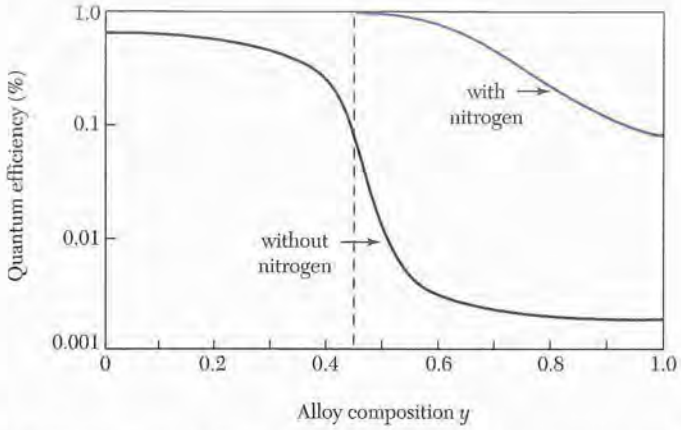


Fig. 8 Quantum efficiency versus alloy composition with and without isoelectronic impurity nitrogen.[4]

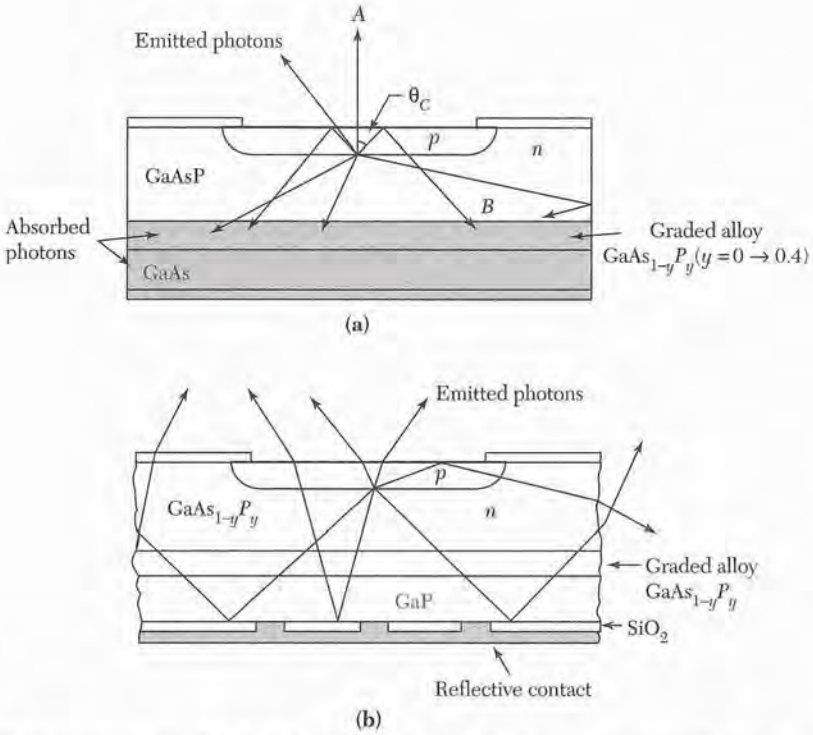


Fig. 9 Basic structure of a flat-diode LED and the effects of (*a*) an opaque substrate ($GaAs_{1-y}P_y$) and (*b*) a transparent substrate (GaP) on photons emitted at the *p-n* junction.[5]

$GaAs_{1-y}P_y$ layer is grown epitaxially to minimize the nonradiative centers at the interface that result from lattice mismatch.

For high-brightness blue LEDs (0.455–0.492 μm), II-VI compounds such as ZnSe, III-V nitride semiconductors such as GaN, and IV-IV compounds such as SiC have been investigated. However, their short lifetimes prevent II-IV–based devices from being

commercialized at present; and the indirect bandgap of SiC results in the low brightness of SiC blue LEDs.

