upsilon \tag{11.35}$$

where υ is the number of atoms per unit area, or the surface density. Therefore

$$F_s = -\frac{D_s}{kT}\gamma_s \Omega \upsilon \frac{\partial K}{\partial s} \tag{11.36}$$

One can see from this expression that a surface flux of atoms will occur along a surface down the curvature gradient, similar to diffusion down a concentration gradient.

Therefore atoms will move from a position of high curvature to low curvature, or from positive curvature to negative curvature.

Example Calculate the surface flux (F_s, in atoms cm^{-1} sec^{-1}) due to a surface curvature difference between points *a* and *b* for a copper film. The radius of curvature at point *a* is 0.5 μm, and the radius of curvature at point *b* is infinity (i.e., a flat surface). The two points are separated by a distance of 0.5 μm along the surface. Use the parameter values given in Table 11–8, with $T = 800K$. State which direction the atoms are moving (*a* to *b*, or *b* to *a*).

Answer The parameter values needed for Eq. (11.43) are

$T = 800K$

$D_s = 0.07 \exp(-0.82/kT)\ cm^2/sec = 0.07 \exp(-0.82/8.62 \times 10^{-5}\ eV/K / 800K)$
$= 4.80 \times 10^{-7}\ cm^2\ sec^{-1}$

$k = 1.381 \times 10^{-23}\ J/K = 1.381 \times 10^{-23}\ J/K \times 10^7\ ergs/J = 1.381 \times 10^{-16}\ ergs/K$

$\gamma_S = 1800\ ergs\ cm^{-2}; \Omega = 1.2 \times 10^{-23}\ cm^3; \upsilon = 1.43 \times 10^{15}\ atoms\ cm^{-2}$

$K_A = 1/R_A = 1/(0.5 \times 10^{-4}\ cm) = 2 \times 10^4\ cm^{-1}$

$K_B = 1/R_B = 1/\infty = 0\ cm^{-1}$

$s = 0.5\ \mu m = 0.5 \times 10^{-4} cm$

$$\frac{\partial K}{\partial s} = \frac{0 - 2 \times 10^4\ cm^{-1}}{0.5 \times 10^{-4}\ cm} = -4 \times 10^8\ cm^{-2}$$

Therefore

$$F_s\ (\text{from } a \text{ to } b) = -\frac{D_s}{kT}\gamma_s \Omega \upsilon \frac{\partial K}{\partial s}$$

$$= -\frac{4.80 \times 10^{-7}\ cm^2\ sec^{-1}}{1.38 \times 10^{-16}\ ergs\ K^{-1}\ 800K} \cdot (1800\ ergs\ cm^{-2})(1.2 \times 10^{-23}\ cm^3)(1.43 \times 10^{15}\ atom\ cm^{-2})(-4 \times 10^8\ cm^{-2})$$

$F_S = 5.4 \times 10^{10}$ atoms cm^{-1} sec^{-1} and the direction of the flux is from a (where the curvature is higher) to b (where the curvature is lower).

The negative of the change in surface current per change in *s*, $-\partial F_s/\partial s$, gives the increase is number of atoms per unit area per unit time for a given location. Multiplying this by the atomic volume, Ω, gives r_n, the speed of movement of the surface element along its normal (the "normal velocity"). Thus

$$r_n = \frac{D_s}{kT}\gamma_s \Omega^2 \upsilon \frac{\partial^2 K}{\partial s^2} \tag{11.37}$$

Eq. (11.37) becomes the governing equation for the surface movement due to surface diffusion controlled reflow. This is usually abbreviated to

$$r_n = B\frac{\partial^2 K}{\partial s^2} \tag{11.38}$$

where

$$B = \frac{D_s}{kT}\gamma_s \Omega^2 \upsilon \tag{11.39}$$

To model the reflow phenomenon, these equations can be applied at each point, at each time step using a finite element program. It is only by doing such simulations that one can see the net affect of the process on the topography and assess the validity of the model.

Such simulations have been done by R. Brain and colleagues [11.41, 11.42] for the reflow of Cu. As will be discussed in Section 11.6, Cu is being developed as a replacement for Al as the main interconnect metal in some applications due to its lower resistivity and better electromigration resistance. In order to improve the filling of contacts and vias with Cu, and to planarize the structure, a reflow of the Cu can be done. Like Al, this can be done either during the deposition of Cu itself, or afterwards in a separate anneal step.

Brain simulated the Cu reflow using the surface diffusion mechanism, that is, Eq. (11.39), using a finite element program to simulate the surface topography movement during an anneal. The surface diffusion mechanism was chosen after comparing relative rates for the other three mechanisms using typical literature values for the parameters. The parameter values Brain used are given in Table 11–8. The simulations of Cu reflow in trenches of various sizes are shown in Figure 11–48. The initial profile was taken from an actual as-deposited Cu film profile, with some asymmetry. For the widest trench, 1 μm wide, one can see that the trench gets filled in with the Cu during the course of the reflow anneal, and the topography becomes much smoother. For narrower trenches, hourglass shapes develop near the bottom of the trenches, with the sides thinning and Cu accumulating at the bottom. Also, Cu agglomerates at the top shoulders, producing overhangs. This illustrates how the geometry of the structure is very important for this process. For longer times, the Cu profile becomes smoother, even for the narrowest

Table 11–8 Cu reflow simulation parameters at 800K [11.42]

D_s	surface diffusivity	$0.07 \exp\left(\frac{-0.82\ \text{eV}}{kT}\right)\ \text{cm}^2\ \text{sec}^{-1}$
γ_s	surface energy per unit area	1800 ergs cm^{-2}
Ω	atomic volume	$1.2 \times 10^{-23}\ \text{cm}^3$
υ	number of atoms per unit area, or surface density	1.43×10^{15} atoms cm^{-2}

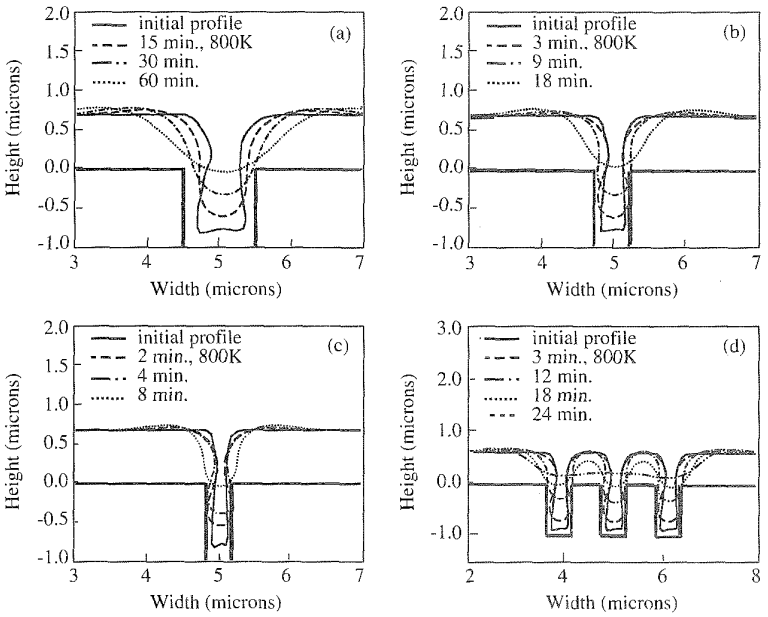


Figure 11–48 Cu reflow simulation for different trench sizes: (a) 1 × 1 μm; (b) 0.5 × 1 μm; (c) 0.33 × 1 μm; and (d) three 0.5 × 1 μm trenches spaced 0.5 μm apart. (After [11.41].)

structures simulated. For the multiple trenches, seen in Figure 11–48(d), similar behavior to the individual trenches is observed, indicating that the reflow process is more or less acting independently at each of the trenches and they do not interact.

However, experimental observations by Brain of this behavior indicate differences from the simulations shown above. For example, the extent of filling due to reflow varied along a trench, and this correlated to the position of grain boundaries. It was possible to qualitatively model this effect by placing the end of a grain boundary at various locations along the profile in a trench and including a grain boundary energy term, similar to the surface energy term. Also evident in the experimental observations were two prominent features: faceting of the Cu along crystallographic orientations during the reflow and nonsymmetric reflow behavior with agglomeration of Cu along one side wall and not the other. Brain proposed that these are due to an anisotropic surface energy. It was assumed in the initial simulations that γ_s is constant in all crystallographic directions. However, previous measurements indicate that this is not the case. The value of γ_s for Cu is not only different for the main directions (<100, 110, and 111>), but varies

significantly for directions between those. Using this more advanced model for γ_s, Brain was able to better match actual experimental results [11.42].

Other types of reflow models are also used in simulators. In some situations, a viscous flow model, similar to that used in modeling 2D oxidation is more appropriate than the surface diffusion model described above. Viscous flow models are often used for silicon dioxide reflow, although surface diffusion models are sometimes used for computational convenience and can give similar results. An example using a viscous flow model is shown in Figure 11–49. In this case the ATHENA simulator is used to simulate PSG reflow over a polysilicon line. This situation might correspond to the first-level dielectric in Figure 11–1. In this case, much higher temperatures can be used for the reflow and the PSG behaves like a viscous incompressible fluid. Reducing surface energy is again the driving force for the reflow. The relevant model parameters in this case for the glass include the per area surface energy, the viscosity, Poisson's ratio, and Young's modulus. Default parameter values were used in the simulation in Figure 11–49.

The reflowing behavior can be very dependent on the specific surface structure and chemistry, such as any native oxide present on a metal film, which will affect the surface energy term. Reflow is also dependent on the bottom interface. Different underlayers can give different reflow results. This can be taken into account by including in the equations the interfacial energy between the reflowing layer and the layer beneath, analogous to the surface energy. For polycrystalline material, grain structure and grain growth also affect reflow. These are more difficult to include in the models.

11.5.4 Grain Growth

All of the materials used for interconnects are polycrystalline. This means that they are made up of regions of single-crystal material in which all the atoms line up perfectly in crystallographic positions, with the regions separated by grain boundaries. This is shown

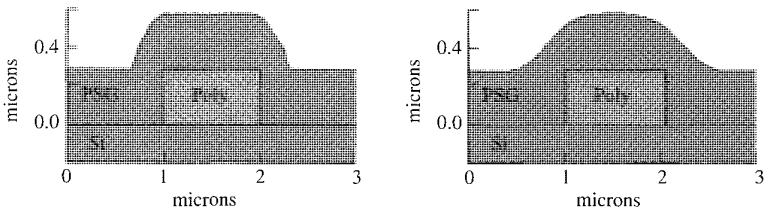


Figure 11–49 ATHENA [11.27] simulation of PSG reflow. The initial structure is on the left, the final structure on the right. The reflow heat cycle was 20 min at 950°C.

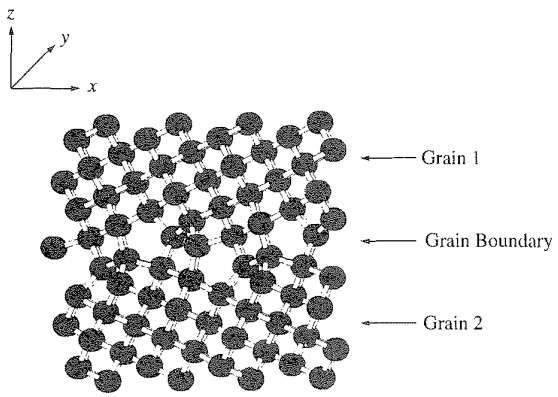


Figure 11–50 Illustration of two side-by-side grains of different crystallographic orientation. The region between them where the atoms cannot line up perfectly is called the grain boundary. This example is a low energy grain boundary in polysilicon. The structure is based on atomic-level calculations [11.43]. Reprinted with permission of the Electrochemical Society.

in Figure 11–50. Within the grains, the atoms are closely packed. Across the grain boundaries, however, the atoms cannot maintain the close packing because of the orientation differences between the grains. Therefore the grain boundary region is often less dense than the interior of the grain. In addition there may be unsatisfied or distorted bonds at the grain boundaries. Finally, charged defects may form preferentially at grain boundaries because these regions represent trapping sites for holes and electrons. These features give the grain boundary regions significantly different properties than the interior of the grains. We will describe simulations involving grains and grain boundaries in this section on grain growth, as well in the following sections on diffusion and electromigration.

When polycrystalline material is heated, grain growth usually occurs such that the average grain size increases. This is also termed grain coarsening. This occurs to minimize the energy of the system associated with grain boundaries. In fact, the lowest energy state of the material would be single-crystal without any grain boundaries at all. While this state is rarely achieved due to kinetic and metastability factors, this is the state the system is driven toward, with the average grain size getting larger. With silicides and polysilicon, especially when these films are deposited amorphous or highly doped, significant grain growth often occurs during annealing. With Al, less growth occurs because large grains are formed during its deposition due to its low melting point and high atomic mobility.

Grain growth is very similar to the process of reflow, as discussed in the previous section. The driving force is the minimization of the "surface" energy. In this case the "surface" is the surface of the grain, or the grain boundary. Since the grain boundaries have higher energies than the bulk or interior of the grains due to the unsatisfied or distorted bonds at the boundaries, the total energy of the system is decreased by reducing the relative amount of grain boundary area. This is done by increasing grain size. Locally, the driving force is the decrease in grain boundary curvature to increase bonding. Decreasing the local grain boundary curvature results in a shorter grain boundary length and a larger average grain size, leading to a lower total energy.

Thin film materials may be deposited as either polycrystalline or amorphous layers. In the latter case, single-crystal regions nucleate and grow during annealing, and eventually the grains will impinge on each other converting the material to polycrystalline material. Annealing of polycrystalline material will cause grain growth to occur to reduce the grain boundary energy. However, not all the grains can grow. For some grains to grow in a polycrystalline film, others must shrink. But overall, the average grain size will increase. We will now consider the models and equations for grain growth in polycrystalline thin films (assuming that the entire layer is composed of grains from the start) and then use simulations to help illustrate the key points. Excellent overview articles can be found in [11.44, 11.45].

We start with the same equation that we started with for reflow, relating the force on an object to its velocity, v:

$$v = M \cdot \text{Force} \tag{11.28}$$

The proportionality constant is the mobility, M. Here v refers to the velocity of the grain boundary, and F is the force on the grain boundary. Grain growth occurs by grain boundary movement or "migration" due to the force on each grain boundary.

To calculate the force on a grain boundary, we can use the fact that the force per area is equal to the pressure difference across the grain boundary between grains. The pressure difference, or force per area, is related to its curvature and surface energy [11.46]:

$$\text{Force per area} = \gamma_{gb} \cdot K \tag{11.40}$$

Here γ_{gb} is the grain boundary energy per unit area (in ergs cm^{-2}) and K is the curvature of the grain boundary at each point along it, defined such that the force is against the curvature, as we will discuss later.

The mobility, M, is usually related to the atomic diffusivity by $M = D/kT$, where D in this case would be the diffusivity of atoms across the grain boundary, or $D_{\text{gb}}^{\perp}$. However since we are dealing with force on a per unit area basis, we must multiply $D_{\text{gb}}^{\perp}/kT$ by A_{gb}, the atomic area of a grain boundary. (One could include the atomic area factor in the force term as was done in the reflow analysis or even leave it as a separate term, but we will include it at part of the mobility term to make it more consistent with treatments in the literature.) The atomic area of grain boundary, A_{gb}, can be written effectively as $\Omega/\lambda^{\perp}$ [11.47], where Ω is the atomic volume in cm^3, and $\lambda^{\perp}$ is the atomic jump distance across the grain boundary in cm (approximately an atomic diameter or lattice constant). A geometrical constant, g, dependent on the grain shape (and on the order of 0.5 to 3) may also be included in the area term. Combining $(D_{\text{gb}}^{\perp}/kT)$ and A_{gb} in the mobility term lumps terms such as $D_{\text{gb}}^{\perp}$, g, and $\lambda^{\perp}$ that are not easily determined into a single parameter. The grain boundary mobility, M_{gb}, is thus defined to be:

$$M_{\text{gb}} = \frac{D_{gb}^{\perp}}{kT} A_{\text{gb}} \tag{11.41}$$

and is in units of $\text{cm}^4\ \text{sec}^{-1}\ \text{erg}^{-1}$. One can think of the mobility in this form as being the grain boundary velocity per unit pressure [11.46], as opposed to the standard velocity

per unit force definition of mobility. The velocity of the grain boundary is equal to the product of the mobility and force terms:

$$v = M_{gb}\,\gamma_{gb}\,K \tag{11.42}$$

Example If the grain boundary mobility, M_{gb}, is 1×10^{-15} cm^4 sec^{-1} erg^{-1}, the grain boundary energy is 300 ergs cm^{-2}, and the curvature, K, is 20 μm^{-1}, what is the grain boundary velocity due to the curvature?

Answer Substituting into Eq. (11.48), we have

$$v = M_{gb}\,\gamma_{gb}K = 1 \times 10^{-15}\ \text{cm}^4\ \text{sec}^{-1}\ \text{erg}^{-1} \cdot 300\ \text{ergs cm}^{-2} \cdot 20 \times 10^4\ \text{cm}^{-1}$$

$$= 6.0 \times 10^{-8}\ \text{cm sec}^{-1} = 0.6\ \text{nm sec}^{-1}$$

The diffusivity of atoms across the grain boundary, $D_{gb}^{\perp}$, follows an Arrhenius temperature dependence. The grain boundary mobility, M_{gb}, therefore depends strongly on temperature, with the exponential temperature dependence of the $D_{gb}^{\perp}$ term dominating over the T in the denominator of Eq. (11.41). In fact, the mobility is often written in the form

$$M_{gb} = M_o \exp\left(-\frac{E_a}{kT}\right) \tag{11.43}$$

where E_a is the activation for atomic diffusion across the grain boundary. M_o would include all the other terms in Eq. (11.41) except for the temperature dependence of $D_{gb}^{\perp}$. Therefore the grain boundary mobility and velocity, and thus the grain growth, increase with temperature.

For semiconductors, the diffusivity across the grain boundary and hence the mobility will also depend on the doping level of the polycrystalline material. This is possibly due to the increase in vacancy concentration with doping [11.48]. This is especially important for polysilicon, which shows a very strong dependence of grain growth on n-type doping concentration, and leads to a $(n/n_i)^m$ dependence of the mobility. The presence of nonelectrically active impurity atoms in a film, especially if they form precipitates, can decrease the grain boundary mobility due to mechanical pinning of the grain boundary. This is called the impurity or particle drag effect and is observed when Al is alloyed with Cu, with Cu precipitates slowing down the grain growth of the Al films.

In two dimensions the curvature, K, at each point is just the inverse of the radius of curvature, R, as given in Eq. (11.32), Shortly we will look at how the curvature varies around grains in a polycrystalline material, as well as the direction of the grain boundary velocity relative to the local curvature. For now, if we simply let the average radius of curvature of all the grain boundaries in a film be equal to the average grain size or

radius, $\bar{r}$, we can derive approximate expressions for the average grain boundary velocity in a polycrystalline film as a function of grain size, and the average grain size as a function of time [11.49]. Letting $\bar{K}$ equal $1/\bar{r}$ in Eq. (11.42) gives the average velocity as

$$\bar{v} = M_{gb}\frac{\gamma_{gb}}{\bar{r}} \tag{11.44}$$

More rigorous derivations, including those based on the average curvature, those involving grain boundary energy per volume calculations, as well as statistical treatments based on grain size distributions, have all resulted in the same expression for the average velocity or growth rate [11.47, 11.50, 11.51, 11.52]. From this expression we see that as the average grain radius increases (and the grain boundary area per volume decreases), the average curvature gets smaller, leading to a lower driving force for growth and a lower average grain boundary velocity. Taking $\bar{v} = d\bar{r}/dt$ in Eq. (11.44) and integrating then leads to

$$\bar{r} = (\bar{r}_o^{\,2} + 2M_{gb}\gamma_{gb}t)^{1/2} \tag{11.45}$$

where $\bar{r}_o$ is the initial grain radius. Therefore the average grain radius, $\bar{r}$, by this model grows approximately as $t^{1/2}$, slowing down with time as the grain growth proceeds. This is quite close to experimental measurements for the average grain size during grain growth. It is also seen experimentally that the average grain size increases with anneal temperature. This is because M_{gb} increases exponentially with temperature.

Grains in a polycrystalline film cannot grow independently of one another—some grow larger while others grow smaller. This makes the analysis and modeling quite a bit more complicated. Figure 11–51 shows a large grain with many smaller grains around it. The curvature of each grain boundary is such that the boundaries are concave toward the smaller grains. This is so the grain boundaries meet at triple points with approximately equal angles of 120°. The movement of the grain boundaries occurs against the local curvature, and the larger grain grows at the expense of the smaller grain [11.53].

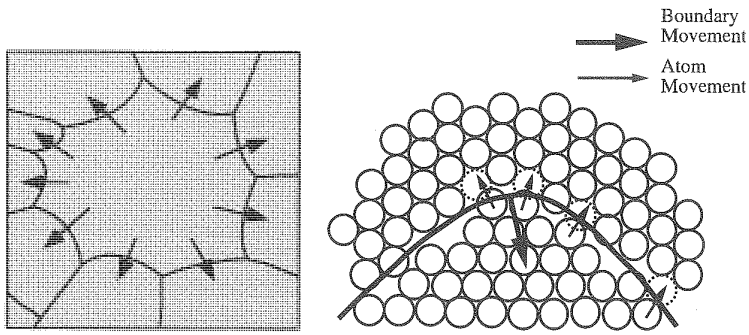


Figure 11–51 Left: Illustration of a large grain with smaller grains around it. The larger grains grow at the expense of smaller grains. Right: Atomic-level illustration of why movement of grain boundaries occurs against the curvature. (After [11.53].)

Figure 11–51 also shows what is happening on an atomistic scale and explains where the driving force originates. The atoms in the smaller grain, below, move across the grain boundary to the larger grain above. They are more stable there because, due to the curvature, they are surrounded by and bonded to more neighboring atoms at equilibrium interatomic spacings. This is very similar to the atomistic mechanism for reflow. In this way the grain boundary is shifted downward with the larger grain growing and the smaller grain shrinking. This movement reduces the curvature of the grain boundary and decreases the total grain boundary area. Eventually the smaller grains are annihilated.