The most promising candidates are GaN (E_g = 3.44 eV) and related III-V nitride semiconductors such as AlGaInN, which have direct bandgaps ranging from 1.95 eV to 6.2 eV with corresponding wavelengths from 0.2 μm to 0.63 μm.[6] Although there are no lattice-matched substrates for the growth of GaN, high-quality GaN has been grown on sapphire (Al_2O_3) using a low-temperature grown AlN as a buffer layer. Figure 10 shows a III-V nitride LED grown on sapphire substrate. Because the sapphire substrate is an insulator, both the n- and p-type ohmic contacts are formed on the top surface. The blue light originates from the radiative recombination in the $Ga_xIn_{1-x}N$ region which is sandwiched between two larger bandgap semiconductors: a p-type $Al_xGa_{1-x}N$ layer and an n-type GaN layer.

There are three loss mechanisms that reduce the quantity of emitted photons: absorption within the LED material, reflection loss when light passes from a semiconductor to air due to differences in refractive index, and total internal reflection of light at angles greater than the critical angle θ_c (Fig. 9a) defined by Snell's law:

$$\boxed{\sin\theta_c = \frac{\overline{n}_1}{\overline{n}_2},} \tag{10}$$

where the light passes from a medium with a refractive index of $\overline{n}_2$, such as GaAs with $\overline{n}_2$ = 3.66 at $\lambda \cong 0.8$ μm, to a medium of $\overline{n}_1$, such as air with $\overline{n}_1$ = 1. For gallium arsenide, the critical angle is about 16°; and for gallium phosphide with $\overline{n}_2$ = 3.45 at $\lambda \cong 0.8$ μm, the critical angle is about 17°.

The forward current-voltage characteristics of a LED is similar to that of the GaAs p–n junction discussed in Chapter 4. At low forward voltages, the diode current is dominated by the nonradiative recombination current due mainly to surface recombination near the perimeter of the LED chip. At higher forward voltages, the diode current is dominated by the radiative diffusion current. At even higher voltages, the series resistance will limit the diode current. The total diode current can be written as

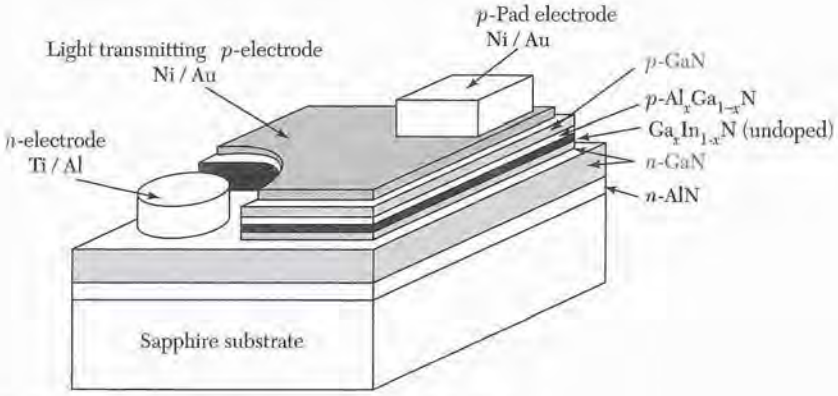


Fig. 10 III-V nitride LED grown on sapphire substrate.[6]

$$I = I_d \exp\left[\frac{q(V - IR_s)}{kT}\right] + I_r \exp\left[\frac{q(V - IR_s)}{2kT}\right], \tag{11}$$

where R_s is the device series resistance and I_d and I_r are the saturation currents due to diffusion and recombination, respectively. To increase the output power of the LED, we must reduce I_r and R_s.

The emission spectra of LEDs are similar to the eye response curve shown in Fig. 6. The spectral width is given by the full width at half maximum (FWHM) intensity. The spectral width generally varies as λ^2_m, where λ_m, is the wavelength at the maximum intensity.[7] Thus, the FWHM becomes larger as the wavelength is increased from visible to infrared. For example, at $\lambda_m = 0.55$ μm (green), FWHM is about 20 nm, but at 1.3 μm (infrared), FWHM is over 120 nm.

Visible LEDs can be used for full-color displays, full-color indicators, and lamps with high efficiency and high reliability. Figure 11 shows the diagrams of two LED lamps.[8] An LED lamp contains an LED chip and a plastic lens, which is usually colored to serve as an optical filter and to enhance contrast. The lamp in Fig. 11*a* uses a conventional diode header. Figure 11*b* shows a package that is suited for a transparent semiconductor, such as gallium phosphide, which emits light through all five facets (four sides and the top) of the LED chip.

Figure 12 shows the basic formats for LED displays. The seven segments (Fig. 12*a*) display numbers from 0 to 9. The 5 × 7 matrix array (Fig. 12*b*) displays alphanumerics (A–Z and 0–9). The displays can be made by monolithic processes similar to those used

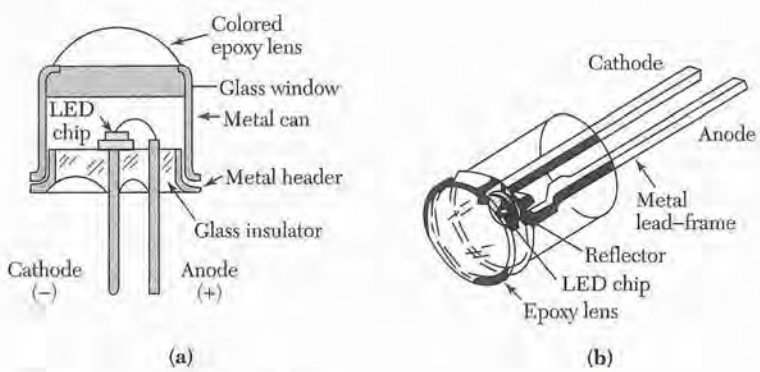


Fig. 11 Diagrams of two LED lamps.[8]

Fig. 12 LED display formats for numeric and alphanumeric: (*a*) 7-segment (numeric); (*b*) 5 × 7 array (alphanumeric).[8]

to make silicon integrated circuits (described in Chapters 10 through 14) or by using an individual LED chip mounted on a reflector to form a bar segment.

There has been interest in the development of white LEDs for general illumination because LEDs are three times as efficient as incandescent lamps and can last 10 times longer. A white LED requires LEDs of red, green, and blue colors. A wide-spread use of white LED will be realized when the costs of these LEDs, especially the blue LEDs, become competitive with the conventional light sources.

Organic LED

In the previous discussions, we considered devices made only from *inorganic* semiconductor materials, such as GaAsP and GaN. In recent years, certain *organic semiconductors* have been studied for electroluminescent applications. The organic light-emitting diode (OLED) is particularly useful for multicolor, large-area flat-panel display because of its attributes of low-power consumption and excellent emissive quality with a wide viewing angle.[9]

Figure 13*a* shows the molecular structures of two representative organic semiconductors.[10] They are the tris (8-hydroxy-quinolinato) aluminum (AlQ_3), which contains six benzene rings connected to a central aluminum atom, and the aromatic diamine, which also contains six benzene rings but with a different molecular arrangement. A basic OLED has a number of layers on a transparent substrate (e.g., glass). Onto the substrate we deposite, in sequence, a transparent conductive anode [e.g., ITO (indium tin oxide)], the diamine as a hole transport layer, the AlQ_3 as the electron transport layer, and the cathode contact (e.g., Mg alloy with 10% Ag). A cross sectional view is shown in Fig. 13*b*.

Figure 13*c* shows the band diagram of the OLED. It is basically a heterojunction formed between AlQ_3 and diamine. Under proper biasing conditions, electrons are injected from the cathode and move toward the heterojunction interface whereas holes are injected from the anode and also move toward the interface. Because of the energy barriers ΔE_C and ΔE_V, these carriers will accumulate at the interface to enhance the chance of radiative recombination.

From Fig. 13*c*, we can specify the design criteria of an OLED: (a) ultrathin layers for low biasing voltage—for example, the total thickness of the organic semiconductor