M. Hillert [11.54] described this mathematically by using the idea of the critical size (used in dissolving and precipitating theory) with Eq. (11.42) for grain boundary movement to obtain

$$\frac{\partial r}{\partial t} = M_{gb}\,\gamma_{gb}\left(\frac{1}{r^{crit}} - \frac{1}{r}\right) \tag{11.46}$$

$\partial r/\partial t$ is the growth rate of an individual grain and r is the radius of the individual grain. Grains with a radius greater than r^{crit} will have a positive growth rate and grow, and grains with a radius smaller than r^{crit} will have a negative growth rate and shrink. What is r^{crit}? Because the sum of all $r\,\partial r/\partial t$ over all the grains must be zero in order to conserve the total size of the 2D system, it can be shown that r^{crit} must equal the average size or radius. Therefore Eq. (11.46) becomes:

$$\frac{\partial r}{\partial t} = M_{gb}\,\gamma_{gb}\left(\frac{1}{\bar{r}} - \frac{1}{r}\right) \tag{11.47}$$

Hillert also showed that these equations lead to an expression for the average grain size, growing as $t^{1/2}$, with an equation very similar to Eq. (11.44). The larger grains thus grow at the expense of the smaller grains, and the average grain size increases with time.

The above models and equations can be used to simulate simple grain growth, or "normal growth." This is in contrast to "abnormal" or "secondary" grain growth which we will briefly discuss later. Frost and colleagues [11.44, 11.55] have developed a two-dimensional computer simulation program for the growth of grains in a thin film polycrystalline layer after nucleation and growth to the polycrystalline state have occurred. Grain growth is modeled by calculating the local grain boundary velocity at each point. Figure 11–52 shows the simulation of normal grain growth using their program. In their algorithm, each grain boundary segment is described by a series of points. Triple points occur where three segments meet. The boundary segments move according to the local boundary velocity which is proportional to the local curvature, as given by Eq. (11.42). After each incremental movement of boundary segments, the triple points are shifted to locations at which the three boundary segment tangents meet at 120°. This sequence of boundary movement and triple point shifting is then continued throughout the simulation.

In Figure 11–52 one can clearly see the growth of grains, with the large grains growing at the expense of the smaller ones, with the latter eventually disappearing and the average grain size increasing with time. All this occurs in the simulation simply by having the grain boundary velocity proportional to the curvature at each point and by

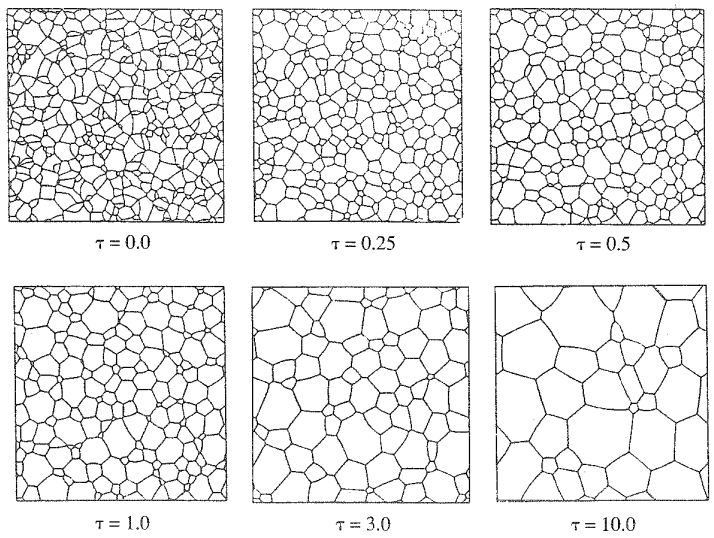


Figure 11–52 Simulation of normal grain growth. The initial structure contained approximately 400 grains of unit average area. τ is the normalized time [11.55]. Reprinted by permission of TMS.

maintaining equilibrium at the triple points. While the initial structure will affect the early growth, a steady-state regime is soon reached which is fairly independent of the initial conditions. In this steady-state regime, the average grain size is determined to be proportional to $t^{1/2}$ in these simulations, in agreement with our earlier treatment.

Walton, Frost, and Thompson [11.56] have also applied their simulations to grain growth in narrow lines, such as that encountered when annealing narrow lines of aluminum interconnects. The simulations are shown in Figure 11–53 which shows top views of a long, narrow film. The same simulation method and equations as above were used, with the additional boundary condition that all intersections between grain boundary segments and the line sidewalls must be at right angles. As the time steps progress, large grains grow again at the expense of smaller grains. Some of these traverse the entire line width, forming stable, rectangular "bamboo" segments. At intermediate times, one can see clusters of grains between the bamboo segments. These are relatively stable and are important in electromigration, as we shall see later. Eventually, though, they disappear and the final structure consists of rectangular bamboo grains of average length approximately 2.3 times the line width.

These simulations are two-dimensional and do not model what is happening in the z, or vertical, direction. This is often a reasonable approximation since many thin films have columnar structures. That is, the grains traverse the entire height of the film and the grain boundaries are constant in the z-direction. And those films that are not columnar as deposited often become columnar or near columnar during grain growth as the average grain size exceeds the film thickness. However, interfaces and surfaces—the boundaries in the third dimension—are still important in columnar structures.

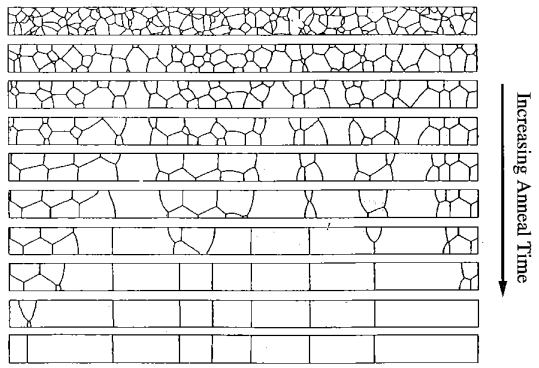


Figure 11–53 Simulation of grain growth within a strip. The grain growth leads to a bamboo structure where all the grains span the line [11.56]. Reprinted by permission of American Institute of Physics.

For many materials, normal grain growth of polycrystalline films proceeds until the average grain diameter is approximately two to three times the film thickness. At this point, normal grain growth stops, or "stagnates." This is believed to be due, at least partly, to a surface effect called grain boundary grooving, as proposed by Mullins [11.39]. Grooves form at the top surface where grain boundaries intersect the surface. This is due to the balancing of grain boundary and surface energies at that intersection. It is similar to triple points which form 120° angles, but this time they are in the vertical plane. Grooving occurs by surface diffusion of atoms in order to minimize the surface energy, such as we discussed in the reflow process, but now with the grain boundary energy also involved. If the force on the grain boundary due to its own curvature is not large enough to overcome the pinning force from the groove, then normal grain growth stops. This tends to occur when the grain diameter is approximately two to three times the film thickness [11.44]. In grain growth simulations, this effect has been included by setting the local grain boundary velocity to zero if the local curvature falls below a critical curvature [11.44].

It has been experimentally observed that the grain size distribution for a thin film that has undergone normal grain growth is often close to a lognormal distribution. A lognormal distribution means that if the natural log of the grain diameter, $\ln(d)$, is plotted against the fraction of grains within a range of sizes, a normal (that is, Gaussian) distribution curve results. This distribution is described by

$$f(d) = \frac{1}{\sigma_d\, d\sqrt{2\pi}} \exp\left[-\left(\frac{1}{\sigma_d\sqrt{2}} \ln \frac{d}{d_{50}}\right)^2 \right] \tag{11.48}$$

where $f(d)$ is the fraction (in percent) of grains with diameter between d and Δd, d_{50} is the median grain diameter, and σ_d is the lognormal standard deviation, usually around 0.4.

An example is shown in Figure 11–54 in a variety of formats with $d_{50} = 1$ μm and $\sigma_d = 0.4$. In (a) the grain size is plotted on the x-axis, and the fraction (in percentage) of

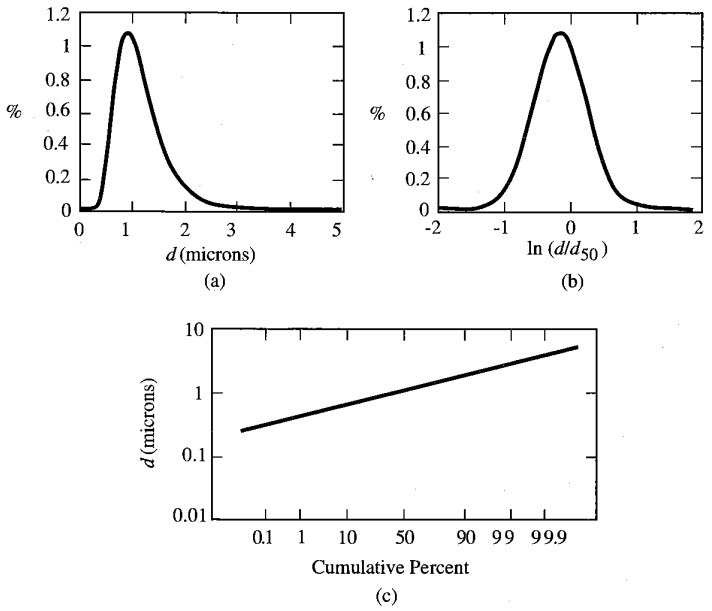


Figure 11–54 Theoretical grain size distribution [following Eq. (11.48) for a lognormal distribution, with median grain size, d_{50}, of 1 μm and σ_d of 0.4] plotted in a variety of formats: (a) percentage of grains with a given grain size, d, plotted versus d; (b) percentage of grains plotted versus $\ln(d/d_{50})$; (c) grain size (on a log scale) plotted versus the cumulative percent of those grains with size d or less.

the number of grains with that size is plotted on the y-axis. The distribution is somewhat skewed in this plot format. (The average grain size for this distribution is 1.08 μm, and the peak of the curve occurs at 0.85 μm). If we plot the fraction of grains with size d versus $\ln(d/d_{50})$, as in (b), we see a normal distribution. Finally we plot d versus the cumulative percentage of grains of size d or less (calculated by integrating the percent versus d curve up to each d value) on a lognormal scale. This shows a straight line, with the 50 percent cumulative percent being at $d = 1$ μm, as expected since that was selected as the median grain size. The slope of the curve is directly related to σ_d of the distribution.

After the normal grain growth stagnates, some grains do sometimes continue to grow. This is due to abnormal or secondary growth. If this happens, the few secondary grains grow quite rapidly until they impinge on one another and eventually all the normal grains are consumed. One reason for this is due to the fact that grains with different crystallographic orientations may have different surface energies. Those grains with lower surface energies will grow faster than, and preferentially to, grains with higher surface energies. When the grain size is on the order of the film thickness (when normal growth starts to stagnate), then the surface effects will become significant compared to the grain boundary effects, and secondary growth becomes important.

When secondary grain growth occurs, the grain size distribution will show a bimodal distribution, with two peaks in plots as in Figure 11–54(a). The first peak corresponds to

the normal growth grains and the second peak corresponds to the secondary growth grains with their larger average grain size. Eventually when the secondary grains completely consume the normal grains, the distribution becomes a standard lognormal distribution again, with just the second peak.

11.5.5 Diffusion in Polycrystalline Materials

Diffusion can be important in back-end processes. For examples, dopants can diffuse from the source/drain regions into polycrystalline silicide layers, depleting the dopants in the silicon and degrading the contact resistance. Another important example is host (or matrix) atom diffusion, such as when Al diffuses during electromigration, reflow, and grain growth. In Chapter 7 we looked at diffusion of dopants in single-crystal silicon quite extensively. Most interconnect materials are polycrystalline.

The key to understanding diffusion in polycrystalline materials is the presence of grain boundaries and the fast diffusion paths that they provide. Figure 11–50 illustrated what a grain boundary might look like at the atomic scale, and one can imagine in a simple way why diffusion of atoms might occur more quickly along grain boundaries. These regions provide more open space, letting atoms move more quickly. Atoms in these regions also tend to be less tightly bound, with unsatisfied or stretched bonds, allowing for faster movement. Calculations show that grain boundaries provide low-energy sites for vacancy formation; therefore one would expect a high concentration of vacancies along grain boundaries. This also leads to faster diffusion as compared to diffusion in the interior of the grain. While the exact mechanisms and reasons are not completely understood, the diffusion along grain boundaries is generally faster and the activation energy lower than in the lattice. The mathematical models for diffusion in polycrystalline materials are based on such a picture of a material containing a mixture of fast diffusion paths along the grain boundaries and slow diffusion paths through the lattice, or the interior or the grains.

Modeling diffusion in polycrystalline materials follows the same basic approach used in single-crystal material, except that in the two types of regions—grain boundary and lattice—the diffusion coefficients are different. In addition, the atoms can move between the two regions depending on their local concentration and the segregation coefficient and transfer coefficient between the grain boundaries and lattice regions. The flux equations for any species in each path are simply

$$F_{gb} = -D_{gb}\left(\frac{\partial C}{\partial x}\right)_{gb} \tag{11.49}$$

$$F_{\ell} = -D_{\ell}\left(\frac{\partial C}{\partial x}\right)_{\ell} \tag{11.50}$$

where D_{gb} refers to the diffusivity in, or along, the grain boundary, and D_{ℓ} is the diffusivity through the lattice, or interior of the grain. These are just forms of Fick's first law, as presented in Chapter 7, for different regions of material.

The complexity in modeling diffusion in polycrystalline materials and variations in modeling techniques arise from the determination of how much transport occurs through each of the two types of regions. This depends on the geometrical layout of the grains, which may not be uniform or regular. It may also change with time because of grain growth. Various levels of mathematical treatment of diffusion in polycrystalline materials have been developed based on assumptions of the geometrical layout.

The simplest model assumes that there is an effective diffusion coefficient for a species that describes the net diffusion through a polycrystalline material. This D_{eff} is just the sum of the diffusion coefficients in each of the two types of region—grain boundary and lattice—each scaled by the relative cross sectional area of each region. The larger the relative cross sectional area of the region, the more atoms transport through that region, and the more weight is given to the diffusion coefficient of that region. Consider diffusion through a structure with square columnar grains of length ℓ, and grain boundary thickness δ, as shown in Figure 11–55, with $\ell >> \delta$. The cross sectional area of each grain is ℓ^2 and the area of the grain boundary (per unit grain), is $\ell \cdot \delta$. The ratio of grain boundary area to grain or lattice area is thus:

$$\frac{\text{gb area}}{\text{grain area}} = \frac{\ell\delta}{\ell^2} = \frac{\delta}{\ell} \tag{11.51}$$

The effective diffusion coefficient in this model is therefore

$$D_{\text{eff}} = \left(1 - \frac{d}{\ell}\right)D_{\ell} + \frac{\delta}{\ell}D_{\text{gb}} \approx D_{\ell} + \frac{\delta}{\ell}D_{\text{gb}} \tag{11.52}$$

Diffusion through the material is described simply by Fick's law:

$$F_{\text{net}} = -D_{\text{eff}}\nabla C \tag{11.53}$$

where F_{net} is the net or total flux through the material, D_{eff} is the diffusion coefficient described by Eq. (11.52), and C is the average concentration over both grain and grain boundary regions. This approach can give only a first-order estimate of diffusion profiles.

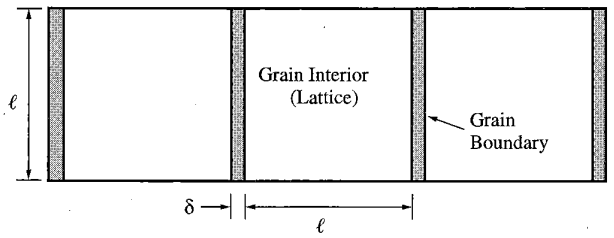


Figure 11–55 Simple representation of grains and grain boundary regions. This is a cross section of a columnar polycrystalline film with square grains. ℓ is the grain width and length and δ is the grain boundary width with $\ell >> \delta$. Diffusion from bottom to top is considered.

Example A columnar polycrystalline material has an average grain size, ℓ, of 0.1 μm and a grain boundary thickness, δ, of 0.001 μm. An impurity atom has a lattice diffusivity, D_ℓ, of 5.0×10^{-14} cm^2 sec^{-1}, and a grain boundary diffusivity, D_{gb}, of 5.0×10^{-11} cm^2 sec^{-1}. What is the effective diffusion coefficient?

Answer Using Eq. (11.52),

$$D_{eff} = D_\ell + \frac{\delta}{\ell} D_{gb} = 5.0 \times 10^{-14}\ \text{cm}^2\ \text{sec}^{-1} + \frac{0.001\ \mu\text{m}}{0.1\ \mu\text{m}} 5.0 \times 10^{-11}\ \text{cm}^2\ \text{sec}^{-1}$$

$$= 5 \times 10^{-14}\ \text{cm}^2\ \text{sec}^{-1} + 5 \times 10^{-13}\ \text{cm}^2\ \text{sec}^{-1} = 5.5 \times 10^{-13}\ \text{cm}^2\ \text{sec}^{-1}$$

Generally it is difficult to separate out the D_{gb} and δ terms from the $D_{gb}\delta$ product, the product equaling 5.0×10^{-18} cm^3 sec^{-1} in this example. One sees that the lattice diffusion adds only about 10% to the effective diffusivity.

Models much beyond this level of complexity require computer simulation techniques. We can illustrate some key ideas using TAURUS, a very flexible process and device simulator [11.28]. We assume two regions in the polycrystalline thin film, grain boundaries and grain interiors. We also assume a simple grain structure with two grains separated by a grain boundary perpendicular to the substrate, as shown in Figure 11–56(a). The grain boundary width is 0.03 μm, which is wider than actual grain boundaries but still illustrates the concepts we wish to demonstrate. We then use TAURUS to solve the coupled diffusion/continuity equations at each point in the mesh. Transport across the interior/grain boundary region is governed by simple segregation, with $k_o = C_{gb}/C_\ell$ equal to 1 in most cases. We do not include any grain boundary motion due to grain growth.

In Figure 11–56(a), the substrate below the two grains is doped at 1×10^{21} cm^{-3}. The segregation coefficient between the substrate and both the grain interior and grain boundary region is set to 100. This keeps the substrate/thin film interface concentration at 1×10^{19} cm^{-3} throughout the diffusion time. The assumed diffusivities in the grain boundary and grain interiors are given in the figure caption, with the diffusivity in the grain boundary 100 times larger than that in the grain interiors in (a). The simulation results after a 30-minute anneal are shown in the figure. (The temperature is not important in these simulations since we assume temperature-independent diffusivities here, although of course we could use Arrhenius expressions.) One can see that the dopant has diffused both up through the grain and up through the grain boundary region—much further in the latter case. One can also see where the dopant has diffused sideways from the grain boundary region into the grain interiors.

Figure 11–56(b) shows a second example in which the diffusion in the grain interior, or lattice, is only a factor of 10 slower than in the grain boundaries. More diffusion in the grains is seen, and more dopant spills into the grains from the grain boundary on

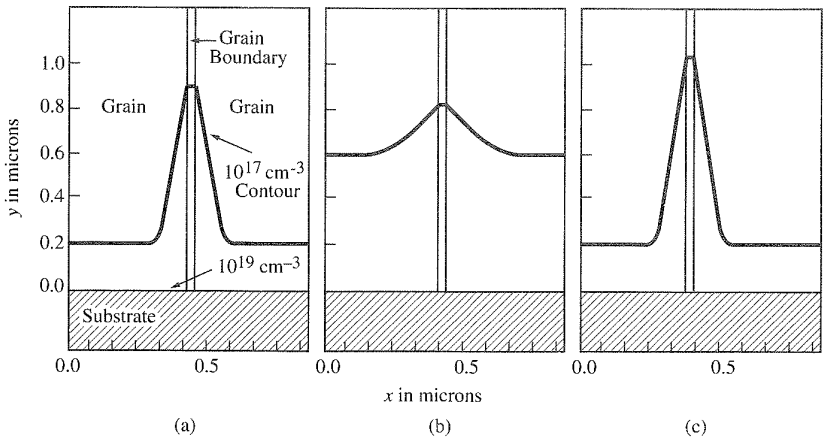


Figure 11–56 TAURUS [11.28] simulations of polycrystalline diffusion where a concentration of 1×10^{19} cm^{-3} is maintained at $y = 0$. Shown is the 1×10^{17} cm^{-3} contour after a 30 minute anneal in each case. (a) The diffusivities in the grains (or lattice) and grain boundary are 1×10^{-14} and 1×10^{-12} cm^2 sec^{-1}, respectively. (b) The diffusivity in the grains is increased to 1×10^{-13} cm^2 sec^{-1}, keeping the diffusivity in the grain boundary at 1×10^{-12} cm^2 sec^{-1}. (c) Same diffusivities as in (a), but the segregation coefficient between the grain boundary and grains is increased from 1 to 1.5.

the sides. Since the activation energy for lattice diffusion is usually greater than that of grain boundary diffusion, the relative importance of lattice diffusion increases with temperature and Figure 11–56(b) might be more representative of a high-temperature diffusion step.

If the segregation coefficient, k_o, between the grain boundary and grain regions is changed, the diffusion profiles change as well. Figure 11–56(c) shows the result when $k_o(\text{gb}/\ell)$ is 1.5 instead of 1 for the same diffusion coefficients as in Figure 11–56(a). Compared to Figure 11–56(a), one sees that the dopant preferentially stays in the grain boundary and therefore diffuses further since the diffusivity in the grain boundary is higher. Changing the segregation coefficient between the grain boundary region and grain region thus has similar effects to changing the relative diffusivities.

It is also possible to include grain growth in these models by merging the models described in Section 11.5.4 with the grain boundary and grain diffusion ideas described here. This is more complex and generally requires numerical simulation.

11.5.6 Electromigration

The phenomenon of electromigration is one of the biggest problems in interconnect technology. Electromigration is the movement of atoms due to the presence of an electrical force. When high currents pass through relatively soft polycrystalline metals with high-atomic diffusivities, such as Al, the electrons can transfer some of their momentum to the metal atoms causing them to move. The force on the atoms is found to be

$$\text{Force} = Z^* q\varepsilon = Z^* q\rho J \tag{11.54}$$

Z^* is the effective ion valence or charge number, a negative number for electron conductors on the order of -4 to -8 for Al and usually empirically determined for polycrystalline films. ε is the electric field due to the current, ρ is the electrical resistivity of the metal, and J is the current density (A cm^{-2}). The atomic flux induced by the current (in the direction of the electron flow, or opposite to the current) equals the force times the mobility times the concentration and is thus

$$F = \frac{DC}{kT} Z^* q\varepsilon = \frac{DC}{kT} Z^* q\rho J \tag{11.55}$$

where D is the atomic diffusivity and C is the atomic concentration. D is related to the motion of atoms usually via vacancy mechanisms. It is thermally activated and of the form

$$D = D_0 \exp\left(-\frac{E_A}{kT}\right) \tag{11.56}$$

so that E_A is the activation energy for diffusion. The atomic diffusivity may also be dependent on the local stress level [11.57] but for simplicity it is often assumed that it is constant with stress. The diffusion of the metal atoms (or vacancies in the opposite direction) in polycrystalline materials predominately occurs along grain boundaries, which we have seen usually have much higher diffusivities than the grain lattice. Therefore D in the above equations usually relates to atomic or vacancy diffusion through the grain boundaries. (Surface and interface diffusion may also be important, but grain boundary diffusion often plays a dominant role, particularly for Al.) One would expect that grain structure, including grain size and the location of grain boundaries, would be important in electromigration behavior.

This current-induced diffusion can lead to void and hillock formation, as illustrated earlier in Figure 11–11. Void formation can lead to an open circuit, while a hillock which extends to make contact with another interconnect line can lead to a short circuit. Either of these will lead to circuit failure. As we discussed in Section 11.4.2, the failure rate is often modeled empirically, with the following equation for the Median Time To Failure (MTTF) repeated here:

$$\text{MTTF} = AJ^{-n} \exp\left(\frac{E_A}{kT}\right) \tag{11.18}$$

where J is the current density, n is typically close to 2, and A is a constant which depends on film structure (grain size, etc.) and processing. E_A is the activation energy for electromigration and is often associated with Al grain boundary diffusion as above; its value ranges from 0.5 to 0.8 eV (the activation energy of lattice diffusion in bulk Al films is about 1.4 eV). The parameters for this equation are usually determined under accelerated testing conditions—higher than usual current and temperature. Accelerated testing is also used to predict when shorts or opens occur in the interconnect lines or when the line resistance reaches a critical value.

The original theoretical basis for Eq. (11.18), especially in regard to the n dependence of the current density, was provided by Black [11.58], and the equation is often referred to as Black's equation. He argued that n should equal 2 based on the momentum exchange between the moving electrons and the metal atoms. Other justifications for n being equal to 2, or at least being in the range of 1–3, have been based on Joule heating considerations, critical vacancy [11.59] or stress [11.60] levels, and on void nucleation and growth models [11.61].

Many models, some rather empirical like Eq. (11.18) and some much more physically based, have been developed to explain and predict electromigration failure behavior, from the atomic level up to the circuit level. Here we will describe a model that explains how interconnect reliability in Al lines depends strongly on grain structure and interconnect dimensions, as well as current density and stress properties, largely following the work of Thompson and colleagues [11.62, 11.63, 11.64, 11.65, 11.66] and Korhonen and colleagues [11.60, 11.67]. This will also serve to tie together many of the concepts described earlier in this chapter.

A very interesting result occurs if Al electromigration failure, in terms of MTTF, is measured as a function of interconnect line width, w. As w decreases to a certain value while keeping all the other parameters and conditions constant, the time to failure decreases. But then as the line width decreases further, the MTTF bottoms out and then increases. The median time to failure can get quite high for narrow lines, which is a positive result for scaling. In fact, the same behavior is seen if MTTF is measured not just as a function of line width, but also of line width divided by the average grain size of the line, or w/d. As w/d decreases, the MTTF decreases. But then at some point—somewhere below w/d equaling one—the MTTF goes way up for small values of w/d. This is shown conceptually in Figure 11–57. So if the line width is kept constant, we see a strong relationship between MTTF and grain size, but it is not a simple relationship. Why does this occur?

As the line width decreases, one might expect that there is more chance for a void to span across the whole line resulting in an increase in line resistance or open circuit. This results in a lower MTTF as w decreases while keeping d constant but would not result in a higher MTTF as w is decreased further. Since electromigration occurs via diffusion along grain boundaries, one might expect that as $1/d$ decreases (i.e., the grain size increases), the metal diffusion would decrease since there are fewer fast diffusion paths available. This would result in a higher MTTF as $1/d$ decreases. These two trends in w

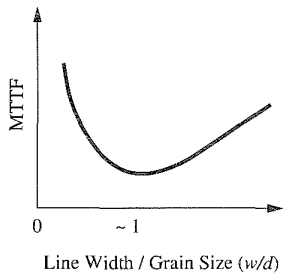


Figure 11–57 Observed relationship between Median Time To Failure (MTTF) due to electromigration versus the ratio of line width to grain size.

and $1/d$ would both be fairly linear relationships, yet in opposite directions. So when combining them, one would conjecture that the MTTF should be fairly independent of w/d. How do we then explain this behavior where the MTTF goes down and then goes up quite rapidly as w/d decreases? We start by going back to the grain growth simulations by Frost's and Thompson's groups that we presented in Section 11.5.4.

In Figure 11–52 we showed the simulation results of Frost and colleagues [11.44, 11.55] for the grain structure after grain growth in a polycrystalline layer. Figure 11–58 shows what happens to a blanket film if it is etched into straight lines of various widths. Shown are top views of lines of various values of w/d_{50} ranging from 1.30 down to 0.31. Some important features are noticed. For the narrowest line ($w/d_{50} = 0.31$), the structure is near-bamboo. When the width of the film and the median grain diameter are about the same (at $w/d_{50} = 1.30$ for example), grains that span the entire width are present. One such spanning grain is seen near the right side in the top line. For very wide lines ($w/d_{50} >> 1.30$ for example), no spanning grains would be present.

These spanning grains greatly affect electromigration behavior. Since electromigration occurs primarily by atomic diffusion along grain boundaries, the spanning grains define regions of atomic flux divergence. Figure 11–59 illustrates this. The flux of atoms will be from left to right in this case due to the electrons flowing left to right, or the current flowing right to left. This atomic flux will be greater in areas where there are no spanning grains due to the presence of grain boundaries there, and less or zero in the area of the spanning grain where there are no grain boundaries, as indicated in the figure. This results in a divergence of flux at the ends of each polygranular cluster, points A and B in the figure. Atoms will diffuse from point A to point B. But few atoms will diffuse to point A or out of point B. This means that metal atoms will be depleted at point A and will build up at point B. For lines that are wide enough so that there are no spanning grains, the only divergence of flux would occur at the ends of the lines. As long as the ends are large areas, such as bonding pads or large contact overlaps, they would not be affected very much by a depletion or buildup of material. (Although when the

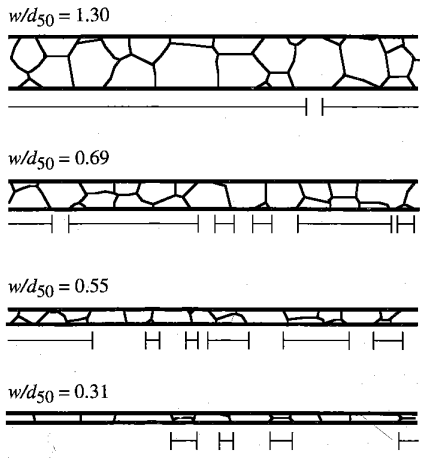


Figure 11–58 Grain structure in lines (top view) after patterning a blanket film, taking random strips of various widths so that w/d_{50} ranges from 1.30 down to 0.31 (w is the line width and d_{50} is the average grain size in the blanket film). Polygranular clusters (the regions between grains that span the entire line) are identified by the line segments below each line. One can see that as w/d_{50} decreases, the number of spanning grains increases and the average length of polygranular clusters decreases. (After [11.64])

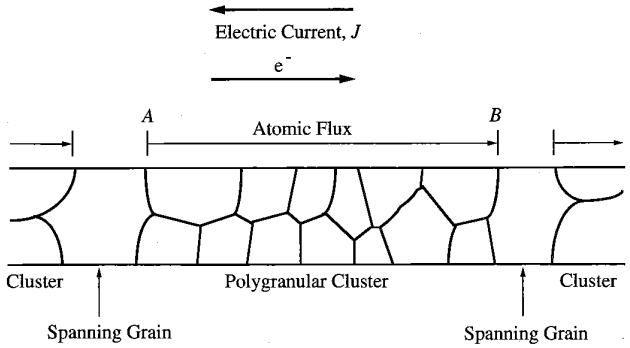


Figure 11–59 Illustration of polygranular cluster, the region between spanning grains. It is in these regions that the Al diffusion is high due to the high-diffusivity paths along the grain boundaries. A divergence of atomic flux will occur at the ends of polygranular clusters, causing metal atom depletion at point *A* and metal atom buildup at point *B*. (After [11.68].)

line ends in a tungsten via with a small contact overlap, problems can occur there, and electromigration failures at these points have become more important recently.) But in narrow lines where spanning grains are present, the depletion and buildup can lead to void and hillock formation at the ends of each spanning grain, ultimately leading to interconnect failure. Figure 11–12 showed an example of this.

The regions between the spanning grains are termed "polygranular clusters," since they contain multiple, not single, grains across the width at each point. The regions of polygranular clusters are indicated on the simulated lines in Figure 11–58. One can see that the number of spanning grains increases, and hence the average length of the polygranular clusters decreases, as the w/d_{50} value decreases. One might then expect that as the w/d_{50} ratio decreases, the chance of electromigration would get higher due to the increasing spanning grains and decrease the MTTF. It does, explaining one part of the w/d_{50} versus MTTF behavior. But then something else comes into play.

The additional effect is a back flux of atoms that opposes the forward flux. This back flux is due to a stress gradient that occurs due to the depletion and buildup of metal atoms at the flux divergence points. Figure 11–60 illustrates this. When the metal atoms diffuse due to the electromigration force, they get depleted at the start of a polygranular cluster region and build up at the end of the region. Because the metal line is constrained within the encapsulating dielectric materials, this creates a tensile stress at the start (point *A* in Figures 11–59 and 11–60) and a compressive stress at the end (point *B*). A stress gradient or field results that induces a back flux of the atoms from point *B* to point *A*. The excess atoms do not want to be where there is a compressive stress but would rather be where they are depleted. Therefore there is a flux term that opposes the original electromigration flux due to the stress gradient. (This is similar to an n/p junction in a semiconductor. The electrons and holes initially diffuse due to concentration gradients in the two carriers, but an electric field develops which wants to pull them

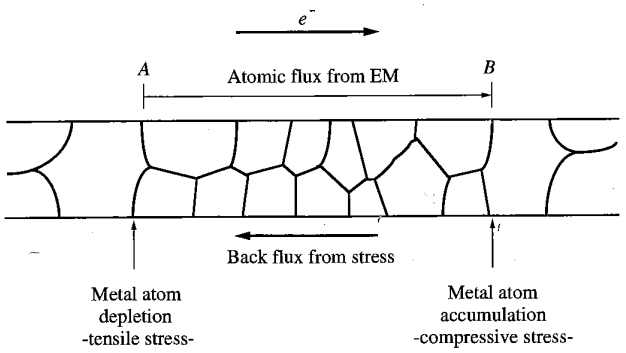


Figure 11-60 Polygranular cluster region showing stress buildup due to depletion and accumulation of the metal atoms at the ends of the region. This stress causes a back flux of the metal atoms which opposes the flux due to electromigration. (After [11.68].)

back. Ultimately a steady state is reached where the two fluxes balance each other.) The net flux due to these two opposing fluxes is now

$$F = \frac{DC}{kT}\left(Z^* q\rho J + \Omega \frac{\partial \sigma}{\partial x}\right) \tag{11.57}$$

where Ω is the atomic volume of the metal atom and σ is the stress, normal to the grain boundary. The first term in the equation is the original electromigration flux and the second term is due to the stress gradient and acts in the opposite direction.

We can solve for the steady-state condition where the net flux becomes zero and see what the stress, σ, is for that condition. To do this we set $F = 0$ in Eq. (11.57), leading to

$$\frac{\partial \sigma}{\partial x} = \frac{-Z^* q\rho J}{\Omega} \tag{11.58}$$

Letting $G = Z^* q\rho J/\Omega$, and integrating Eq. (11.58) gives

$$\sigma = -Gx + C \tag{11.59}$$

where C is the integration constant. If we assume that the stress gradient is symmetric around the center of the polygranular cluster segment of length L_p such that $\sigma = 0$ at $x = L_p/2$, we can solve for the constant, and the steady-state solution for the stress becomes

$$\sigma = G\left(\frac{L_p}{2} - x\right) = GL_p\left(\frac{1}{2} - \frac{x}{L_p}\right) \tag{11.60}$$

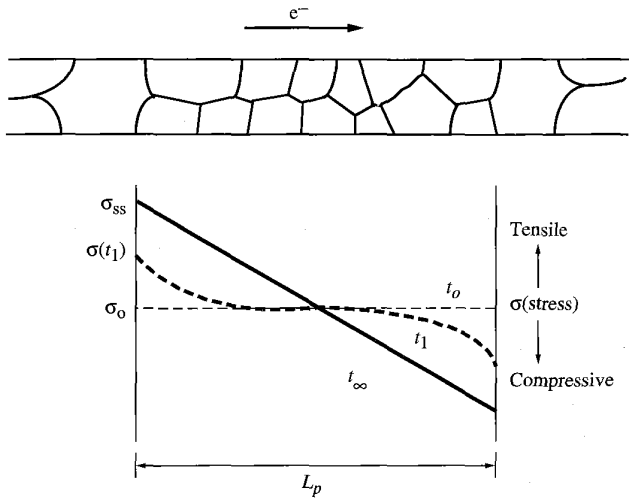


Figure 11–61 Schematic illustration of buildup of stress within polygranular cluster with time due to electromigration. In the bottom graph, stress (σ) is plotted versus distance along the cluster of length L_p for various times. Positive stress is tensile stress, and negative stress is compressive stress. Eventually (as $t \rightarrow t_\infty$), a steady-state is approached where the forward flux is equal to the backward flux due to stress. (After [11.68].)

This steady-state solution is shown in Figure 11–61.

Thus the flux due to electromigration causes depletion and accumulation of metal at the two ends of polygranular clusters, or the ends of lines in the extreme case. Stresses build up within the encapsulated metal which cause a back flux. Ultimately the back flux equals the forward flux, and no more electromigration occurs. This is the steady-state condition, and the steady-state stress of Eq. (11.60) is the stress that would be present at that time. If the metal, in its environment, can withstand and sustain this stress then failure or damage will not occur. But if it cannot, then failure is likely. The stress at which damage or failure occurs is called the critical stress, σ_{crit}. For an encapsulated metal line, which is usually the situation since metal interconnects are invariably surrounded by dielectric layers, this could correspond to the stress needed to rupture the encapsulating material. Or it could be the stress needed to initiate void formation. Not only are ruptures or voids formed, but this can also release the stress, and further electromigration of Al can occur. Even for an unencapsulated line, a stress develops which creates a back flux. The surface energy of the metal or the native metal oxide, as well as the substrate below it, will act as restraining forces. However, the critical stress in these cases would be less than for a thick dielectric layer encapsulation.

Therefore a criterion for failure of an interconnect under electromigration conditions could be if the steady-state stress is equal to or greater than some critical stress for

plastic deformation, encapsulent rupture, large void formation, and so on. The maximum steady state stress would occur at $x = 0$ or $x = L_p$. Using Eq. (11.60), the stress at these points is

$$\sigma_{max} = \pm \frac{GL_p}{2} = \pm \frac{Z^* q \rho J}{\Omega} \frac{L_p}{2} \tag{11.61}$$

If failure occurs when $\sigma_{max} = \sigma_{crit}$, then

$$\sigma_{crit} = \pm \left(\frac{Z^* q \rho J}{\Omega} \frac{L_p}{2} \right)_{crit} \tag{11.62}$$

Since Z^*, ρ, and Ω are material constants, this defines a critical JL_p product, with magnitude:

$$|JL_p|_{Crit} = \left| \frac{\sigma_{crit} 2\Omega}{Z^* q \rho} \right| \tag{11.63}$$

If the actual $|JL_p|$ is less than this value, then the maximum stress that develops is less than the critical stress and no failure will occur due to electromigration. But if $|JL_p|$ is above this value, failure will occur. This means that for a certain polygranular cluster length, no electromigration damage will occur below a critical current density. (Some electromigration of Al will occur, but it will stop due to the back flux before any critical stress level for damage occurs.) Likewise, for a given current density, no electromigration damage will occur below a critical polygranular cluster length, $(L_p)_{crit}$. Figure 11–62 illustrates this, where the steady-state stress is shown for two polygranular clusters—one long and one short. For the long cluster, which is above the critical length for the given current density, the steady-state stress at the end is greater than the critical stress. For the short cluster, which is below the critical length, the steady-state stress is less than the critical stress. In an interconnect line, if all the polygranular clusters are shorter than $(L_p)_{crit}$, the stresses that develop will all be less than the critical stress, and the line should not fail by this mechanism.

We can now explain the behavior for Al lines described in Figure 11–57, and repeated in Figure 11–63. For very large w/d_{50} values, no spanning grains are present. If there are large contact or pad areas at the ends of the line, and no significant points of flux divergence present in the line, the time to failure can be relatively high. But as w/d_{50} decreases, some spanning grains occur. This would be the case for $w/d_{50} = 1.3$ as shown in Figure 11–58. This results in divergences in the flux, causing stress to develop that can be greater than the critical stress for failure (with $L_p > (L_p)_{crit}$ for most or all of the polygranular clusters). The lines would fail and the median time to failure would decrease. Now as w/d_{50} decreases more, more spanning grains are present, but the lengths of the polygranular cluster regions between them are shorter, many of them shorter than $(L_p)_{crit}$. The stresses that develop, which create the back fluxes and which in turn eventually stop the electromigration flux, are in many cases smaller than the critical stress. For small enough values of w/d_{50} (0.31 for example), bamboo or near-bamboo structures are present. The polygranular clusters are few, and for some lines, all the clusters are

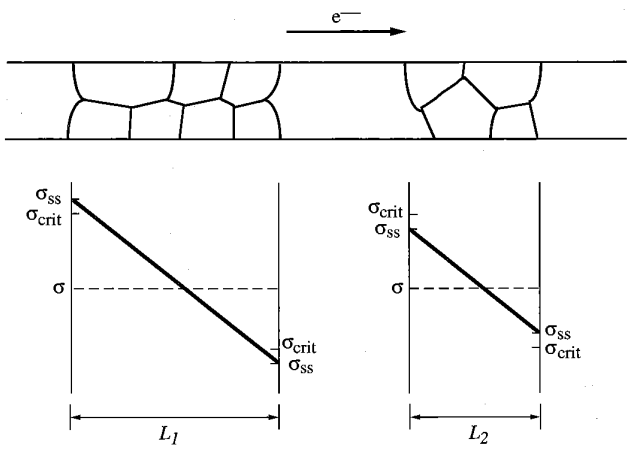


Figure 11–62 Steady-state stress distribution for two polygranular clusters of different lengths, L_1 and L_2. For the cluster of length L_1, the maximum steady-state stress (which occurs at the end of the cluster) is greater than some critical stress for line failure. For the cluster of length L_2, the maximum steady-state stress is less than the critical stress for line failure. Failure would occur in the first cluster, but not in the second. (After [11.68].)

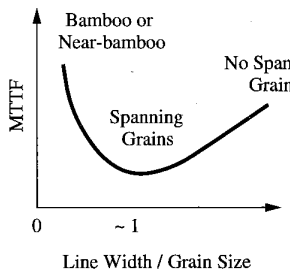


Figure 11–63 Relationship between MTTF due to electromigration versus the ratio of line width to grain size, with grain structure regimes indicated.

shorter than $(L_p)_{crit}$ and the lines will not fail by this mechanism. The MTTF thus increases dramatically. Simulations by Thompson and colleagues [11.62, 11.63, 11.64] based on this model have captured this behavior quite well.

An analysis very similar to that above was originally developed by Blech [11.69] back in 1976. He derived an expression equivalent to Eq. (11.63). (In Blech's analysis the stress is defined in terms of $\Delta\sigma$, the difference in stress between the two ends, which would be equivalent to 2σ in our analysis above.) Blech did not consider polygranular clusters, but just the entire interconnect length, L. In 1976 interconnect lines were usually not narrow enough to have many, if any, spanning grains. Without them, the whole line acts like a single polygranular cluster, and significant depletion and accumulation can occur at the

two ends of each line. This depletion and accumulation can be noticeable if there are not large areas, such as bonding pads, at the ends to act as virtually infinite sources or sinks of Al. In Blech's derivation, L is therefore used instead of L_p, but with the same results. For a given current density, Eq. (11.63) can be rewritten in terms of a critical length as

$$(L_p)_{\text{crit}} = \left|\frac{2\sigma_{\text{crit}}\Omega}{Z^* q \rho J}\right| \tag{11.64}$$

and has become known as the Blech length. L_P can refer to the actual line length for wide lines without any spanning grains, as well as to polygranular cluster length for narrow lines. Wide lines (or polygranular clusters in narrow lines) shorter than the Blech length will not incur damage or failure due to electromigration. But for a wide line longer than this which does not have large areas at the ends to serve as sources and sinks, electromigration damage can occur, usually in the form of void formation and hillock growth at the two ends, with a significant transport of metal atoms from one end of the line to the other. Typical values for the Blech length would be 5–20 μm for an Al line and about 0.5 MA cm^{-2} current density, depending on the exact conditions. Classic experiments by Blech and Herring [11.69, 11.70] accompanying their analysis nicely confirmed the concepts of electromigration-induced stress, back fluxes, critical line lengths, and critical current densities.

Example A wide Al interconnect line (with no spanning grains and no large Al sources at the ends) is 10 μm long, and a current density of 0.3 MA cm^{-2} is passed through it. If the critical stress for failure is 0.4×10^9 dyne cm^{-2}, should this line fail due to electromigration?