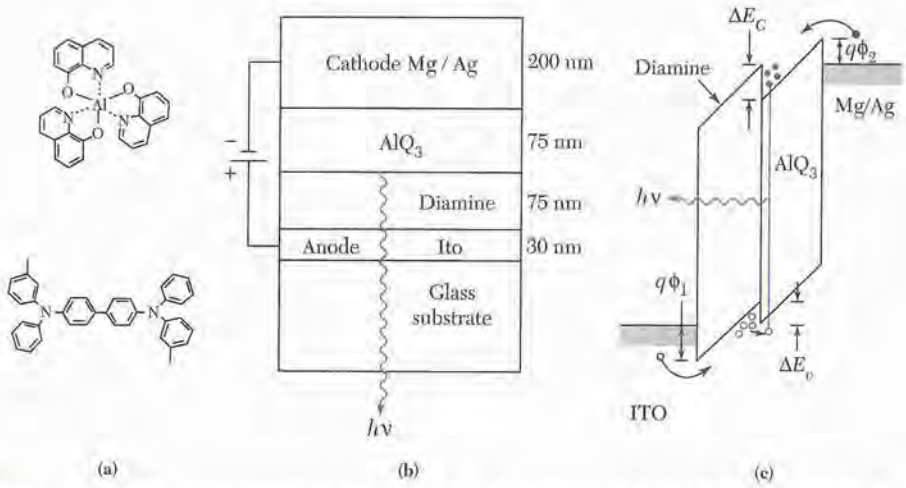


Fig. 13 (*a*) Organic semiconductors. (*b*) OLED cross sectional view. (*c*) Band diagram of an OLED.

layers shown is only 150 nm; (b) low injection barriers—the barrier height $q\phi_1$ for hole injection and the barrier height $q\phi_2$ for electron injection must be low enough to allow large carrier injections for high-current density operation, and (c) proper bandgaps for the required color. For AlQ_3, the emitted light is green color. By choosing different organic semiconductors with different bandgaps, various colors including red, yellow, and blue can be obtained.

9.2.2 Infrared LED

Infrared LEDs include gallium arsenide LEDs, which emit light near 0.9 μm, and many III-V compounds, such as the quaternary $Ga_xIn_{1-x}As_yP_{1-y}$ LEDs, which emit light from 1.1 to 1.6 μm.

An important application of infrared LEDs is in opto-isolators, where an input or control signal is decoupled from the output. Figure 14 shows an opto-isolator having an infrared LED as the light source and a photodiode as the detector. When an input signal is applied to the LED, light is generated and subsequently detected by the photodiode. The light is then converted back to an electrical signal as a current that flows through a load resistor. Opto-isolators transmit signals at the speed of light and are electrically isolated because there is no electrical feedback from the output to the input.

Another important application of infrared LEDs is for transmission of an optical signal through an optical fiber, as in a communication system. An optical fiber is a waveguide at optical frequencies. The fiber is usually drawn from a preform of glass to a diameter of about 100 μm. It is flexible and can guide optical signals over distances of many kilometers to a receiver, similar to the way a coaxial cable transmits electrical signals.

Two types of optical fibers are shown in Fig. 15. One type of fiber has a cladding layer of relatively pure fused silica (SiO_2) surrounding a core of doped glass (e.g., germanium doped glass) that has a higher refractive index than the cladding layer.[11] This type of fiber is called a *step index fiber*. The light is transmitted along the length of the fiber by internal reflection at the step in the refractive index. The critical angle for internal reflection is about 79° for $\overline{n}_1 = 1.457$ (cladding layer) and $\overline{n}_2 = 1.480$ (core, 20% Ge-doped) as calculated from Eq. 10. Note that different rays will propagate with different path lengths (Fig. 15*a*). A light pulse reaching the end of a step index fiber will result in a pulse spread. In a *graded-index fiber* (Fig. 15*b*), the index decreases from the core center by a parabolic law. Now, rays traversing toward the cladding have a high velocity (due to lower refractive index) than rays along the center of the core. The pulse spread is significantly reduced. As the light is transmitted along the optical fiber, the light signal will be attenuated. However, due to the transparency of ultrapure silica used for the fiber material in the wavelength region from 0.8 to 1.6 μm, the attenuation is quite low and

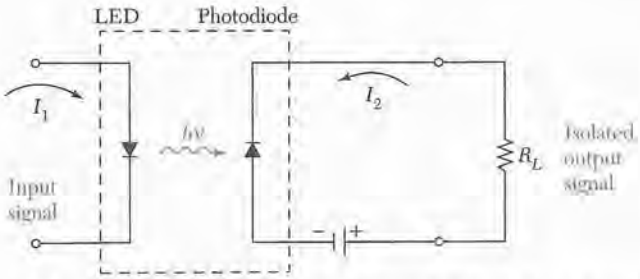


Fig. 14 An opto-isolator in which an input signal is decoupled from the output signal.