Answer For a wide line with no spanning grains, the whole line can be considered as one polygranular cluster. Therefore, if the line length (or line length times the current density) is greater than some critical value, then the line should fail. Either Eq. (11.63) or (11.64) can be used. Using Eq. (11.63), L_p can refer to the polygranular cluster length for narrow lines with spanning grains, or to the actual line length for wide lines without any spanning grains. For Al, the atomic volume, Ω, is 16×10^{-24} cm^3, Z^* is on the order of -4 to -8, so we will use -6, and the resisitivity of Al, ρ, is about 3.0×10^{-6} Ω cm. Therefore $|JL_p|_{\text{crit}}$, where L_p is the line length in this case, is

$$|JL_p|_{\text{crit}} = \left|\frac{0.4 \times 10^9 \text{ dyne cm}^{-2} * 2 * 16 \times 10^{-24} \text{ cm}^3}{-6 * 1.6 \times 10^{-19}\text{C} * 3 \times 10^{-6}\ \Omega \text{ cm}} * \left(\frac{\text{J}}{10^7 \text{ dyne cm}}\right)\right|$$

$$= 444 \text{ A cm}^{-1}$$

The actual $|JL_p|$ equals 0.3×10^6 A cm^{-2} times 10×10^{-4} cm, which equals 300 A cm^{-1}.

Therefore $|JL_p| < |JL_p|_{crit}$ and the line will not fail.

One could also use Eq. (11.64) and compare the actual line (or polygranular cluster) length to the critical line length. Substituting in the numbers, or noting that $(L_P)_{crit}$ just equals $|JL_p|_{crit}$ divided by $|J|$, leads to

$$(L_P)_{crit} = 444 \text{ A cm}^{-1} / 0.3 \times 10^6 \text{ A cm}^{-2} = 1.5 \times 10^{-3} \text{ cm} = 15 \text{ μm}$$

Since the actual line length of 10 μm is less than the critical line length, the line should not fail. The stress that builds up for this length is less than the critical stress for failure.

Korhonen and colleagues [11.67] have taken the stress development even further. They have solved for the stress buildup as a function of time, with the following result:

$$\frac{\partial \sigma}{\partial t} = \frac{\partial}{\partial x}\left[\frac{D\Omega\kappa}{kT}\left(\frac{\partial \sigma}{\partial x} + \frac{Z^* J \rho q}{\Omega}\right)\right] \tag{11.65}$$

where κ is the modulus relating a change in atomic concentration with stress. Eq. (11.65) is then solved numerically for the stress, σ. One solution was shown schematically in Figure 11–61. Eventually the steady-state solution is reached, also shown in that figure. In other modeling work by Thompson and colleagues. [11.65, 11.66], the MTTF is calculated by determining the time it takes for the stress level to build up to a critical value in randomly picked interconnect lines from grain growth simulations. It has also been shown that nearby polygranular clusters can interact with each other and affect the electromigration behavior [11.66].

Also important in the modeling and prediction of reliability is the variation in MTTF, that is, the standard deviation. This is needed to predict the initial failures in an interconnect line. In fact, the standard deviation is seen to increase as bamboo or near-bamboo grain structure is approached due to a certain percentage of lines having polygranular cluster regions longer than the critical length. By using a statistically large number of randomly chosen lines from the grain structure simulation, the standard deviation of MTTF can also be predicted with this model.

The model presented above for electromigration based on stress and polygranular clusters does a good job of explaining the observed phenomena regarding electromigration, line dimensions, and grain structure. The model does not deal with the specific failure mechanisms, such as void formation, hillock formation, or dielectric cracking, other than saying that failure occurs when some critical stress is reached. There is much work today in developing models that account for specific failure mechanisms as well as other phenomena and aspects of electromigration.

For example, failure may not happen right when a critical stress is reached, but sometime later when a void grows and reaches a critical size. Joule heating can occur, especially in areas where voids form and the current density is increased locally. In fact, self-healing may occur due to temperature increases around voids. We have mentioned that Cu is added to Al lines to reduce electromigration, and Cu solutes and precipitates affect the Al diffusion along grain boundaries. Electromigration in Al(Cu) films usually

occurs after an incubation period—the time it takes for the Cu itself to electromigrate from the anode side, thus depleting itself and allowing the Al to then electromigrate. Flux divergences at *W* vias can also be common failure sites. And in bamboo structures electromigration failure can still occur. In that case lattice, surface or interface diffusion can be important and must be considered.

Many models are being developed which address these and other issues. But the simple model presented above does serve to illustrate the basic concepts of electromigration, and it ties together a lot of ideas we discussed earlier in the chapter, including grain growth and grain structure, diffusion in polycrystalline materials, mechanical behavior, electrical behavior, and circuit reliability.

11.6 Limits and Future Trends in Technologies and Models

We have seen how back-end technology has evolved over the years, and it will certainly continue to change in the future. There are several driving forces for this change. One is the fact that the lateral dimensions will keep on shrinking in order to build more devices and more complex circuits per unit area. However, the vertical dimensions usually do not scale as quickly, resulting in an ever-increasing aspect ratio. This makes it harder to adequately fill spaces and planarize the topography. Another driving force is the trend toward lower thermal budgets, requiring lower processing temperatures for shorter times. Of course the overriding driving force is the need for faster circuit speeds, necessitating smaller interconnect delays. All these may require new structures, processes, and even new materials in back-end structures.

The back-end structures used in today's silicon-based integrated circuit technology have come a long way from the simple Al/SiO_2 system originally developed. Yet considering all the circuit and processing requirements and issues that have been encountered as the technology has progressed, the structures and processes have remained relatively simple. Part of this is attributable to the fact that a relatively few number of materials, Ti, $TiSi_2$, and/or TiN in particular, used in conjunction with Al(Cu) global interconnects and W contacts and vias, can be used to create very robust multilevel metallization systems. Another reason why the processing has remained relatively simple is that simple processing is in itself a goal. Over the years many solutions to problems have been proposed and even demonstrated but have not been adopted in manufacturing because they were too complex.

One possible change in the back-end structure is to replace the W contacts and vias with Al contacts and vias. This is consistent with the philosophy of fewer materials and types of processing. It also solves some problems associated with W contacts and vias. Electromigration problems in Al often occur in the vicinity of W contacts and vias. This may be partly due to mechanical stresses that can develop at W/Al interfaces which can enhance electromigration or void and hillock formation. But it is also because the different material in the contact or via serves as a point for flux divergence. Al also has a lower electrical resistivity than W which can be important especially in circuits where current flow between different metal layers is significant. Via alignment and overetching may also cause problems. Using Al as the contact or via material would reduce many of these prob-

lems. But the problems of Al contacts and vias that originally caused people to use W plugs for contact and vias would have to be solved. These problems are mostly related to filling of small, deep vias and contacts and planarized structures. Possible solutions to this are reflowing Al, either after deposition or during the deposition itself, or using ionized PVD Al deposition. Damascene processes, either single or dual, are being investigated and developed for Al in order to obtain planarized Al via and interconnect levels.

Smaller, deeper spaces and holes will be a problem for all back-end materials, not just Al. While many of these materials are still being deposited by sputtering techniques, especially the metals, the trend has been to try to use CVD. This is because better filling and conformal coverage are often achieved by CVD. However there are still problems with CVD deposition of metals which must be solved. These include contamination and difficulty with doping (Al with Cu, for example). New sputter techniques such as ionized and hot sputter depositions are also being developed, but they can be complicated processes or require a high temperature. New processes, such as the dual damascene, may put more emphasis on filling or covering spaces with metal, and less on filling with dielectrics.

CVD is being considered for silicides not only for better coverage, but for another reason as well. CVD deposition of silicides offers the advantage of little or no silicon consumption as compared with the common method of sputtering Ti and reacting it with the underlying Si. This is becoming important as junctions in the silicon are getting shallower. However, in order to obtain self-aligned silicide structures with the reactive method, silicon must be consumed. Some selective CVD techniques for silicides are being developed, but these also consume some Si. Another possible solution is the selective epitaxial deposition of Si in the source/drain regions prior to refractory metal deposition, with the epitaxial Si serving as the Si source for the silicidation process.

Another driving force in future process development is the lowering of the thermal budget. Less diffusion of dopants in the underlying Si is one reason for this. Another is the thermal stability of some of the new low-K dielectric materials being considered. Currently the highest back-end processing steps are the silicide formation anneal and the BPSG reflow anneal, both at about 800°C. For the latter, new dielectric deposition techniques are being used, notably High-Density Plasma (HDP) deposition of SiO_2, which eliminates the need for a reflow anneal. This technique is quickly replacing the normal CVD deposition and reflow of BPSG, in addition to being used for intermetal dielectric deposition. Chemical-Mechanical Polishing (CMP) methods for the dielectric layer are being used to obtain planar topographies, and more emphasis is being placed on this room-temperature process.

In $TiSi_2$ processing, a high-temperature step is needed to transform the high resistivity C49 phase to the low resistivity C54 phase. This transformation becomes even more difficult as the line thickness and width decrease to submicron dimensions. Several techniques or options are being investigated or considered to reduce this problem and allow lower processing temperatures. These include preamorphizing the Si surface by As or Ge implantation or implanting or depositing thin layers of W or Mo prior to Ti deposition. Another option is to use $CoSi_2$ instead of $TiSi_2$, sometimes with a Ti under- or overlayer. Other silicides being studied include PdSi and NiSi, which form low-resistivity silicides at low temperatures.

Multilayer interconnect structures will continue to be used, but again there will be a

trade-off between simpler structures with fewer layers and more complex structures to address some of the emerging problems. For example, to reduce electromigration failures as lines become thinner one can sandwich a TiN layer in the middle of the Al layer [11.71]. This not only provides another shunt layer but also prevents any voids from spanning the whole line. Other shunt or barrier layers are also being considered in addition to the conventional Ti or TiN layers including Mo, Ta, W-Si-N layers.

Even with all the improvements being considered and developed for the Al/SiO_2-based interconnect system, there is still a crisis looming ahead which will require some major changes in back-end structures and processing. The speed of circuits is becoming more and more dominated by the back-end structure as technology progresses. For the next generations of circuits, the interconnect delay will be prohibitively large and new technology will be required [11.9].

To consider how one may improve the interconnect speed, we turn to the simple expression from Section 11.1 for the delay time for global interconnects

$$\tau_L = 0.89 \cdot K_I K_{ox} \varepsilon_o \rho L^2 \left(\frac{1}{H x_{ox}} + \frac{1}{W L_S} \right) \tag{11.3}$$

The goal is to decrease τ_L, or at least to keep it from increasing too much as some of the dimensions change. One can see that if the metal interconnect height, H, width, W, spacing, L_S, or oxide thickness, x_{ox}, are smaller, then the delay time increases. As we discussed at the beginning of the chapter, these dimensions have generally gotten smaller as the technology has evolved, leading to in increase in interconnect delay with each technology generation. One recent trend has been to keep the metal thickness H constant to minimize the reduction in resistivity while shrinking the lateral dimensions. This does slow down the increase in RC delay as the lateral dimensions are scaled. However, if the thickness is kept constant while the lateral dimensions decrease, the aspect ratios of the features increase making the deposition and etching more difficult. And the benefits to the RC delay diminish above aspect ratios of about 2 due to the increase in line-to-line capacitance [11.7]. The second term in Eq. (11.3) will eventually dominate and the increase in RC delay will again approach 2x per generation.

Another option is to actually increase the height, width, and spacing of those interconnects where the major delays occur. These would be the longest lines and would probably be the higher levels of interconnects. Such a scheme is commonly used today, with the higher-level interconnects kept thicker and wider. The intermetal spacing, L_S, is also kept larger, which decreases the line-to-line capacitance. One disadvantage of such "reverse scaling" is that fewer metal lines can be fabricated on any level where the line widths and spacings are larger, and additional metal levels may be needed to achieve the same functionality.

Additional strategies to reduce the relative interconnect delay time will be required as the active devices get smaller and the circuits gets faster. Perhaps L, the length of the global interconnect lines, can be shortened. This will require circuit architectural changes, perhaps using buffers or repeaters. Eq. (11.3) suggests another change that can be made: changing the physical constants of the materials, ρ and K_{ox}, the metal resistivity and the dielectric constant, respectively. By decreasing these, the interconnect RC delay will be decreased.

Replacing SiO_2 with dielectric materials which have lower dielectric constants is being aggressively studied and developed for future use. Decreasing the dielectric constant would decrease the interconnect delay by decreasing both the line-to-substrate and line-to-line capacitances. Decreasing the capacitance also minimizes interconnect cross talk. As shown in Table 11–6, many low-K dielectrics are being considered. These include fluorinated SiO_2, polymers, and inorganic/organic hybrid polymers, as well as aerogels (SiO_2 with holes in them) and air bridges (no dielectric material at all!). The main problems with many of these materials are thermal and mechanical stability: They tend to evaporate or melt when heated and crack easily when subjected to stresses.

Replacing Al with a metal conductor having a lower electrical resistivity is another way of decreasing interconnect delay. For long interconnect lines where the total delay is dominated by the interconnect RC product, Eq. (11.3) indicates a linear relationship between resistivity and delay time. However, shorter or thicker lines may not be enhanced as much by reducing the interconnect resistance if the driver and load terms dominate the delay [11.4, 11.72]. The line inductance may also limit how much improvement can be made by using low resistance interconnect materials [11.5]. Alternately, one may reduce the thickness of the interconnect by using a lower-resistance metal in place of Al, keeping the same net line resistance. While R would stay the same, this would decrease the line-to-line capacitance and the total delay. Cross talk would also be decreased with the thinner interconnects.

Three elements do have lower resistivities than Al: silver, gold, and copper. Silver has corrosion problems and poor electromigration resistance, and gold has just marginally lower resistivity than Al along with device contamination problems. Copper has a lower resistivity, 1.72 $\mu\Omega$-cm versus 2.7 for Al. Perhaps more importantly, Cu has much better electromigration resistance than Al, although more research must be carried out. Higher electromigration resistance also allows for circuit operation at higher current densities in the interconnects, resulting in potentially faster circuit speed. For these reasons, many manufacturers have been developing new processes that use Cu as an interconnect metal.

However, there are problems with using copper interconnects which have hindered its development. Copper is difficult to plasma etch since the byproducts, copper halides, are nonvolatile at low temperatures. Corrosion of copper is a problem, especially if the halides are not all removed. A damascene process, single or dual, is the obvious solution to this problem since the copper in this process does not have to be plasma etched, just polished back using CMP. Another problem with copper interconnects is that Cu is a harmful contaminant in Si devices, and Cu atoms can move quickly through SiO_2 and organic dielectrics. Cladding layers, of TiN, Si_3N_4 or Ta-based alloys, for example, must be used as barriers around the copper to prevent this. But this can reduce the effectiveness of using the lower-resistivity material if cladding layers of higher-resistivity material must be used on all sides of the interconnect, thus reducing the volume of the Cu. One alternative to this is to treat the surface of the dielectric, using NH_3 plasma for example, which creates a diffusion barrier out of the dielectric instead of out of the metal interconnect [11.73]. Another disadvantage of copper is that it is not self-passivating which can lead to corrosion problems, unlike Al which forms a thin protective Al_2O_3 skin in air.

Yet, as in the past with other new technologies, the problems with Cu are being solved and Cu will be common in many IC processes in the near future. These processes will make use of the damascene process, Cu electroplating, and CMP. But questions still remain. Will Cu be used only on upper metal layers or will it replace Al entirely, keeping with the "keep it simple" philosophy? Also, which circuits will truly benefit from Cu interconnects? Perhaps only high-performance logic circuits will. Memory chips may use Al interconnects for some time to come.

Even if new materials are used for interconnects and dielectrics, there is a limit to which they will help the circuit speed. Copper has only a 40% lower resistivity than Al so that the improvement in signal delay is a maximum of 40%—less if additional cladding is required. The lowest dielectric constant obtainable is 1 (that of air), so there is a limit to which this can be improved. So even by using Cu as the interconnect instead of Al, and air as the dielectric instead of SiO_2, the maximum decrease in signal delay is about 85%. Eventually another evolution is required if the progression to even faster integrated circuits is to continue. Other types of interconnects may be needed such as superconductors, optical interconnects, or RF transmitters to broadcast signals across chips. And other circuit architecture will probably be needed, with less dependence on long global interconnect lines.

Even though the technologies will evolve and the materials may change, the basic models and concepts as presented in the chapter should still apply. Whether the main global interconnect material is Al or Cu, the silicide is $TiSi_2$ or $CoSi_2$, the contact material is W or Al, or the dielectric material is SiO_2 or a low-K polymer, the basic models for silicide formation, CMP, reflow, grain growth, grain boundary diffusion, and electromigration will still be applicable.

11.7 Summary of Key Ideas

In this chapter we have looked at back-end technology—the structures and fabrication techniques to electrically connect devices to each other and to the outside world. As structures and devices get smaller and chips larger, the importance of interconnects and back-end structures increases—even dominates—over the active devices in terms of circuit performance. The back-end structures include a variety of components: contacts to active areas, local interconnects, global interconnects, vias between interconnect levels, first-level dielectrics, and intermetal dielectrics to physically and electrically separate the different interconnects, and finally a passivation layer to protect the entire circuit. We saw how the back-end structure has evolved considerably from the simple $Si/SiO_2/Al$ structures in the first integrated circuits. Today's technology is characterized by multilevel interconnects, each level itself made up of multilayer metal structures with multilayer dielectrics separating them. Adhesion, contact, barrier, and shunt layers are used in the interconnects, while barrier and planarization layers are utilized in the dielectric structures. There have been many driving forces at work in this development. Some of the most important include the need for additional interconnect levels, the need for planarized structures, lower processing temperatures, and the minimization of reliability problems. In addition, there has always been an effort to make the process as simple and cost effective as possible.

We described the typical processing steps that are commonly used in back-end technology. These include the formation of local interconnects such as silicide structures; the fabrication of the first-level dielectrics like reflowed BPSG; contact formation using tungsten plugs and the etchback process or Chemical-Mechanical Polishing (CMP); global interconnects with stacked multi-layer structures; intermetal dielectric deposition and planarization, using different combinations of PECVD oxide, spin-on-dielectrics, high-density plasma CVD gap-fill oxides, and CMP; via formation with either W or Al plugs; and final passivation with PECVD oxide/nitride layers.

Back-end structures and technology will continue to evolve as devices get smaller, processing temperatures decrease, and even higher performance is required. We will soon be switching to copper, low-K dielectric interconnect systems for many types of high-performance circuits, or to something even more radically different in the future. Nevertheless, many of the same basic processing concepts and ideas will be important.

11.8 References

[11.1]. K.C. Saraswat and F. Mohammadi, "Effect of Scaling of Interconnections on the Time Delay of VLSI Circuits," *IEEE Trans. Electron Dev.*, vol. ED-29, p. 645, 1982.

[11.2]. A. Wilnai, "Open-Ended RC Line Model Predicts MOSFET IC Response," *EDN*, p. 53, 1971.

[11.3]. G. K . Rao, *Multilevel Interconnect Technology*, McGraw-Hill, 1993, p. 24.

[11.4]. H.B. Bakaglu and J.D. Meindl, "Optimal Interconnection Circuits of VLSI," *IEEE Trans. Elect. Dev.*, vol. ED-32, p. 903, 1985.

[11.5]. T.Schreyer, Y. Nishi, and K.C. Saraswat, "A Complete RLC Transmission Line Model of Interconnect Delay," *Proc. of 1988 Symp. on VLSI Technology*, p. 95, 1988.

[11.6]. M.T. Bohr and Y.A. El-Mansy, "Technology for Advanced High-Performance Microprocessors," *IEEE Trans. Elec. Dev.*, vol. 45, p. 620, 1998.

[11.7]. M.T. Bohr, "Interconnect Scaling—the Real Limiter to High Performance ULSI," *IEDM Technical Digest,* p. 241, 1995.

[11.8]. A.K. Stamper, V. McGahay, and J.P. Hummel, "Intermetal Dielectrics—a Five Year Outlook," *Proc. of Diel. for ULSI Multilevel Inter. Conf.*, p. 13, 1997.

[11.9]. National Technology Roadmap for Semiconductors, SIA, 1997.

[11.10]. S. M. Sze, *Physics of Semiconductor Devices*, 2d ed., John Wiley & Sons, 1981, p. 305.

[11.11]. C.Y. Chang, Y.Y. Fang, and S.M Sze, "Specific Contact Resistance of Metal-Semiconductor Barriers," *Solid State Elect.*, vol. 14. p. 541, 1971.

[11.12]. M.A. Nicolet, "Diffusion Barriers in Thin Films," *Thin Solid Films,* vol. 19, p. 52, 1978.

[11.13]. S. Wolf, *Silicon Processing for the VLSI Era, Vol. 2, Process Integration*, Lattice Press, 1990, p. 126.

[11.14]. C.R.M. Grovenor, *Microelectronic Materials*, Institute of Physics Publishing, 1989, p. 278.

[11.15]. M. Wittmer, "Barrier Layers: Principles and Applications in Microelectronics," *J. Vac. Sci. Tech. A.*, vol. 2, p. 273, 1984.

[11.16]. S Bauguess, L.H. Liu, M.L. Dreyer, M. Griswold, and E. Hurley, "The Effects of Accelerated Stress Conditions on Electromigration Failure Kinetics and Void Morphology," *Mat. Sci. Soc. Symp. Proc.*, vol. 428, p. 93, 1996.

[11.17]. S.P. Maruarka, *Silicides for VLSI Applications*, Academic Press, 1983.

[11.18]. D. Pramanik, "CVD Dielectric Films for VLSI," *Semicond. Int'l.*, June 1988, p. 94.

[11.19]. C. H. Ting and T.E. Seidel, "Methods and Needs for Low K Material Research," *Mat. Sci. Soc. Symp. Proc.*, vol. 381, p. 3, 1995.

[11.20]. S.J. Fang, A. Barda. T.Janecko, W. Little, D. Outley, G. Hempl, S. Joshi, B. Morrison, G.B. Shinn, and M Birang, "Control of Dielectric Chemical Mechanical Polishing (CMP) Using An Interferometry Based Endpoint Sensor," *Proc. of the Inter. Interconnect Tech. Conf.*, 1998, p. 76.

[11.21]. D. K. Schroder, *Semiconductor Material and Device Characterization*, J. Wiley & Sons, 1990.

[11.22]. W.M. Loh, S.E. Swirhun, T.A. Schreyer, R. M. Swanson, and K.C. Saraswat, "Modeling and Measurement of Contact Resistances," *IEEE Trans. Elec. Dev.*, vol. ED-34, p. 512, 1987.

[11.23]. W.M Loh, "Modeling and Measurement of Contact Resistances," Ph.D. thesis, Stanford University, 1987.

[11.24]. W.T. Lynch and K.K. Ng, "A Tester for the Contact Resistivity of Self-Aligned Silicides," *IEDM Tech. Dig.*, p. 352, 1988.

[11.25]. R. Venkatraman, "Plasticity and Flow Stresses in Al Thin Films on Si," Ph.D. thesis, Stanford University, 1992.

[11.26]. SPEEDIE is a topography simulator developed at Stanford Univeristy.

[11.27]. ATHENA is a process simulator from Silvaco Inc.

[11.28]. TSUPREM-IV and TAURUS are process (and device) simulators from Avant! Inc.

[11.29]. S. Cea and M.E. Law, "Two Dimensional Simulation of Silicide Growth and Flow," *Proc. of IEEE Int'l. Workshop on Numerical Modeling of Processes and Devices for Int. Circuits, NUPAD V,* 1994, p. 113.

[11.30]. S. Cea, "Multidimensional Viscoelastic Modeling of Silicon Oxidation and Titanium Silicidation," Ph.D. thesis, University of Florida, 1996.

[11.31]. S. Cea, unpublished.

[11.32]. FLOOPS is a process simulator developed by Mark Law at University of Florida.

[11.33]. C.A. Pico and M.G. Lagally, "Kinetics of Titanium Silicide Formation on Single Crystal Si: Experiment and Modeling," *J. Appl. Phys,* vol. 64, p. 4957, 1988.

[11.34]. V. Svilan, J.M.E. Harper, C. Cabral, L.A. Clevenger, "Stress Evolution During the Formation and Transformation of Titanium Silicide," in *Thin Films: Stresses and Mechanical Properties V. Symposium*, S.P. Baker, editor, Materials Research Society, 1995, Pittsburgh, p. 167.

[11.35]. H. Norstrom, K. Maex, and P. Vandenabeele, "Thermal Stability and Interface Bowing of Submicron $TiSi_2$/Polycrystalline Silicon," *Thin Solid Films*, vol. 198, p. 53, 1991.

[11.36]. H. Norstrom, K. Marx, A. Romano-Rodriquez, J. Vanhellemont, and L Van den hove, "Formation of $CoSi_2$ and $TiSi_2$ on Narrow Poly-Si Lines," *Microelectron. Eng.*, vol. 14. p. 327, 1991.

[11.37]. F.W. Preston, "The Theory and Design of Plate Glass Polishing Machines," *J. Soc. Glass Tech.,* vol. 11, p. 214, 1927.

[11.38]. J. Warnock, "A Two-Dimensional Process Model for Chemimechanical Polish Planarization," *J. Electrochem. Soc.*, vol. 138, p. 2398, 1991.

[11.39]. W.W. Mullins, "Theory of Thermal Grooving," *J. Appl. Phys.*, vol. 28, p 333, 1957.

[11.40]. W.W. Mullins, "Flattening of a Nearly Plane Solid Surface Due to Capillarity," *J. Appl. Phys.*, vol. 30, p 77, 1959.

[11.41]. R.A. Brain, D.S. Gardner, D.B. Fraser, and H.A. Atwater, "The Effect of Grain Boundaries on Surface Diffusion Mediated-Planarization of Polycrystalline Cu Films," *Mat. Res. Soc. Symp. Proc.*, vol. 389, p. 107, 1995.

[11.42]. R.A. Brain, "Capillary-Driven Reflow of Thin Cu Films with Submicron, High Aspect Ratio Features," Ph.D. thesis, California Institute of Technology ,1996.

[11.43]. S.P. Chen, J.D. Kress, A.F. Voter, and R.C. Albers, "Electrically Inactive Poly-Silicon Grain Boundaries," *Electrochem. Soc. Proc.*, vol. 96-4, p. 359, 1996.

[11.44]. H.J. Frost, "Microstructural Evolution in Thin Films," *Materials Characterization*, vol. 32, p. 257, 1994.

[11.45]. C.V. Thompson, "Grain Growth in Thin Films," *Ann. Rev. of Mat. Sci.*, vol. 20, p. 245, 1990.

[11.46]. W.W. Mullins, "Two-Dimensional Motion of Idealized Grain Boundaries," *J. Appl. Phys.*, vol. 27, p. 900, 1956.

[11.47]. C.V. Thompson, "Secondary Grain Growth in Thin Films of Semiconductors: Theoretical Aspects," *J. Appl. Phys.*, vol. 58, p. 763, 1985.

[11.48]. H.J. Kim and C.V. Thompson, "Kinetic Modeling of Grain Growth in Polycrystalline Silicon Films Doped with Phosphorus of Boron," *J. Electrochem. Soc.*, vol. 135, p. 2312, 198

[11.49]. D.A. Porter and K.E. Eastering, *Phase Transformations in Metals and Alloys*, 2d ed., Chapman and Hall, 1992, p. 139.

[11.50]. W.W. Mullins, "The Statistical Self-Similarity Hypothesis in Grain Growth and Particle Coursening," *J. Appl. Phys.*, vol. 59, p. 1341, 1986.

[11.51]. J.E. Burke, "Some Factors Affecting the Rate of Grain Growth in Metals," *Trans. Amer. Inst. of Mining and Metall. Engineers,* vol. 180, p. 73, 1949.

[11.52]. S.K. Kurtz and F.M.A. Carpay, "Microstructure and Normal Grain Growth in Metals and Ceramics. Part 1, Theory," *J. Appl. Phys.*, vol. 51, p. 5725, 1981.

[11.53]. L.H. Van Vlack, *Elements of Materials Science and Engineering*, Addison-Wesley, 1989.

[11.54]. M. Hillert, "On the Theory of Normal and Abnormal Grain Growth," *Acta Metall.*, vol. 13, p. 227, 1965.

[11.55]. H.J. Frost and C.V. Thompson, "Computer Simulation of Microstructrual Evolution in Thin Films," *J. of El. Mater.*, vol. 17, p. 447, 1988.

[11.56]. D.T. Walton, H.J. Frost, and C.V. Thompson, "Development of Near-Bamboo and Bamboo Microstructures in Thin-Film Strips," *Appl. Phys. Lett.,* vol. 61, p. 40, 1992.

[11.57]. Y.J. Park and C.V. Thompson, "The Effects of the Stress Dependence of Atomic Diffusivity on Stress Evolution due to Electromigration," *J. Appl. Phys.*, vol. 82, p. 4277, 1997.

[11.58]. J.R. Black, "Electromigration—A Brief Survey and Some Recent Results," *IEEE Trans. El. Dev.*, vol. ED – 16, p. 338, 1969.

[11.59]. M. Shatzkes and J.R. Lloyd, "A Model for Conductor Failure Considering Diffusion Concurrently with Electromigration Resuling in a Current Exponent of 2," *J. Appl. Phys.*, vol. 59, p. 3890, 1986.

[11.60]. M.A. Korhonen, P. Borgeson, K.N. Tu, and C.Y. Li, "Stress Evolution due to Electromigration in Confined Metal Lines," *J. Appl. Phys.*, vol. 73, p. 3790, 1993.

[11.61]. J.R. Lloyd, "Electromigration Failure," *J. Appl. Phys.*, vol. 69, p. 7601, 1991.

[11.62]. C.V. Thompson and J.R. Lloyd, "Electromigration and IC Interconnects," *MRS Bulletin*, December 1993, p. 19.

[11.63]. C.V. Thompson, Y.C. Joo, and B.D. Knowlton, "Modeling of the Structure and Reliability of Near-Bamboo Interconnects," *Mat. Res. Soc. Symp. Proc.*, vol. 391, p. 163, 1995.

[11.64]. Y.C. Joo and C.V. Thompson, "Analytic Model for the Grain Structures of Near-Bamboo Interconnects," *J. Appl. Phys.*, vol 76, p. 7339, 1994.

[11.65]. B.D. Knowlton, J.J. Clement, R.I. Frank, and C.V. Thompson, "Coupled Stress Evolution in Polygranular Clusters and Segments in Near-Bamboo Interconnects," *Mat. Res. Soc. Symp. Proc.*, vol. 391, p. 189, 1995.

[11.66]. B.D. Knowlton, J.J. Clement, and C.V.Thompson, "Simulation of the Effects of Grain Structure and Grain Growth on Electromigration and the Reliability of Interconnects," *J. Appl. Phys*, vol. 81, p. 6073, 1997.

[11.67]. M.A. Korhonen, P. Borgeson, D.D. Brown, and C.Y. Li, "Microstructure Based Statistical Model of Electromigration Damage in Confined Line Metallizations in the Presence of Thermally Induced Stressses," *J. Appl. Phys.*, vol. 74, p. 4995, 1994.

[11.68]. C.V. Thompson, "Electromigration, Microstructure, and IC Interconnect Reliability: Metallurgy Meets TCAD," presented at Stanford University, Nov. 1995.

[11.69]. I.A. Blech, "Electromigration in Thin Aluminum Films on Titanium Nitride," *J. Appl. Phys.*, vol. 47, p. 1203, 1976.

[11.70]. I.A. Blech and C. Herring, "Stress Generation by Electromigration," *Appl. Phys. Lett.*, vol. 29, p. 131, 1976

[11.71]. D. Gardner, "Layered and Homogeneous Films of Al and Al/Si with Ti and W for Multilevel Interconnects," *IEEE Trans. Elect. Dev.*, vol. ED-32, p. 174, 1985.

[11.72]. S. Bothra, B. Rogers, M. Kellam, and C.M. Osburn, "Analysis of the Effects of Scaling on Interconnect Delay in ULSI Circuits," *IEEE Trans. on Elect. Dev.*, vol. 40, p. 591, 1993.

[11.73]. K. Mikagi, H. Ishikawa, T. Usami, M. Suzuki, K. Inoue, N. Oda. S. Chikaki, I. Sakai, and T. Kikkawa, "Barrier Metal Free Copper Damascene Interconnections Technology Using Atmospheric Copper Reflow and Nitrogen Doping in SiOF Film," *IEDM Technical Digest*, vol. 96, p. 365, 1996.

11.9 Problems

11.1. Replot the delay time versus chip area curve (Figure 11–4) for $F_{min} = 0.25$ μm, for the following cases (all on the same graph), assuming no driver or load elements.

a. With Al as the interconnect metal (assume the resistivity of Al is 3.0 μΩ cm) and SiO_2 as the dielectric (assume $K_{ox} = 3.9$).

b. With Al as the interconnect metal but with a new dielectric material, whose dielectric constant, K, is equal to 2.2.

c. With Cu as the interconnect metal (resisitivity is 1.7 μΩ cm) and a low-K dielectric material (K is equal to 2.2).

11.2. Calculate the percentage increase in the interconnect RC delay according to Eq. (11.3) if the thicknesses H and x_{ox}, remain constant while the lateral dimensions W and L_S scale with (and equal) F_{min}. Assume F_{min} is decreased from 0.5 to 0.35 μm and H and x_{ox} equal 0.5 μm. Also assume that the interconnect length L remains constant.

11.3. Al spiking of Si would not be a problem if the Si diffusion in Al were small and little Si diffused into the Al, leaving voids in the Si for the Al to fill. Just how much does the Si diffuse into the Al? If the diffusivity of Si in Al at 450°C is about 40 $\mu m^2\ min^{-1}$, calculate how far the Si will diffuse into the Al after 60 minutes at 450°C? (Calculate the distance in which the Si concentration falls to 50% of the surface concentration.)

11.4. For 0.35-μm technology, the junction depth (before silicidation) is about 100 nm. If you want to leave 50 nm of Si after silicidation to ensure low leakage current, how much $TiSi_2$ is formed and how much Ti is needed if all is consumed?

11.5. A source/drain implant is done with arsenic with a dose of $1 \times 10^{15}\ cm^{-2}$ and an energy of 40 keV. A titanium silicide layer is then formed on top of the source/drain regions to reduce the sheet resistance of those regions. This is done by depositing Ti on the surface and annealing. 55 nm of silicide is formed, which consumes the top surface of the Si in the source/drain regions and as a result reduces the amount of As dopant in the Si. What is the peak concentration of As in the Si in the source/drain regions after the silicidation process? (Assume that the implant is done directly into the Si with no oxide and that no dopant diffusion or segregation occurs during the silicidation.)

11.6. What are two reasons why the damascene process (single damascene version) might be used instead of the normal masked plasma etch process?

11.7. An oxide layer is deposited over 0.5-μm-thick metal lines, as in Figure 11–17. The resulting step height in the oxide layer is 0.4 μm. What is the degree of planarization?

11.8. 0.5-μm-thick Al lines are formed on a flat surface. A layer of SiO_2 (at least 1 μm thick) is then deposited over this. A Degree Of Planarization (DOP) of 1 is desired but the actual DOP obtained was 0.4. A CMP etchback of the oxide is then done to obtain a net degree of planarization of 1. If the CMP etch rate is 0.12 μm per minute, what is the minimum CMP etch time needed to obtain this net DOP of 1? Assume no variations in thicknesses or CMP etch rates.

11.9. A damascene process is used to fabricate a tungsten via through a SiO_2 dielectric layer. First the SiO_2 dielectric layer is deposited, with a thickness of 1 μm. A via hole is etched in the SiO_2, and W is blanket deposited by CVD. Right after the W deposition, the Degree Of Planarization (DOP) equals 1.0, and the thickness of the W directly above the dielectric layer is equal to 0.8 μm. A plasma etchback of the W layer is now done to remove the W that covers the top of the dielectric layer, and leaving the W only in the via hole. If the etch rate of the W in the etchback process is 5.0 nm sec^{-1}, and the etch selectivity of W with respect to SiO_2 is 4:1, what is the profile of the structure after 180 sec of etching? Specify the heights of the W and SiO_2 layers in nm. (Neglect any variations in thicknesses or etch rate.)

11.10. A SiO_2 layer is deposited by LPCVD over a single 1-μm-wide, 0.5-μm-high metal line. The deposition flux and time are such that 0.25 μm of SiO_2 would be deposited on a flat surface. The sticking coefficient for this process is 0.01 and the arrival angle distribution parameter n is equal to 1. Very conformal deposition is obtained, but the DOP is 0. Would changing n or S_C improve (increase) the DOP? Explain.

11.11. SiO_2 is deposited on a Si wafer at 800°C. How much stress is induced in the film due to the cooling of the wafer from 800°C to 25°C? Is the stress tensile or compressive? Assume no intrinisic stress in the film and no stress relaxation processes.

11.12. Consider a surface that has a region that is concave and as well as a region that is convex. During reflow, will atoms move from a concave region to a convex region, or from a convex region to a concave region? Explain both in terms of the force equations for reflow as well as in terms of the atomistic picture.

11.13. If the grain boundary mobility, M_{gb}, is 1×10^{-15} cm^4 sec^{-1} erg^{-1}, the grain boundary energy is 300 ergs cm^{-2}, and the curvature, K, is 20 μm^{-1}, calculate the average grain radius after 60 minutes of grain growth. Assume the initial grain size is zero.

11.14. An Al interconnect line has polygranular clusters that are all equal or less than 5 μm long, and a current density of 0.5 MA cm^{-2} is passed through it. If the critical stress for failure is 0.4×10^9 dyne cm^{-2} (and let Z^* equal -6), should this line fail due to electromigration? What is the critical current density below which electromigration failure will not occur for this line?

Appendices

A.1 Standard Prefixes

Prefix	Symbol	Value
peta-	P-	10^{15}
tera-	T-	10^{12}
giga-	G-	10^{9}
mega-	M-	10^{6}
kilo-	k-	10^{3}
milli-	m-	10^{-3}
micro-	μ-	10^{-6}
nano-	n-	10^{-9}
pico-	p-	10^{-12}
femto-	f-	10^{-15}

A.2 Useful Conversions

1 atm = 760 torr (exactly)
1 atm = 1.013×10^5 Pa
1 Pa = 1 J m^{-3} = 10 dyne cm^{-2}, 1 MPa = 10^7 dyne cm^{-2}
1 eV = 1.60219×10^{-19} J
1 cal = 4.184 J
1 J = 10^7 erg
1 erg = 2.39×10^{-8} cal
1 N = 1 J m^{-1} = 10^5 dyne
1 V = 1 ohm amp
1 J = 1 volt coulomb = 1 coulomb amp ohm
1 J = 10^7 dyne cm
1 coulomb = 1 amp sec = 1 farad volt
1 amp = 1 farad volt sec^{-1}
1 nm = 10^{-3} μm = 10^{-7} cm = 10^{-9} m = 10 Å
T K = T°C + 273.15

A.3 Physical Constants

Quantity	Symbol	Value
Avogadro constant	N_{AV}	6.02204×10^{23} mole^{-1}
Boltzmann constant	k	1.38066×10^{-23} J K^{-1} = 8.617×10^{-5} eV K^{-1}
Electron charge	q	1.60218×10^{-19} C
Electron rest mass	m_o	9.1×10^{-28} g
Gas constant	R	1.987 cal mole^{-1} K^{-1}
Permittivity in vacuum	ε_0	8.854×10^{-14} F cm^{-1}
Planck constant	h	6.626×10^{-34} J sec = 4.136×10^{-15} eV sec
Reduced Planck constant	$\hbar\ (=h/2\pi)$	1.055×10^{-34} J sec
Speed of light in vacuum	c	2.998×10^{10} cm sec^{-1}
Thermal voltage at 300 K	kT/q	0.0259 V

A.4 Physical Properties of Silicon

Property	Value
Band gap at 300K	1.107 eV
Electron majority carrier mobility at 300K	1500 cm^2 volt^{-1} sec^{-1}
Hole majority carrier mobility at 300K	450 cm^2 volt^{-1} sec^{-1}
Crystal structure	Cubic diamond
Melting point	1417°C
Density	2.328 gm cm^{-3}
Knoop hardness	950–1150 kg mm^{-2}
Thermal conductivity k_S at 300K	0.358 cal (sec cm °C)$^{-1}$
Latent Heat of Fusion	430 cal gm^{-1}
Emissivity at high temperatures	≈ 0.55
Linear thermal expansion coefficient $\Delta L/L\Delta T$	2.6×10^{-6} °C^{-1}
Young's modulus	$Y_{100} = 1.3 \times 10^{12}$ dyne cm^{-2} $Y_{111} = 1.9 \times 10^{12}$ dyne cm^{-2}
Lattice constant	0.543095 nm
Specific heat	0.7 Joule gm^{-1} °C^{-1}
Thermal diffusivity	0.92 cm^2 sec^{-1}
Atomic density	5×10^{22} cm^{-3}
Atomic weight	28.09
Breakdown field	$\approx 3 \times 10^5$ V cm^{-1}
Effective density of states (valence band) N_V	1.04×10^{19} cm^{-3}
Effective density of states (conduction band) N_C	2.8×10^{19} cm^{-3}
Density of states effective mass of electrons m_e^*	1.08 m_o ($m_o = 9.1 \times 10^{-28}$ g)
Density of states effective mass of holes m_h^*	0.81 m_o
Peak electron velocity v_{SAT}	1×10^7 cm sec^{-1}
Peak hole velocity v_{SAT}	1×10^7 cm sec^{-1}
Electron affinity χ	4.05 eV
Intrinsic carrier concentration n_i at 300 k	1.45×10^{10} cm^{-3}
Intrinsic resistivity at 300 k	2.3×10^5 Ωcm
Relative permittivity ε_r	11.9

A.5 Properties of Insulators Used in Silicon Technology

	SiO_2	Si_3N_4	SiO_xN_y
Structure	Amorphous	Amorphous	Amorphous
Resistivity (Ωcm)	$10^{14} - 10^{16}$	$\approx 10^{14}$	
Density (gm cm^{-3})	2.27	3.1	
Dielectric constant, K	3.8 – 3.9	7.5	5 – 6
Dielectric strength ($V\ cm^{-1}$)	$5–10 \times 10^6$	$\approx 10^7$	$\approx 5 \times 10^6$
Energy gap (eV)	≈ 9	≈ 5	
Expansion coeff. ($°C^{-1}$)	5×10^{-7}		
Refractive index	1.46	2.05	1.6 – 1.9
Thin film stress (dyne cm^{-2})	$2–4 \times 10^9$	$9–10 \times 10^9$	$1–6 \times 10^9$
Thermal conductivity (W/cm°C)	0.014		
Infrared absorption peak (μ)	9.3	11.5 – 12	9 – 12
Etch rate in buffered HF (nm min^{-1})	100	0.5–1	2–40
Melting point (°C)	≈ 1700		
Electron mobility ($cm^2\ volt^{-1}\ sec^{-1}$)	20–40		
Hole mobility ($cm^2\ volt^{-1}\ sec^{-1}$)	$\approx 2 \times 10^{-5}$		
Surface conductivity (Ω/sq)	$10^{16}–10^{19}$		

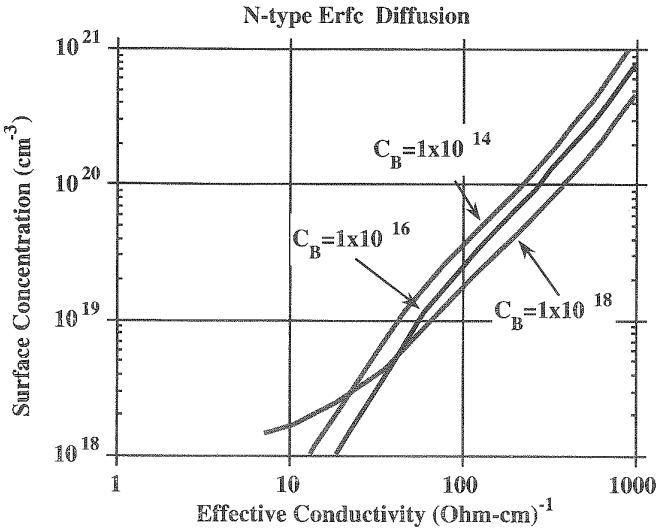
A.6 Color Chart for Deposited Si_3N_4 Films Observed Perpendicularly under Daylight Fluorescent Lighting