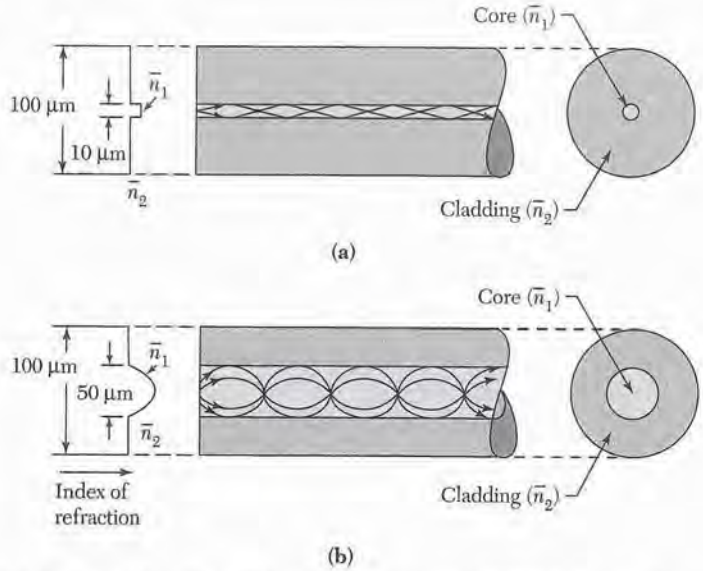


Fig. 15 Optical fibers. (*a*) Step-index fiber having a core with slightly larger reflective index. (*b*) Graded-index fiber having a parabolic grading of the reflective index in the core.[11]

is proportional to λ^{-4}. Typical attenuations are about 3 dB/km at a wavelength of 0.8 μm, 0.6 dB/km at 1.3 μm, and 0.2 dB/km at 1.55 μm.

A simple point-to-point optical-fiber communication system is shown in Fig. 16, where the electrical input signals are converted to optical signals using an optical source (LED or laser). The optical signals are coupled into the fiber and transmitted to the photodetector, where they are converted back to electrical signals.

The surface-emitting infrared InGaAsP LED used for optical-fiber communication is shown[12] in Fig. 17. The light is emitted from the central surface area and coupled into the optical fiber. The use of heterojunctions (e.g., InGaAsP-InP) can increase the efficiency that results from the confinement of the carrier by the layers of the higher-bandgap semiconductor InP, surrounding the radiative-recombination region InGaAsP. We con-

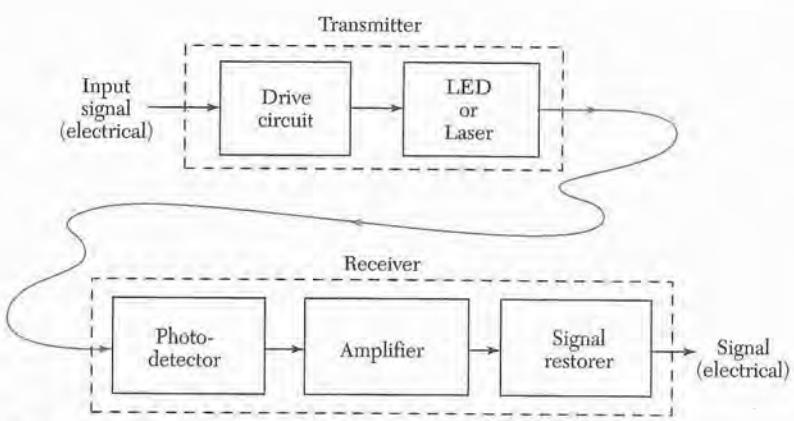


Fig. 16 Basic elements of an optical fiber transmission link.