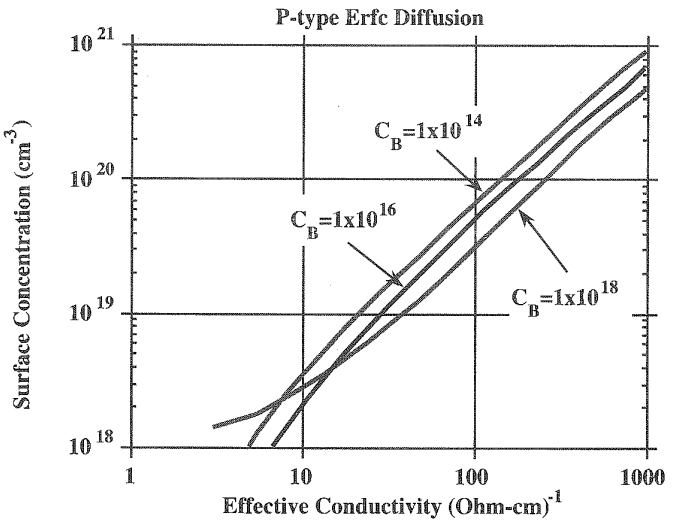
Film Thickness (nm)	Color	Film Thickness (nm)	Color
10	Light brown	95	Light blue
17	Medium brown	105	Very light blue
25	Brown	115	Light blue-brown
34	Brown-pink	125	Light brown-yellow
35	Pink-purple	135	Very light yellow
43	Purple	145	Light yellow
53	Very dark blue	155	Yellow
60	Dark blue	165	Bright yellow
70	Medium blue	175	Intense yellow

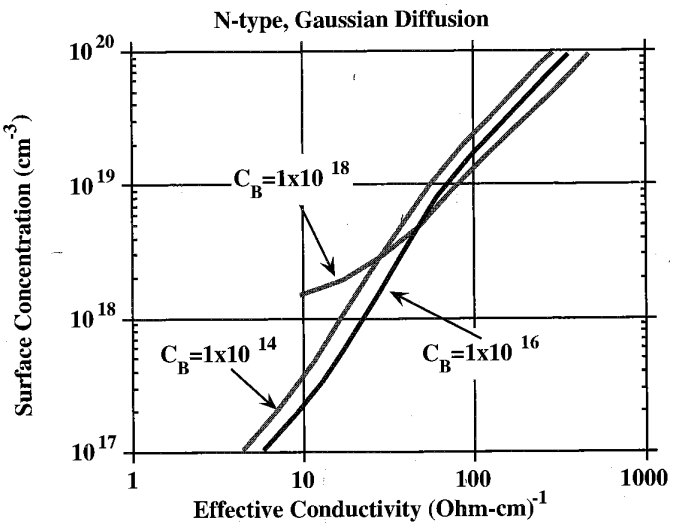
A.7 Color Chart for Thermally Grown SiO_2 Films Observed Perpendicularly under Daylight Fluorescent Lighting [6.7]

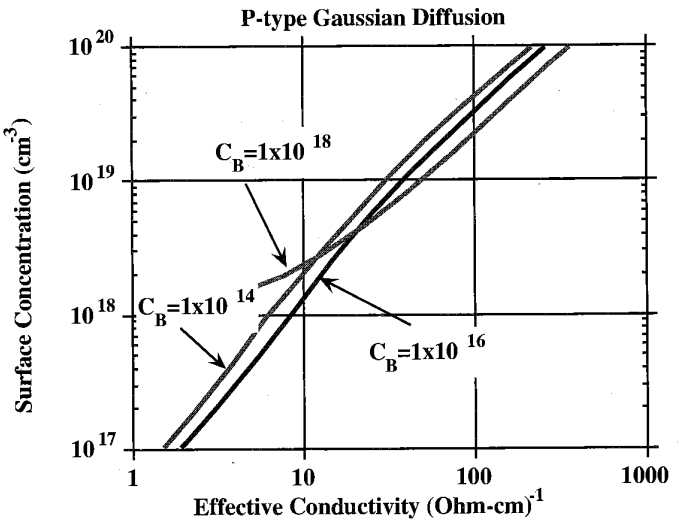
Film Thickness (nm)	Color	Film Thickness (nm)	Color
50	Tan	630	Violet-red
70	Brown	680	"Bluish"
100	Dark violet to red violet	720	Blue green to green
120	Royal blue	770	"Yellowish"
150	Light blue to metallic blue	800	Orange
170	Metallic to light yellow green	820	Salmon
200	Light gold or yellow	850	Light red-violet
220	Gold with slight yellow orange	860	Violet
250	Orange to melon	890	Blue
270	Red-violet	920	Blue-green
300	Blue to blue-violet	950	Dull yellow-green
310	Blue	970	Yellow
320	Blue to blue-green	990	Orange
340	Light green	1000	Carnation pink
350	Green to yellow-green	1020	Violet-red
360	Yellow-green	1050	Red-violet
370	Green-yellow	1060	Violet
390	Yellow	1070	Blue-violet
410	Light orange	1100	Green
420	Carnation pink	1110	Yellow-green
440	Violet-red	1120	Green
460	Red-violet	1180	Violet
470	Violet	1190	Red-violet
480	Blue-violet	1210	Violet-red
490	Blue	1240	Carnation pink to salmon
500	Blue-green	1250	Orange
520	Green	1280	"Yellowish"
540	Yellow-green	1320	Sky blue to green-blue
560	Green-yellow	1400	Orange
570	Yellow to "yellowish"	1450	Violet
580	Light orange or yellow to pink	1460	Blue-violet
600	Carnation pink	1500	Blue

A.8 Irvin's Curves






N-type, Gaussian Diffusion
Surface Concentration (cm-3)
10 20
10 19
10 18
10 17
CB=1x10 18
CB=1x10 14
CB=1x10 16
1
10
100
1000
Effective Conductivity (Ohm-cm)-1


P-type Gaussian Diffusion
Surface Concentration (cm-3)
10 20
10 19
10 18
10 17
CB=1x10 18
CB=1x10 14
CB=1x10 16
1
10
100
1000
Effective Conductivity (Ohm-cm)-1

A.9 Error Function

Properties

The error function is defined by

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp(-\eta^2) d\eta$$

so that

$$\operatorname{erfc}(z) = 1 - \operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_z^\infty \exp(-\eta^2) d\eta$$

The error function is an odd function:

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

which also means

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

The limit as $z \to \infty$ is

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

which also means

$$\operatorname{erf}(-\infty) = -1$$

The derivative of the error function is

$$\frac{d}{dz} \operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \exp(-z^2)$$

A useful approximation with better than 4% accuracy:

$$\operatorname{erfc}(z) \approx \frac{1.1 \exp(-z^2)}{(|z| + \sqrt{z^2 + 1.21})}$$

(from the NBS Applied Math Series 55–1964, "Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables.")

Asymptotic approximations:

$$\operatorname{erf}(z) \to \frac{\exp(-z^2)}{z\sqrt{\pi}} \text{ for } z \gg 1$$

$$\operatorname{erf}(z) \to \frac{2z}{\sqrt{\pi}} \text{ for } z \ll 1$$

z	erf(z)	z	erf(z)	z	erf(z)
0.00	0.00	0.34	0.369365	0.68	0.663782
0.01	0.011283	0.35	0.379382	0.69	0.670840
0.02	0.022565	0.36	0.389330	0.70	0.677801
0.03	0.033841	0.37	0.399206	0.71	0.684666
0.04	0.045111	0.38	0.409009	0.72	0.691433
0.05	0.056372	0.39	0.418739	0.73	0.698104
0.06	0.067622	0.40	0.428392	0.74	0.704678
0.07	0.078858	0.41	0.437969	0.75	0.711156
0.08	0.090078	0.42	0.447468	0.76	0.717537
0.09	0.101281	0.43	0.456887	0.77	0.723822
0.10	0.112463	0.44	0.466225	0.78	0.730010
0.11	0.123623	0.45	0.475482	0.79	0.736103
0.12	0.134758	0.46	0.484655	0.80	0.742101
0.13	0.145867	0.47	0.493745	0.81	0.748003
0.14	0.156947	0.48	0.502750	0.82	0.753811
0.15	0.167996	0.49	0.511668	0.83	0.759524
0.16	0.179012	0.50	0.520500	0.84	0.765143
0.17	0.189992	0.51	0.529244	0.85	0.770668
0.18	0.200936	0.52	0.537899	0.86	0.776100
0.19	0.211840	0.53	0.546464	0.87	0.781440
0.20	0.222703	0.54	0.554939	0.88	0.786687
0.21	0.233522	0.55	0.563323	0.89	0.791843
0.22	0.244296	0.56	0.571616	0.90	0.796908
0.23	0.255023	0.57	0.579816	0.91	0.801883
0.24	0.265700	0.58	0.587923	0.92	0.806768
0.25	0.276326	0.59	0.595936	0.93	0.811564
0.26	0.286900	0.60	0.603856	0.94	0.816271
0.27	0.297418	0.61	0.611681	0.95	0.820891
0.28	0.307880	0.62	0.619411	0.96	0.825424
0.29	0.318283	0.63	0.627046	0.97	0.829870
0.30	0.328627	0.64	0.634586	0.98	0.834231
0.31	0.338908	0.65	0.642029	0.99	0.838508
0.32	0.349126	0.66	0.649377		
0.33	0.359279	0.67	0.656628		

z	erf(z)	z	erf(z)	z	erf(z)
1.00	0.842701	1.34	0.941914	1.68	0.982493
1.01	0.846811	1.35	0.943762	1.69	0.983153
1.02	0.850838	1.36	0.945561	1.70	0.983790
1.03	0.854784	1.37	0.947312	1.71	0.984407
1.04	0.858650	1.38	0.949016	1.72	0.985003
1.05	0.862436	1.39	0.950673	1.73	0.985578
1.06	0.866144	1.40	0.952285	1.74	0.986135
1.07	0.869773	1.41	0.953852	1.75	0.986672
1.08	0.873326	1.42	0.955376	1.76	0.987190
1.09	0.876803	1.43	0.956857	1.77	0.987691
1.10	0.880205	1.44	0.958297	1.78	0.988174
1.11	0.883533	1.45	0.959695	1.79	0.988641
1.12	0.886788	1.46	0.961053	1.80	0.989091
1.13	0.889971	1.47	0.962373	1.81	0.989525
1.14	0.893082	1.48	0.963654	1.82	0.989943
1.15	0.896124	1.49	0.964898	1.83	0.990347
1.16	0.899096	1.50	0.966105	1.84	0.990736
1.17	0.902000	1.51	0.967277	1.85	0.991111
1.18	0.904837	1.52	0.968413	1.86	0.991472
1.19	0.907608	1.53	0.969516	1.87	0.991821
1.20	0.910314	1.54	0.970586	1.88	0.992156
1.21	0.912956	1.55	0.971623	1.89	0.992479
1.22	0.915534	1.56	0.972628	1.90	0.992790
1.23	0.918050	1.57	0.973603	1.91	0.993090
1.24	0.920505	1.58	0.974547	1.92	0.993378
1.25	0.922900	1.59	0.975462	1.93	0.993656
1.26	0.925236	1.60	0.976348	1.94	0.993923
1.27	0.927514	1.61	0.977207	1.95	0.994179
1.28	0.929734	1.62	0.978038	1.96	0.994426
1.29	0.931899	1.63	0.978843	1.97	0.994664
1.30	0.934008	1.64	0.979622	1.98	0.994892
1.31	0.936063	1.65	0.980376	1.99	0.995111
1.32	0.938065	1.66	0.981105		
1.33	0.940015	1.67	0.981810		

z	erf(z)	z	erf(z)	z	erf(z)
2.00	0.995322	2.34	0.999065	2.68	0.999849
2.01	0.995525	2.35	0.999111	2.69	0.999858
2.02	0.995719	2.36	0.999155	2.70	0.999866
2.03	0.995906	2.37	0.999197	2.71	0.999873
2.04	0.996086	2.38	0.999237	2.72	0.999880
2.05	0.996258	2.39	0.999275	2.73	0.999887
2.06	0.996423	2.40	0.999312	2.74	0.999893
2.07	0.996582	2.41	0.999346	2.75	0.999899
2.08	0.996734	2.42	0.999379	2.76	0.999905
2.09	0.996880	2.43	0.999411	2.77	0.999910
2.10	0.997021	2.44	0.999441	2.78	0.999916
2.11	0.997155	2.45	0.999469	2.79	0.999920
2.12	0.997284	2.46	0.999497	2.80	0.999925
2.13	0.997407	2.47	0.999523	2.81	0.999929
2.14	0.997525	2.48	0.999547	2.82	0.999933
2.15	0.997639	2.49	0.999571	2.83	0.999937
2.16	0.997747	2.50	0.999593	2.84	0.999941
2.17	0.997851	2.51	0.999614	2.85	0.999944
2.18	0.997951	2.52	0.999635	2.86	0.999948
2.19	0.998046	2.53	0.999654	2.87	0.999951
2.20	0.998137	2.54	0.999672	2.88	0.999954
2.21	0.998224	2.55	0.999689	2.89	0.999956
2.22	0.998308	2.56	0.999706	2.90	0.999959
2.23	0.998388	2.57	0.999722	2.91	0.999961
2.24	0.998464	2.58	0.999736	2.92	0.999964
2.25	0.998537	2.59	0.999751	2.93	0.999966
2.26	0.998607	2.60	0.999764	2.94	0.999968
2.27	0.998674	2.61	0.999777	2.95	0.999970
2.28	0.998738	2.62	0.999789	2.96	0.999972
2.29	0.998799	2.63	0.999800	2.97	0.999973
2.30	0.998857	2.64	0.999811	2.98	0.999975
2.31	0.998912	2.65	0.999821	2.99	0.999976
2.32	0.998966	2.66	0.999831	3.00	0.999978
2.33	0.999016	2.67	0.999841		

A.10 List of Important Symbols

A—area, in units of cm^2
C—capacitance, in units of F or farads
C—concentration, in units of cm^{-3}
D—diffusivity (experimentally measured), in units of $cm^2\ sec^{-1}$
d—diffusivity of mobile species, in units of $cm^2\ sec^{-1}$
E—energy, in units of eV
F—flux, in units of $cm^{-2}\ sec^{-1}$
F_{min}—minimum feature size, in units of nm
I—current, in units of amps
I—optical intensity in units of $J\ cm^{-2}\ sec^{-1}$
I—interstitial (Si interstitial in Si crystal, unless otherwise designated)
I—number of impurities in crystal growth model
J—current density, in units of amps cm^{-2}
m—mass, in units of gm
M—particle or atomic mobility (D/kT), in units of $cm^2\ eV^{-1}\ sec^{-1}$
n—electron concentration, in units of cm^{-3}
n—index of refraction, unitless
N–density, in units of atoms cm^{-3}
P -pressure, in units of atm or torr
Q—dose, in units of cm^{-2}
R—resistance, in units of ohms or Ω
t—time, in units of sec
T—temperature, in units of K (unless noted by °C)
v—velocity, in units of cm sec^{-1}
v—deposition or etch rate (velocity of surface), in units of cm sec^{-1}
V—voltage or potential difference, in units of volts
V—volume, in units of cm^3
V—vacancy (in Si unless otherwise designated)
$\mathcal{E}$—electric field, in units of V cm^{-1}
ε—permittivity, in units of F cm^{-1}
μ—carrier mobility , in units of $cm^2\ volt^{-1}\ sec^{-1}$
σ—conductivity, in units of S m^{-1}
σ—stress, in units of dynes cm^{-2} or Pa
τ—lifetime, in units of sec
ρ—electrical resistivity, in units of ohm cm
ρ_S—sheet resistance, in units of ohms per square
ρ_C—specific contact resistivity, in units of ohm cm^2

A.11 List of Common Acronyms

AC—alternating current
A/C—amorphous/crystalline interface in Si
AES—Auger electron spectroscopy
AFM—atomic force microscopy
APCVD—atmospheric pressure chemical vapor deposition
AR—aspect ratio
ARC—antireflection coating
ARDE—aspect ratio dependent etching

BCC—body-centered cubic unit cell
BICMOS—bipolar and CMOS devices on the same chip
BJT—bipolar junction transistor
BOE—buffered oxide etch (same as BHF, buffered HF)
BPSG—borophosphosilicate glass

CVD–chemical vapor deposition
CZ—Czochralski crystal growth method
DC—direct current
DI—deionized, or very pure, water
DLTS—deep level transient spectroscopy
DNQ—diazonaphthoquinone, a resist material
DOF—depth of focus
DOP—degree of planarization
DRAM—dynamic random-access memory
DUV—deep ultraviolet

E-beam—electron beam
EBSD—electron back-scattering diffraction
ECR—electron cyclotron resonance
EGS—electronic grade silicon
EM—electromigration
EMP—electron microprobe
EOR—end of range defects
ERF—error function
ERFC—complementary error function
EUV—extreme ultraviolet

FC—float-zone crystal growing technique
FCC—face-centered cubic unit cell
FET—field effect transistor
FIB—focused ion beam
FTIR—Fourier transfrom infrared spectroscopy
FZ—float-zone crystal growth method

GOI—gate oxide integrity

HDP—high-density plasma
HEPA—high-efficiency particulate air (filter)
HF—high frequency
HMDS—Hexamethyldisilane, an adhesion promoter

IC—integrated circuit
ICP—inductively coupled plasma
IMD—intermetal dielectric
IMP—ionized metal deposition (same as IPVD)
INSOL—insoluble portion of polmer base in CA resist
IPVD—ionized PVD (same as IMP)

LDD—lightly doped drain
LF—low frequency
LLS—localized light scatterer
LOCOS—local oxidation of silicon
LPCVD—low-pressure chemical vapor deposition
LSS—implant range theory developed by Lindhard, Scharff, and Schiott

MBE—molecular beam epitaxy
MGS—metallurgical grade silicon
MOS—metal oxide semiconductor
MTF—modulation transfer function
MTTF—median time to failure

NMOS—N channel MOS transistor
NPN—bipolar transistor with N-type emitter and collector
NTD—neutron transmutation doping
NTRS—National Technology Roadmap for Semiconductors

OISF—oxidation-induced stacking fault
OPC—optical phase correction
OPD—optical path difference

PAC—photoactive compound
PAG—photo-acid generator
PEB—postexposure bake
PECVD—plasma-enhanced chemical vapor deposition
PMOS—P channel MOS transistor
PSG—phosphosilicate glass
PSM—phase shift mask
PVD—physical vapor deposition

QBD—charge to breakdown

RBS—Rutherford backscattering
RCA—a wafer cleaning process
RF—radio frequency, usually 13.56 Hhz
RIE—reactive ion etching
RTA—rapid thermal annealing
RTO—rapid thermal oxidation

SACVD—subatmospheric CVD
SC-1, SC-2—wafer cleaning solutions, part of RCA cleaning process
SCM—scanning capacitance microscope
SEM—scanning electron microscopy
SIA—Semiconductor Industry Association
SIMOX—process for SOI, named for "separation by implanted oxygen"
SIMS—secondary ion mass spectrometry
SMIF—standard mechanical interface (box)
SOG—spin-on glass
SOD—spin-on dielectric
SOI—silicon on insulator
SOL—soluble portion of polmer base in CA resist
SPEEDIE—Stanford Profile Emulator for Etching and Deposition in IC Engineering
SRH—Shockley, Read, Hall recombination
STM—Scanning tunneling microscope
SUPREM—Stanford University Process Engineering Models

TCA—trichloroethane
TCAD—technology computer-aided design
TED—transient-enhanced diffusion
TEM—transmission electron microscopy
TEOS—tetraethooxysilane, $Si(OC_2H_5)_4$
TXRF—Total X-ray fluorescence

ULSI—ultralarge-scale integration—usually $> 10^6$ components
UV—ultraviolet light, light with a wavelength of about 0.2 – 0.4μ

VLSI—Very large scale integration—usually $> 10^5$ components

XES—X-ray electron spectroscopy
XPS—X-ray photo-electron spectroscopy
XRF—X-ray fluorescence
XRD—X -ray diffraction

A.12 Tables in Text

1.1 SIA Roadmap for silicon technology—circuit requirements
1.2 Bandgaps and intrinsic carrier concentrations
3.1 Defect etches for silicon
3.2 Solid/liquid segregation coefficients for dopants in silicon
3.3 Energy levels of V and I in silicon
3.4 Point-defect concentrations in silicon
4.1 SIA Roadmap for silicon technology—defect requirements
4.2 SIA Roadmap for silicon technology—yield requirements
4.3 Oxidation-reduction reactions and potentials
5.1 SIA roadmap for silicon technology—lithography requirements
6.1 SIA roadmap for silicon technology—dielectrics requirements
6.2 Silicon oxidation B and B/A rate constants
6.3 Silicon oxidation rate constants for parallel oxidation model
6.4 Typical silicon oxidation parameters in SUPREM IV
7.1 SIA Roadmap for silicon technology—doping requirements
7.2 Implantation versus solid-gas-phase doping
7.3 Intrinisic diffusivities of dopants in silicon
7.4 Furnace and RTA diffusion and annealing characteristics
7.5 Diffusivity parameters of dopants in silicon
7.6 Values of f_I and f_V for silicon and dopants in silicon

9.1 Common deposition methods for thin films
9.2 Typical modeling parameter values for various deposition techniques
10.1 Common wet chemical etchants
10.2 Boiling points of typical etch products
10.3 Common plasma etch gases for films used in IC fabrication
11.1 SIA Roadmap for silicon technology—back-end requirements
11.2 Resistivities and melting points of interconnect materials
11.3 Barrier layer properties
11.4 Mechanical properties of interconnect materials
11.5 Properties of silicides
11.6 Materials for interlevel dielectrics
11.7 Silicidation *B* and *B/A* rate constants used in simulations
11.8 Cu reflow simulation parameters

A.13 Answers to Selected Problems

1.3. 4.6pA.
1.4. $2.15 \times 10^5\ \Omega$cm at room temperature.
1.10. 4.2×10^{-17} Farads.
3.1. $148°C cm^{-1}$.
3.2. 0.109 gm.
3.3. f = 0.995.
3.4 $3.82 \times 10^{14}\ cm^{-3}$.
4.1. 7.7%.
5.3. 0.15 μm.
5.4 $\lambda = 0.4\ \mu m$
6.15. $x_{<100>} = 0.314\ \mu m, x_{<111>} = 0.346\ \mu m, t_{<100>} = 1.09$ hours, $t_{<111>} = 1.04$ hours
6.16. t = 1.66 hours
6.18. t = 2.25 hours.
7.2 16.66 μm.
7.3. (a) $1.75 \times 10^{14}\ cm^{-2}$ (b) 0.44 μm (c) 75.75Ω/square
7.9. 0.72 μm.
7.17. 7.09.
8.2. $Q = 1 \times 10^{14}\ cm^{-2}$.
8.4 $R^*_P + 4.3\Delta R^*_P$.
8.8 0.58 μm.
8.15. 94.4%.

9.6 1.4 $\mu m\ min^{-1}$.

9.9 0.94 $\mu m\ min^{-1}$

9.11 (a) 433 cm, 1.2×10^{-2} collisions; (b) 1.44 cm, 3.5 collisions; (c) 0.013 cm, 392 collisions; (d) 1.7×10^{-5} cm, 3.0×10^{5} collisions

10.2. 0.5 μm, 0.125 μm.

10.9. Chemical component = 0.55, Physical component = 0.45.

10.10. $2.5 \times 10^{18}\ cm^2\ sec^{-1}$

11.5. $1.3 \times 10^{20}\ cm^{-3}$

11.8 2.5 min.

Index

A

Abnormal, or secondary, grain growth, 761
Activation, 484
Active regions:
  formation of, 52–56
Adhesion, 737
Aerial image, 251
AES (Auger Electron Spectroscopy), 174, 176, 651
Airy's disk, 211–13, 215
Al
  deposition, 565–66
  plasma etching, 649–50
  reflow, 705
  spiking, 692
Alkali ion contamination, 38, 153, 310
Alloy junction bipolar transistors, 8–9
Amorphous layer formation, 479
Amorphous layer regrowth, 482
Anisotropic arrival angle distribution, 544
Anisotropic etching, 75–76, 611
Anisotropy, 75–76, 611, 616
Antimony, 127, 452
Antireflection Coating (ARC), 230, 281
APCVD, *See* Atmospheric Pressure Chemical Vapor Deposition (APCVD)
Application-Specific Integrated Circuits (ASICs), 2
Arrival angle distribution, 543, 567, 587, 654
Arsenic, 72–74, 344, 477–78, 487
Arsine, 561
Aspect Ratio (AR), 510
Aspect Ratio Dependent Etching (ARDE), 641, 663
ATHENA (Silvaco), 272, 274, 312, 345, 347, 403, 598–600, 653, 671–74, 737, 744–45, 753
Atmospheric Pressure Chemical Vapor Deposition (APCVD), 512, 513–25
Atomic Force Microscopy (AFM), 299, 726
Atomic Layer Deposition (ALD), 601
Auger Electron Spectroscopy (AES), 174, 176, 651
Autodoping, 522–25

B

Back-end technology, 681–786
  basic concepts, 687–715
  contacts, 688–95
  defined, 681
  dielectrics, 707–15
  historical development, 687–88
  interconnects/vias, 695–707
  manufacturing methods/equipment, 715–25
    contact formation, 719–21
    first-level dielectric processing, 718–19
    global interconnects, 721–22
    IMD deposition and planarization, 723–24
    silicided gates and source/drain regions, 716–18
    via formation, 724
  measurement methods, 725–37
    chemical and structural measurements, 732–34
    electrical measurements, 726–32
    mechanical measurements, 734–37
    morphological measurements, 726
  models/simulation, 737–76
    Chemical-Mechanical Polishing (CMP), 744–46
    diffusion in polycrystalline materials, 762–65
    electromigration, 765–76
    grain growth, 753–62
    reflow, 746–53
    silicide formation, 738–44
  technologies/models, limits and future trends in, 776–80
Band model, 18–21
  introduction of dopants, 19
Bandgap temperature dependence, 27
Barrel etcher, 628
Barrier layers:
  properties, 696
  refractory metal compounds as, 695
  refractory metal silicides as, 694
Base collector junction, 40
Bell Telephone Laboratories, 7–9
BESOI, 145–46

Bias-sputter deposition, 530, 549–51
Bias sputtering, 549–51
Bias Temperature Stressing, 311
BICMOS technology, 42–43
BICs (Boron Interstitial Clusters), 497
Bipolar Junction Transistors (BJTs), 4, 39–41
Bird's beak effect, 292
Blanket tungsten deposition, 567
Bleaching, 230
Body-Centered Cubic (BCC) cells, 94
Boltzmann distribution function, 25
Bonded and etch-back technology for SOI (BESOI), 145
Bond model, 18–19
  introduction of dopants, 19
Boron, 9, 61–62, 72, 127, 414–15, 428–29, 452, 469, 478
Boron Interstitial Clusters (BICs), 497
Boron TED, 497
Bosch process, 675
BPSG (borophosphosilicate glass), 85–86, 709, 718
Brandt-Kitagawa effective charge, 473
Brownian motion, 157–58
Buffered HF (BHF)/Buffered Oxide Etch (BOE), 612, 618, 733
Buried layers, 65–70

C

Capacitance-Voltage (CV) method, 301, 399
Carbon in silicon, 142
Carrier lifetime, 31, 177
Cathode dark space, 541
Cathode sheath, 541
Cesium, 396
Channeling, 461
Charge to breakdown, 311
Charged point defects, 134–35, 436
Chemical cleaning, 159, 184
Chemical etching, in plasma systems, 622–25
Chemical-Mechanical Polishing (CMP), 58, 86–87, 107, 279, 570, 675, 713–14, 744–46
Chemical Vapor Deposition (CVD), 102–5, 289, 511–30, 699
  Atmospheric Pressure Chemical Vapor Deposition (APCVD), 512, 513–25
  High-Density Plasma Chemical Vapor Deposition (HDPCVD), 530, 551, 564, 586–87
  Low-Pressure Chemical Vapor Deposition (LPCVD), 512, 525–27, 582–83
  Plasma-Enhanced Chemical Vapor Deposition (PECVD), 512, 527–30, 584–87
  sub-atmospheric CVD (SACVD), 564
Chemically amplified (CA) resist, 225
Child's law, 468
Chip size, 1–3
Chip yield, 182
Chlorosilane, 102
CIM (Computer Integrated Manufacturing) tools, 180
Clean factories, 157–59, 165, 169–73, 181–84
Clean rooms, 193–95
Cleaning chemistry, 159, 184
Climb process, 99–100
CMOS technology, 12, 42–43, 49–92
  process flow, 50–90
    active region formation, 52–56
    buried and epitaxial layers, 65–70
    field implants under LOCOS regions, 63–65
    gate formation, 71–76
    N and P well formation, 60–63
    process options for active region and well formation, 63–70
    shallow trench isolation, 57–60
    source/drain formation, 80–82
    substrate selection, 51–52
    tip/extension (LDD) formation, 76–80
CMOS technology, process flow:
  contact and local interconnect formation, 82–84
    multilevel metal formation, 84–90
CMP, *See* Chemical–Mechanical Polishing (CMP)
Cold-wall reactors, 512
Collimated sputter deposition, 551–54
Color charts, 300 + appendix page
Complementary MOS, *See* CMOS technology
Computer-Aided Design (CAD), 6, 203
Concentration dependent diffusion, 409, 436
Conductivity, 371
Conformal coverage, 509, 589
Contact formation, 719–21
Contact printing, 208
Contacts, 681, 688–95
  resistance, 727
  resistivity, 689
  spiking, 692
Contaminant, 152, 157, 173
Cosine arrival angle distribution, 544
Cosine emission, or diffuse emission, 534
Covalent bonding, 14
Critical area, chips, 152
Critical Modulation Transfer Function (CMTF), 227–28
Critical particle size, 153
Crookes dark space, 541
Cross-bridge Kelvin structure, 728–29
Crystalline, 13
Crystalline precipitate, 101
Crystalline semiconductor materials, 13
Crystals:
  carbon in silicon, 142
  defects in, 97–101

diameters, 104
manufacturing methods/equipment, 109–11
measurement methods, 111–21
defect etches, 117–18
electrical measurements, 111–12
electron microscopy, 119–21
Fourier transform infrared spectroscopy (FTIR), 118–19
Hall effect measurements, 115–17
hot point probe, 112–13
physical measurements, 117
sheet resistance, 113–15
models/simulation, 121–44
Czochralski (CZ) crystal growth, 122–28
point defects, 131–38
zone refining and FZ growth, 128–31
oxygen in silicon, 138–42
pulling apparatus, 110
raw materials and purification, 101–2
structure of, 94–97
technologies/models, limits and future trends in, 144–46
volume defects in, 101
wafer preparation/specification, 105–9
Cu deposition, 570–71
Cu interconnects, 779
Current crowding, 727
Curvature (surface), 748
CVD, *See* Chemical Vapor Deposition (CVD)
CV method, 301
CV profiling, 399
Czochralski (CZ) crystal growing apparatus, 103
Czochralski (CZ) crystal growth, 122–25, 537
dopant incorporation during, 125–28
Czochralski method, 97, 102

D

Damage annealing, 479–82
Damage production, 476–79
Damascene process, 87, 704, 724
dual 724
Dash etch, 118
DC sputter deposition, 539–46
Deal-Grove model, 267, 293, 312–13, 322–29, 344–45, 349, 515, 741
Deep Level Transient Spectroscopy (DLTS), 180, 488
Deep P+ diffusions, 41
Deep UV (DUV), 222
resists, 225–26
Defect density numbers, 152–53
Defect etches, 117–18
Degenerate semiconductors, 26
Degree of anisotropy, 616
Degree Of Planarization (DOP), 703–4
Deionized water (DI), 172
Denuded zone, 162, 168, 170
DEPICT (Avant!), 247
Deposition, 509–608
basic concepts, 511–54
Chemical Vapor Deposition (CVD), 511–30
Atmospheric Pressure Chemical Vapor Deposition (APCVD), 512, 513–25
High-Density Plasma Chemical Vapor Deposition (HDPCVD), 530, 586–87
Low-Pressure Chemical Vapor Deposition (LPCVD), 512, 525–27, 582–83
Plasma-Enhanced Chemical Vapor Deposition (PECVD), 512, 527–30, 584–87
common deposition methods (table), 555
historical development, 511–12
manufacturing methods, 554–71
Al deposition, 565–66
Cu deposition, 570–71
epitaxial Si deposition, 556–58
polycrystalline Si deposition, 558–61
SiO_2 deposition, 563–65
Si_3N_4 deposition, 561–63
Ti deposition, 566
TiN deposition, 568–69
$TiSi_2$ deposition, 567–68
Ti-W deposition, 566
W deposition, 567
WSi_2 deposition, 567–68
measurement methods, 572
models, 573–600
for deposition simulations, 573–74
in physically based simulators, 574–81
Physical Vapor Deposition (PVD), 511–12, 530–54
evaporation method, 511, 531–39
sputter deposition method, 511, 539–54
simulations, 590–601
of deposition using a physically based simulator, 590–98
step coverage, 509
technologies/models, limits/future trends in, 601–2
Depth of focus, 214
Devitrification, 292
Diazoquinone molecules, 224
DIBL (Drain-Induced Barrier Lowering), 373, 499
Dibroane, 561
Dichlorosilane, 102
Dielectrics, 707–15
first level, 681, 707, 718
intermetal dielectrics (IMDs), 681, 707–8, 723
Diffraction, 209–12
Diffusion, 371–451
basic concepts, 374–92

diffused layers, design/evaluation of, 389–92
dopant solid solubility, 375–77
error-function solution near a surface, 384–85
from a macroscopic viewpoint, 377–79
Gaussian solution:
in an infinite median, 380–81
near a surface, 381–82
historical development, 374–92
intrinsic diffusion coefficients of dopants in silicon, 386–88
manufacturing methods and equipment, 392–95
measurement options, 395–403
capacitance voltage, 399
inverse electrical measurements, 402–3
sheet resistance, 398–99
SIMS (Secondary Ion Mass Spectrometry), 396–97
spreading resistance, 397–98
TEM cross section, 399–400
2D electrical measurement using scanning probe microscopy, 400–402
models/simulations, 403–39
activation energy for self-diffusion and dopant diffusion, 426
charge state effects, 436–39
chemical equilibrium formulation for dopant-defect interactions, 432–34
dopant-defect interactions, 426–32
dopant diffusion and I and V, 422–26
Fick's laws modifications to account for concentration-dependent diffusion, 409–13
Fick's laws modifications to account for electric field effects, 406–9
interfacial dopant pileup, 415–17
numerical solutions of the diffusion equation, 403–5
oxidation-enhanced or -retarded diffusion, 419–22
physical basis for diffusion at atomic scale, 417–19
segregation, 413–15
simplified expression for modeling, 434–36
successive diffusion steps, effect of, 388–89
technologies/models, limits and future trends in, 439–42
Dill model, 267–69
Dimethylaluminumhydride (DMAH), 566
Direct bandgap semiconductor, 30
Direct fluxes, topography models, 575–77, 654
Dishing, 745
Dislocation free wafers, 98
Dislocation loops, 99, 481
Dislocations, 98–100, 110
Dissociative mechanism, 187
DLTS (Deep Level Transient Spectroscopy), 180, 488
DNQ resists, 240–41, 259–64
g-line/i-line, 259–63
Dopant activation, 484–86
Dopants, 13–17
heavy doping effects, 21–22
shallow donors, 19–20
Double crucible method, 144
Drain formation, 80–82
DRAM, 155–56, 182, 201
Drive-in diffusion, 375
Dry etching
See Plasma etching (or etching plasma)
Dual damascene process, 724
Dual diffusion mechanism, 422
Dual Pearson distribution, 461–63
DUV resist, 225
Dynamic random access memory, *See* DRAM

E

Edge bead, 240
Effective diffusivity, 434
Electric field effects, 406
Electrical measurements, of silicon wafers, 111–12
Electroless deposition, 570–71
Electromigration, 88–89, 696–98, 765–76
Electron Back-Scattering Diffraction (EBSD), 733
Electron beam lithography, 273–75
Electron Cyclotron Resonance (ECR), 530, 632–33
Electron microscopy, 119–21
Electron Spectroscopy for Chemical Analysis (ESCA), 732–33
Electronic Grade Silicon (EGS), 102
Electronic stopping, 273–75
Electroplating, 570
Elemental semiconductors, 14
Ellipsometer, 45, 300, 572, 733
Emission, 514
Emission flux, 578
Emitted flux, physically based topography models, 579–80
End-Of-Range (EOR) defects, 481, 484
Enhancement/Depletion (E/D), 42
Epitaxial Lateral Overgrowth (ELO), 146
Epitaxial layers, 65–70
Epitaxial silicon deposition, 519, 556–58
Erosion, 746
Error function, 382–85
Etchback planarization, 711
Etching, 609–81
anisotropic etching, 611
anisotropy, 611, 616
basic concepts, 612–37
bias, 613
directionality, 610–11
dry etching, *See* Plasma etching
historical development, 612–37
isotropic etching, 611, 616

manufacturing methods, 637–50
plasma etching aluminum, 649–50
plasma etching conditions/issues, 638–43
plasma etching polysilicon, 647–49
plasma etching silicon dioxide, 644–47
plasma etch methods for various films, 643–44
measurement methods, 650–53
models:
for etching simulation, 653–56
linear etch model, 656–63
saturation/adsorption model for ion-enhanced etching, 663–69
plasma etching, 619–37
in barrel etchers, for, 628–30
chemical etching, 622–25
in high-density plasma systems, 632–34
ion-enhanced etching, 622, 626–28
mechanisms, 621–28
in parallel plate systems—plasma mode, 630–31
in parallel plate systems—reactive ion etching (RIE) mode, 630–31
physical etching, 625–26
sputter etching and ion milling, 634–36
system design, 619–20
selectivity, 609–11, 613
uniformity, 641–42
vertical etching, 611, 666
wet etching, 609, 612–19
Buffered HF (BHF)/Buffered Oxide Etch (BOE), 612, 618
common etchants (table), 619
etch bias, 613
etch rates, 618–19
etch thickness/depth, 613–14
overetching, 614–15
selectivity, 612–13
Etchstop layer, 724
Evaporation method, PVD, 511, 531–39
EVOLVE, 573
Exposure field requirements, lithography, 201
Extension (LDD) formation, 76–80
Extreme Ultraviolet (EUV) lithography systems, 202–3, 277
Extrinsic effects, 21–30
in diffusion, 406–13
in oxidation, 343
Extrinsic gettering sites, 164
Extrusion fill, 554

F

Face-Centered Cubic (FCC) cells, 94–95
Far field diffraction, 212–18
the Huygens-Fresnel principle, 210–11
Fast diffusers, 386
Fermi-Dirac distribution function, 26
Fermi-Dirac probability function, 23, 25
Fermi level, 23–25
Fick's laws, 314, 378–79, 392, 403–417, 429–441
first law, 378, 520
modifications:
to account for concentration-dependent diffusion, 409–13
to account for electric field effects, 406–9
second law, 378–80
Field implants:
defined, 64
under LOCOS regions, 63–65
First-level dielectric processing, 707, 718–19
Fixed oxide charge, 294
Flat band voltage, 309
Float-zone (FZ) crystal growth process, 102–5, 128–31
FLOOPS process simulator, 739
Focused Ion Beam (FIB) milling machines, 399
Focused Ion Beam (FIB) technique, 636
ForceFill technique, 554
Fourier transform infrared spectroscopy (FTIR), 118–19, 732
Four-point probe measurement method, 113, 115, 373
Frank-Turnbull mechanism, 187
Fraunhofer diffraction, 212–18
Frenkel pairs, 476, 479
Frenkel process, 132
Fresnel diffraction, 212, 219–21, 247
FTIR, *See* Fourier transform infrared spectroscopy (FTIR)
Fully coupled diffusion, 426–31
Fully kinetic diffusion, 440–42

G

Gap filling, 509
Gas Immersion Laser Doping (GILD), 440
Gate formation, 71–76
Gate insulator, 288–89
Gate Oxide Integrity (GOI), 153
Gaussian profiles, 380–82, 452–57
Gettering, 156, 161–64, 167–69, 176–80, 186–93
metal atoms:
making mobile, 186–87
trapping at the gettering site, 190–93
metal diffusion to gettering site, 187–90
strategies, for alkali ions/metals, 162
Ghez and van der Meulen model, 324, 328
Glide planes, 100
G-line resists, 223–24
Global interconnects, 721–22, 681
Global planarization, 710
Grain boundaries, 753

Grain boundary diffusion, 762–765
Grain grooving, 760
Grain growth, 753–62
  abnormal or secondary, 761
  normal, 758
Graphite, 110
Gravitational sedimentation, 158
GROFILMS, 602
Grown junction bipolar transistors, 8
Growth kinetics:
  dependence on crystal orientation, 329–32
  dependence on pressure, 328–29
  mixed ambient, 332–33

H

Hall effect measurements, 115–17
Hall mobility, 117
Hall scattering factor, 117
Han and Helm model, 323–24, 332
Heavy doping effects, 21–22, 343, 406
Heavy ions, distribution of, 452
Henry's law, 314, 324
HEPA filters, 158, 160, 165
Hexamethyldisilane (HMDS), 239–40
High-Density Plasma Chemical Vapor Deposition (HDPCVD), 58, 530, 551, 564, 586–87, 713, 719
High-Density Plasma (HDP) etch system, 632–634
High Efficiency Particulate Air (HEPA) filters, 158, 160, 165
High-energy implants, 466–68
High-Frequency CV curve, 311
High K dielectrics, 290
High-temperature sputter deposition, 554, 566
Hillock growth, 696–98
Hot electron problems, 77
Hot point probe, 112–13
Hot-sputter deposition, 554, 566
Hot-wall reactors, 512
Huygens-Fresnel principle, 210
Hydrogen passivation, 356–57

I

Ideal cosine emission, 535
Ideal gas law, 314
I-line resists, 223–24
I-line steppers, 207
Implant - see Ion Implantation
Impurities, 97
Indirect fluxes, topography models, 577, 654
Indium, 9, 388, 412
Inductively Coupled Plasma (ICP), 530, 633
Infrared heating, 240
Infrared temperature sensors, 125
Inhibitors for etching, 627, 640
"In situ" doped poly, 74
Integrated circuit (IC):
  components, 1
  historical perspective, 1–46
  increasing complexity of, 1
  and the planar process, 7–13
  power supply level, 76–77
Interconnects
  capacitance, 682
  global, 681, 721–22
  local, 681, 699–702
  resistance, 682
  scaling, 683
Interface charges, 352–357
Interface trapped charge, 294–95
Intermetal dielectrics (IMDs), 681, 707–8, 723
Interstitialcy, 98
Intrinsic doping coefficients, 387–88
Intrinsic electron concentration, 16, 21
Intrinsic Fermi level, 23
Intrinsic gettering, 52, 192, 195
  sites, 164
Intrinsic region, 22
Intrinsic stress, 734
Ion beam heating, 469
Ion-enhanced etching, in plasma systems, 622, 626–28, 663–669
Ion implantation, 15–17, 60–61, 190, 346, 375, 451–508
  basic concepts, 451–63
  historical development, 451–63
  ion implantation, effects on devices, 497–99
  manufacturing methods/equipment, 463–69
    high-energy implants, 466–68
    ion beam heating, 469
    ultralow energy implants, 468–69
  measurement methods, 469–70
  models/simulations, 470–99
    damage annealing, 479–82
    damage production, 476–79
    dopant activation, 484–86
    local electronic stopping, 474–75
    nonlocal electronic stopping, 473–74
    nuclear stopping, 471–73
    solid-phase epitaxy, 482–84
    total stopping powers, 475
    Transient-Enhanced Diffusion (TED), 486–87
  technologies/methods, limits and future trends in, 499–500
Ionized metal plasma (IMP) deposition, 554
Ionized PVD (IPVD), 551–54
Ionized sputter deposition, 551–54
Ion milling, 634–36

Ion pairing model, 191–92
Irvin's curves, 389
Isolation methods, 53–59
Isotropic arrival angle distribution, 544
Isotropic etching, 75–76, 611, 616