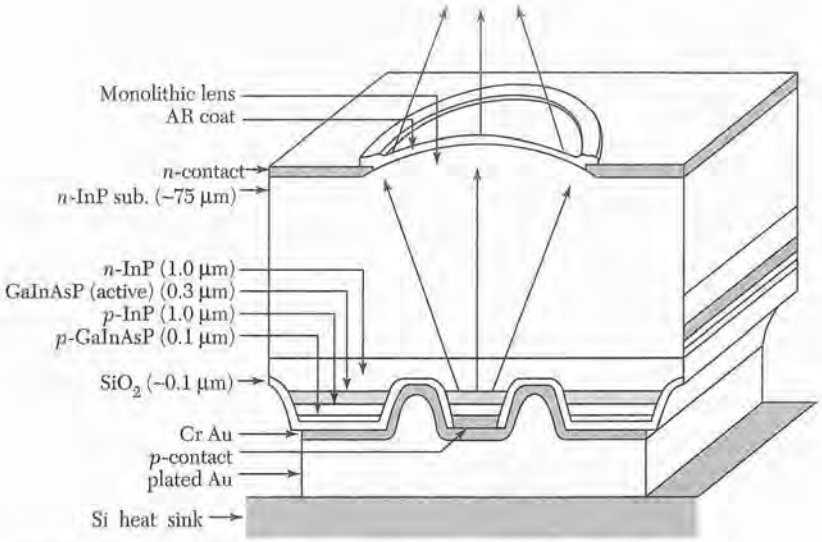


Fig. 17 Small-area mesa-etched GaInAsP/InP surface-emitting LED structure.[12]

sider the carrier confinement in Section 9.3. The heterojunction can also serve as an optical window to the emitted radiation because the higher-bandgap–confining layers do not absorb radiation from the lower-bandgap–emitting region.

The electrical input signal is generally modulated at high frequencies. This signal causes direct modulation of the injected current in an LED. Parasitic elements such as the depletion-layer capacitance and series resistance can cause a delay of carrier injection into the junction and a delay in the light output. The ultimate limit on how fast one can vary the light output depends on the carrier lifetime, which is determined by various recombination processes, such as the surface recombination discussed in Chapter 3. If the current is modulated at an angular frequency ω, the light output $P(\omega)$ is given by

$$P(\omega) = \frac{P(0)}{\sqrt{1+(\omega\tau)^2}}, \tag{12}$$

where $P(0)$ is the light output at $\omega = 0$, and τ is the carrier lifetime. The *modulation bandwidth* Δf is defined as the frequency at which the light output is reduced to $1/\sqrt{2}$ that at $\omega = 0$, that is,

$$\Delta f \equiv \frac{\Delta\omega}{2\pi} = \frac{1}{2\pi\tau}. \tag{13}$$

EXAMPLE 2

Calculate the modulation bandwidth of a GaAs-based LED with $\tau = 500$ ps.

SOLUTION From Eq. 13,

$$\Delta f = \frac{1}{2\pi \cdot 500 \cdot 10^{-12}} = 318 \text{ MHz}.$$

9.3 SEMICONDUCTOR LASER

Semiconductor lasers are similar to the solid-state ruby laser and helium-neon gas laser in that the emitted radiation is highly monochromatic and produces a highly directional beam of light. However, the semiconductor laser differs from other lasers in that it is small (on the order of 0.1 mm long) and is easily modulated at high frequencies simply by modulating the biasing current. Because of these unique properties, the semiconductor laser is one of the most important light sources for optical-fiber communication. It is also used in video recording, optical reading, and high-speed laser printing. In addition, semiconductor lasers have significant applications in many areas of basic research and technology, such as high-resolution gas spectroscopy and atmospheric pollution monitoring.

9.3.1 Semiconductor Materials

All lasing semiconductors have direct bandgaps. This is expected because the momentum is conserved and the radiative-transition probability in a direct-bandgap semiconductor is high. At present, the laser emission wavelengths cover the range from 0.3 to over 30 μm. Gallium arsenide was the first material to emit laser radiation, and its related III-V compound alloys are the most extensively studied and developed.

The three most important III-V compound alloy systems are $Ga_xIn_{1-x}As_yP_{1-y}$, $Ga_xIn_{1-x}As_ySb_{1-y}$, and $Al_xGa_{1-x}As_ySb_{1-y}$ solid solutions. Figure 18 shows the bandgaps plotted against the lattice constant for various III-V binary semiconductors and their intermediate ternary and quaternary compounds.[13] To achieve a heterostructure with negligible interface traps, the lattices between the two semiconductors must be matched closely.