K

Kick-out mechanism, 187
Kinchin-Pease formula, 476–77
"Knock on," use of term, 396
Kohler illumination, 236

L

Langmuir adsorption, 664
Langmuir-Knudsen theory, 536
Latchup, 65, 467
Latent image, 253–59
Lateral straggle, 453
Lattice constant, 27, 94
Lattice damage, and sputter etching, 635
Lattice defects, 33
Law of mass action, 21
LDD devices, 77–78
Lenz's law, 144
Light Emitting Diodes (LEDs), 30
Light ions, distribution of, 452
Lightly Doped Drain (LDD) devices, 77–78
Linear etch model, 656–63
Linear Parabolic model, 313
Line defects, 98
Linewidth standards, 244
Lithography, 107, 201–87
  advanced mask engineering, 271–72, 277–78
  basic concepts, 203–34
  contact printing, 208
  contact and proximity systems, 219–21, 247
  diffraction, 209–12
  exposure field requirements, 201
  Extreme Ultraviolet (EUV) lithography systems, 202–3, 277
  historical development, 203–34
  light sources, 206–7
  mask engineering, 230–34
    Optical Proximity Correction (OPC), 230–32
    Phase Shift Masks (PSMs), 230, 232–34
  measurement methods, 241–46
    etched features, 244–46
    mask features/defects, 242–44
    resist patterns, 244
  models/simulation, 246–72
    optical intensity pattern in the photoresist, 253–59
    photoresist developing, 267–70
    photoresist exposure, 259–64
    photoresist postbake, 270–71
    postexposure bake (PEB), 264–66
    wafer exposure systems, 247–53
  optical, 202–3
  photoresists, 221–30, 238–41
    basic properties/characterization of, 227–30
    Deep UV (DUV) resists, 225–26
    g-line resists, 223–24
    i-line resists, 223–24
  placement accuracy, 201
  projection printing, 209
  projection systems, 212–18
  proximity printing, 208–9
  ray tracing, 209–12
  resolution requirements, 201
  technologies/models:
    advanced mask engineering, 277–78
    electron beam lithography, 273–75
    limits and future trends in, 272–81
    new resists, 278–81
    X-ray lithography, 275–77
  wafer exposure systems, 208–9, 234–41, 247–53
Loading effect in etching, 641
Local electronic stopping, 474–75
Local interconnects, 75, 82–84, 681, 699–702
Localized light scatterers (LLS), 153
Local Oxidation of Silicon, *See* LOCOS process, 561–62
Local planarization, 710
LOCOS process, 53, 289, 291–92, 333, 348–49, 561–62
  poly-buffered, 55–56
LOCOS regions, 54–59
  field implants under, 63–65
Long throw collimaton, 553
Long throw sputtering, 553
Lot tracking, 108
Low-Frequency CV curve, 306
Low K deposited dielectrics, 290
Low-Pressure Chemical Vapor Deposition (LPCVD), 512, 525–27, 582–83
  systems, 54, 78, 289, 582–83
Low temperature oxide (LTO), 563
LSS range theory, 470

M

Mack's model, 267–269
Macroscopic loading, 641
Magnetic CZ ingots, 144
Magnetron sputter deposition, 551
Mask count, 3
Mask engineering, 230–34
  advanced, 271–72, 277–278

Optical Proximity Correction (OPC), 230–32
Phase Shift Masks (PSMs), 230, 232–34
Masking, 54
Mask inspection, 204
Mask penetration, 455
Mask transmittance, 248
Mass action, law of, 21
MEMS (Micro-Electro-Mechanical Systems), 675
Metal
diffusivities in Si, 163
mechanical properties (table), 697
resistivities (table), 688
vapor pressures, 536
Metal atoms, basic properties of, 163–64
Metal Oxide Semiconductor (MOS) transistors, 4
Metallurgical Grade Silicon (MGS), 101–2
Micralign, 234
Microdiffraction stress analysis, 736
Microscopic or micro-loading, 641
Microtrenching, 635
Microwave heating, 240
Miller indices, 95
Mixed ambient growth kinetics, 332–33
Mobile oxide charge, 295
Modern scientific discovery, 43–45
Modern silicon integrated circuit, schematic cross section of, 12–13
Modulation Transfer Function (MTF), 215–18
Molecular Beam Epitaxy (MBE), 468, 558
Molecular Dynamics (MD) calculations, 602
Molecular flow, 536
Monte Carlo calculations, 602
Monte Carlo simulations, 471
Monte Carlo simulator, 475, 489–90, 654
Moore's law, 1
MOS Capacitance-Voltage measurements (MOS CV), 177
MOS capacitor, 301–12
MOS gate oxide, 72–73
MOS (Metal Oxide Semiconductor), 36–39
Multilevel metal formation, 84–90, 702–3

N

National Technology Roadmap for Semiconductors (NTRS), 2–4, 49, 686
Native point defects, 97–98
National Bureau of Standards, 244
Near field diffraction, 212, 219–21, 247
Nerst-Einstein relationship, 172
Neutron Transmutation Doping (NTD), 131
NMOS technology, 41–42
Nitride growth kinetics, 347
Non-ideal cosine emission, 535
Nonlocal electronic stopping, 473–74
N-type dopants, 17
Nuclear stopping, 471–73
Nucleation surface, 102
Numerical Aperture (*NA*), 213
Numerical solutions of diffusion equations, 403

O

Off-axis illumination, 236–37
Ohmic contacts, 688–92
One-dimensional defects in crystals, 98–99
Open circuit voltage decay method, 178
Optical intensity pattern in the photoresist, 253–59
Optical lithography, 202–3
Optical Path Difference (OPD), 252–53
Optical Proximity Correction (OPC), 230–32, 271
Overall chip size, 1
Overetch, 614
Oxidation, 184, 287–370
basic concepts, 290–96
growth kinetics:
dependence on crystal orientation, 329–32
dependence on pressure, 328–29
mixed ambient, 332–33
models/simulation, advanced point defect based models for oxidation, 39–43
polysilicon oxidation, 345–47
Si_3N_4 growth and oxidation kinetics, 347–50
substrate doping effects, 343–45
2D SiO_2, 333–39
historical development, 290–96
manufacturing methods/equipment, 296–98
measurement methods, 298–312
electrical measurements, 301–12
optical measurements, 299–301
physical measurements, 299
models/simulation, 312–59
complete oxidation module simulation, 357–59
linear parabolic model, 313–22
models for planar oxidation kinetics, 322–26
silicide oxidation, 350–52
Si/SiO_2 interface changes, 352–57
thin oxide SiO_2 growth kinetics, 326–28
Si_3N_4 growth and oxidation kinetics, 347–50
technologies/models, limits and future trends in, 359–61
Oxidation enhanced diffusion (OED), 339, 422
Oxidation furnaces, 297–98
Oxidation Induced Stacking Faults (OISF), 100
Oxidation of silicon nitride, 347
Oxidation Retarded Diffusion (ORD), 339, 422
Oxide trapped charge, 295
Oxygen in silicon, 138–42
diffusivity, 140
oxygen atoms, 139

oxygen donors, formation of, 139
precipitation, 139–40

P

Parallel oxidation model, 323
Particle detection, 157, 169, 181
Passivation, 10, 650, 725
Passive barriers, 693
Pauli exclusion principle, 23
Pearson's distribution, 460
Pellicle, 204
Periodic table, portion relevant to semiconductor materials/doping, 15
Perkin-Elmer Corp., 234
Phase Shift Masks (PSMs), 230, 232–34
Phosphine, 561
Phosphorus, 62, 72–74, 344, 415
implant, 62–63
Phosphosilicate glass (PSG), 85, 162, 709
Photo-Acid Generator (PAG), 222–23
Photoconductive decay, 177
Photolithography, 11–12
ensuring resist adhesion, 238–39
postbake, 241
postexposure bake (PEB), 240–41, 264–66
prebake, 240
spinning the resist onto the wafer, 239–40
Photoresist, 54, 61–63, 71, 84, 154, 221–30, 238–41
basic properties/characterization of, 227–30
Deep UV (DUV) resists, 225–26
g-line resists, 223–24
i-line resists, 223–24
Photoresist developing, 267–70
Photoresist exposure, 259–64
DUV resists, 264
g-line/i-line DNQ resists, 259–63
Photoresist postbake, 270–71
Physical etching, in plasma systems, 625–26
Physical measurements, and characterization of silicon wafers, 117
Physical Vapor Deposition (PVD), 511–12, 522, 530–54
evaporation, 511, 531–39
sputter deposition, 511, 539–54
bias-sputter deposition, 549–51
collimated sputter deposition, 551–54
DC sputter deposition, 539–46
high-temperature deposition, 554
ionized sputter deposition, 551–54
magnetron sputter deposition, 551
reactive sputter deposition, 547
RF sputter deposition, 547–49
Placement accuracy, lithography, 201
Planarization, 85, 710–714
etchback, 711
global, 710
local, 710
Planar process:
and integrated circuits, 7
invention of, 9–10
Planck's law, 207
Plasma, 529, 539, 619
Plasma density, 621, 638
Plasma-Enhanced Chemical Vapor Deposition (PECVD), 512, 527–30, 584–87
coverage/filling, 529–30
reactions/processes in PECVD systems, 529
typical equipment configuration, 528–29
Plasma etching, 76, 619–37
See also etching, plasma
Plasma potential, 541
Plasma sheath, 541, 620
PMMA, 279–81
PMOS technology, 41–42
PN diodes, 33–35
Point defects, 97–98, 131–38, 417, 442
charged, 134
chemical pumping of, 427–28
diffusivities, 133
OED, ORD, 339–43, 422–26
Poly-buffered LOCOS process, 55–56
Polycrystalline, 13
diffusion, 762–65
materials, 93–94
semiconductor materials, 13
Polygranular clusters, 768
Polysilicon, 75, 558–561
deposition, 558–61
doping, 561
oxidation, 345–47
plasma etching, 647–49
resistivity, 561
Postbake, 241
Postexposure bake (PEB), 240–41, 264–66
DUV resists, 266
g-line/i-line DNQ resists, 264–65
Prebake, 240
Predeposition, preferred method for, 374–75
Printing:
contact, 208
projection, 209
proximity, 208–9
Projection aligners, 234
Projection printing, 209
PROLITH (Finle Technologies), 247
Proximity effect, 274

Proximity printing, 208–9
PSG (phosphosilicate glass), 85–86, 162, 709
$PtSi_2$, 693
P-type dopants, 17, 20, 24
Pure silicon, and band model, 19
PVD, *See* Physical Vapor Deposition (PVD)

Q

QCM (Quartz-Crystal Microbalance), 572
Quartz, 109–10

R

Radiation damage, and sputter etching, 635
Rapid Thermal Annealing (RTA), 99, 393–95
Rapid Thermal Oxidation (RTO), 99, 298
Rapid Thermal Processing (RTP), 99
Ray tracing, 209–12
RBS (Rutherford Backscattering Spectrometry), 133, 175, 483, 732
RCA clean, 156–57, 160, 184, 193–94
Reactive Ion Beam Etching (RIBE), 636
Reactive Ion Etching (RIE), 280, 631–32
Reactive sputter deposition, 547
Redeposited flux, topography models, 579
Reemission, 514
Reflow, 86, 709, 718, 746–53
Refractory metal compounds, as barrier layers, 695
Refractory metal contact-adhesion layers, 694
Refractory metal silicides, as barrier layers, 694
Reliability of interconnects, 730–32, 767
Resistivity, 18, 372, 688
Resolution requirements, lithography, 201
Reverse Short-Channel Effect (RSCE), 499
RF sputter deposition, 547–49
RIE, *See* Reactive Ion Etching (RIE)
RIE "lag", 641
RTA, *See* Rapid Thermal Annealing (RTA)
RTO, *See* Rapid Thermal Oxidation (RTO)
RTP, *See* Rapid Thermal Processing (RTP)
Rutherford Backscattering Spectrometry (RBS), 133, 175, 483, 732

S

Sacrificial barriers, 693
Salicide, 568, 700
SAMPLE, 573
Saturation/adsorption model for ion-enhanced etching, 663–69
Saucer pits, 118
SCALPEL system, 274
Scanning capacitance microscope (SCM), 400–401
Scanning Electron Microscope (SEM), 170, 173, 651
Scanning systems, 235–37
Scanning Tunneling Microscope (STM), 400
SCM, *See* Scanning capacitance microscope (SCM)
Scotch tape method for adhesion, 737
Secondary defects, 481
Secondary grain growth, 761
Secondary Ion Mass Spectrometry (SIMS), 117, 175–76, 396–97, 733
Segregation, 125, 344, 413–15
  and float-zone process, 129
Selective Si deposition, 557
Selective W deposition, 567
Selectivity in etching, 609–11, 613
Self-aligned silicide, 568, 770
SEM, *See* Scanning Electron Microscope (SEM)
Semiconductor devices, 33–41
  bipolar junction transistors (BJTs), 39–41
  MOS transistors, 36–39
  PN diodes, 33–35
Semiconductor Industry Association (SIA), 2
Semiconductor manufacturing, 151–99
  basic concepts, 154–64
  clean factories, 157–59, 165, 169–73, 181–84
  gettering, 161–64, 167–69, 176–80
  historical development, 154–64
  measurement methods, 169–80
    clean factories, 169–73
    gettering, 176–80
    wafer cleaning, 173–76
  methods/equipment, 165–69
    clean factories, 165
    gettering, 167–69
    wafer cleaning, 166–67
  models/simulation, 180–93
    clean factories, 181–84
    gettering, 186–93
    wafer cleaning, 184–85
  technologies/models, limits/future trends in, 193–94
  wafer cleaning, 154, 159–61, 166–67, 173–76, 184–85
Semiconductors, 13–33
  bandgap in, 27
  band model, 18–19
  bond model, 18–19
  defined, 13
  degenerate semiconductors, 26
  direct bandgap semiconductors, 30
  dopants, 13
  electrical properties of, 18–19
Semiconductor surfaces, behavior of, 7
Semiconductor technology families, 41–43
  CMOS devices, 42–43
  NMOS devices, 41–42
  PMOS devices, 41–42
SEM (Scanning Electron Microscope), 120

Separation by implanted oxygen (SIMOX), 145
SHADE, 573
Shallow acceptor, 20
Shallow donors/acceptors, 19–20, 134
Shallow junctions, 373–74
Shallow trench isolation (STI), 57–60, 63
Sheet resistance, 113–15, 373, 398–99
Shockley, William, 7, 44, 451
Shockley-Read-Hall (SRH) recombination process, 30–31, 33, 37
Short channel effects, 78
Short–distance interconnects, *See* Local interconnects
Si_3N_4 growth and oxidation kinetics, 347–50
Sidewall spacers, 76, 78
Silane, 102, 558
Silicide
  formation-deposition, 567–68, 699–702, 716–18, 738–44
  oxidation, 350–52
  properties, 700
Silicon, 8–9, 131, 287
  carbon in, 142–44
  diamond crystal structure, 95–96
  future projections for, 3
  LOCOS process, 53
  oxygen in, 138–42
  as representative semiconductor, 14
  traps in, 30–32
Silicon crystal orientations, 96–97
Silicon dioxide
  deposition, 563–65
  growth, see Oxidation
  plasma etching, 644–47
Silicon interstitials, 142–43
Silicon lattice attom, missing, 97
Silicon nitride, 109–10
  deposition, 561–63
  films, 562–63
  growth and oxidation kinetics, 347–50
Silicon On Insulator (SOI) structures, 145–46
Silicon oxidation, 290
Silicon tetrachloride, 102
Silicon wafers:
  basic properties of, 93
  characterization of, 117
  cleaning, 154, 159–61, 166–67, 173–76, 184–85, 193–94
  dislocation free wafers, 98
  electrical measurements of, 111–12
  exposure systems, 234–41
  preparation/specification, crystals, 105–9
  tracking, 108
SIMBAD, 573, 602
SIMOX, 145
SIMS, *See* Secondary Ion Mass Spectrometry (SIMS)
Single-crystal wafers, 93
Si/SiO_2 interface changes, 352–57
Sirtl etch, 118
Slow diffusers, 386
Smart-Cut process, 146
Solid-phase epitaxy, 482–84
Solid solubility data, 377
Source/drain formation, 80–82
Spacer formation, 642
Spatial coherence of practical light sources for lithography, 217–18
Spatial frequencies of the diffraction pattern, 248
Specific contact resistivity, 689
Spectrometry:
  Rutherford Backscattering Spectrometry (RBS), 133, 175, 483, 732
  SIMS (Secondary Ion Mass Spectrometry), 175–76, 396–97
  sputter neutral mass spectrometry, 397
Spectrophotometers, 45
Spectroscopic ellipsometry, 572
Spectroscopy:
  AES (Auger Electron Spectroscopy), 174, 176, 651
  DLTS (Deep Level Transient Spectroscopy), 180
  Electron Spectroscopy for Chemical Analysis (ESCA), 732–33
  Fourier transform infrared spectroscopy (FTIR), 118–19, 732
  XES (X-ray Electron Spectroscopy), 174, 176
  XPS (X-ray Photoelectron Spectroscopy), 175
SPEEDIE, 573–74, 582, 590–98, 653, 710, 737
Spiking (contact), 692
Spin-On-Dielectric (SOD), 714–15
Spin-On-Glass (SOG), 512, 711–12, 714–15, 723
Spontaneous chemical etching. 627
Spreading resistance, 117, 397–98
Sputter deposition, 511, 539–554
  processes in, 542
Sputtered flux, topography models, 579–80
Sputter etching, 549–51, 634–36
Sputtering, 530
  targets, 543
  yield, 543, 580, 656
Sputter neutral mass spectrometry, 397
SRH (Shockley, Read, Hall) recombination, 30, 32–33, 190
Stacking faults, 100, 132, 192
Sticking coefficient, 538, 578, 624
  table of, 588
Stefan-Boltzmann constant, 123
Step coverage, 509–10, 538, 589
Steppers, 12, 54, 207, 235–37
STM, *See* Scanning Tunneling Microscope (STM)

Stress
during electromigration, 769–72
in thin films-measurement, 734–37
Stress dependent oxidation, 333–39
Stuffed barriers, 693
Sub-atmospheric CVD (SACVD), 564
Substrate doping effects, 343–45
SUPREM IV, 312, 333, 334, 336–37, 341, 347, 350, 357, 403, 405, 418, 420, 428, 435, 738
Surface analysis, 173
Surface curvature, 748
Surface diffusion fluxes, topography models, 577–78
Surface energy, 747
Surface inversion, 37
Surface recombination velocity, 32
Surface SIMS, 397
Swirl defects, 118, 133

T

TAURUS-TOPOGRAPHY, 653, 670–71, 737
TCA (trichloroethane), 297
Technology Computer-Aided Design (TCAD) simulators, 44, 91, 181
TED, see Transient Enhanced Diffusion (TED)
Temperature dependence of bandgap, 26–28
Temperature gradients, 99
TEM (Transmission Electron Microscope), 120–21, 293
Tetraethooxysilane (TEOS) oxide deposition, 563, 710, 718
Tetrakis–diethyamino titanium (TDEAT), 569
Tetrakis–dimethyamino titanium (TDMAT), 569
Thermal energy, and Si-Si bonds, 14
Thermal expansion, 734
Thermal stress, 734
Thermionic emission, 689
Thomas-Fermi model of the atom, 472
Threshold voltage, 42, 71
Ti deposition, 566–67
TiN barrier, 695, 720
TiN deposition, 568–69
TiN local interconnect, 83, 702, 717
Tip or extension (LDD) formation, 76–80
$TiSi_2$ deposition, 567–68, 716–718
Ti-W barrier, 694
Ti-W deposition, 566
Total stopping powers, 475
Total X-Ray Fluorescence (TXRF), 733
Transfer length, 727
Transient-Enhanced Diffusion (TED), 82, 375, 389, 393, 469, 484–97
atomic-level understanding of, 488–97
Transmission Electron Micrograph (TEM), Si/SiO_2 interface, 293
Traps, in silicon, 30–31
Trenching, 635
Trichlorosilane, 102
Tri-isobutyl-Al (TIBA), 566
TSUPREM IV, 312, 403, 416, 431, 462, 464, 743
Tungsten deposition, 567, 720
Tungsten plugs, 567, 704–5
Tungsten silicide, 567, 701
Tunneling, 289, 689

U

Ultralow energy implants, 468–69
Ultrashallow SIMS, 397
Ultrasonic cleaning, 160–61
Ultra violet (UV) light-sensitive part, of resist, 54
Unintentional doping:
of deposited layers, 522
methods used to produce atom flux in, 511

V

Vacancy, 97
diffusion, 417–26
oxidation, 339–45
Vacuum deposition method, PVD, 511
Valence electrons, 14
Van der Pauw method, 373
Vapor pressures of metals, 536
Vertical etching, 611, 666
Very Large Scale Integrated (VLSI) circuit process, 49, 89, 220
Via formation, 724
Vias, 681, 706
View Angle, 577
Viscoelastic flow, 336
VLSI circuit process flow, 49
Voids, 510
Volume defects in crystals, 101

W

Wafer cleaning, 154, 159–61, 166–67, 173–76, 184–85
Wafer exposure systems, 234–41
demagnification in, 204
Wafer preparation/specification, crystals, 105–9
Wafer tracking, 108
W deposition, 567, 720
Wet etching, 609, 612–19
Buffered HF (BHF)/Buffered Oxide Etch (BOE), 612, 618
common etchants (table), 619
etch bias, 613
etch rates, 618–19
etch thickness/depth, 613–14

overetching, 614–15
selectivity, 612–13
Wolff rearrangement, 224
WSi_2 deposition, 567–68

X
XES (X-ray Electron Spectroscopy), 174, 176
XPS (X-ray Photoelectron Spectroscopy), 175
X-ray Diffraction (XRD), 418, 733, 736
X-ray Fluorescence (XRF), 175–76
X-ray lithography, 275–77
X-ray Photoelectron Spectroscopy (XPS), 651, 732–33
XRF (X-ray Fluorescence), 175–76

Y
Yield Models, 181-84

Z
Zernike polynomial, 252
Zone refining:
defined, 129
and FZ growth, 128–31

PRENTICE HALL
Upper Saddle River, NJ 07458
http://www.prenhall.com