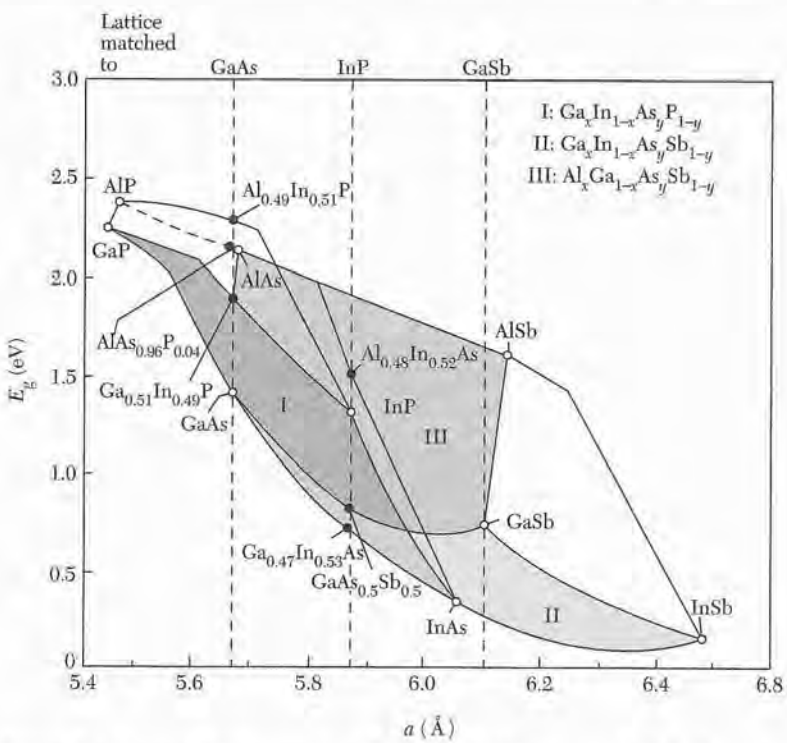


Fig. 18 Energy bandgap and lattice constant for three III-V compound solid alloy system.[13]

If we use GaAs (a = 5.6533 Å) as the substrate, the ternary compound $Al_xGa_{1-x}As$ can have a lattice mismatch less than 0.1%. Similarly, with InP (a = 5.8687 Å) as the substrate, the quaternary compound $Ga_xIn_{1-x}As_yP_{1-y}$ also can have a nearly perfect lattice match, as indicated by the center vertical line in Fig. 18.

Figure 19*a* shows the bandgap of ternary $Al_xGa_{1-x}As$ as a function of aluminum composition.[1] The alloy has a direct bandgap up to x = 0.45, then becomes an indirect-bandgap semiconductor. Figure 19*b* shows the compositional dependence of the refractive index. For example, for x = 0.3, the bandgap of $Al_xGa_{1-x}As$ is 1.789 eV, which is 0.365 eV larger than that of GaAs; its refractive index is 3.385, which is 6% smaller than that of GaAs. These properties are important for continuous operation of semiconductor lasers at and above room temperatures.

9.3.2 Laser Operation

Figure 20 shows schematic representations[14] of the band diagram under forward-bias , the refractive index profile, and the optical-field distribution of light generated at the junction of a homojunction laser (Fig. 20*a*) and a double-heterostructure (DH) laser (Fig. 20*b*).

Population Inversion

As discussed in Section 9.1.1, to enhance the stimulated emission for laser operation we need population inversion. To achieve population inversion in a semiconductor laser, we consider a p-n junction or a heterojunction formed between degenerate semiconductors. This means that the doping levels on both sides of the junction are high enough that the Fermi level E_{FV} is below the valence band edge on the p-side and E_{FC} is above the conduction band edge on the n-side. When a sufficiently large bias is applied (the band diagrams in Fig. 20), high injection occurs, that is, large concentrations of electrons and holes are injected into the transition region. As a result, the region d (Fig. 20) contains a large concentration of electrons in the conduction band and a large concentration of holes in the valence band; this is the required condition for population inversion. For band-to-band

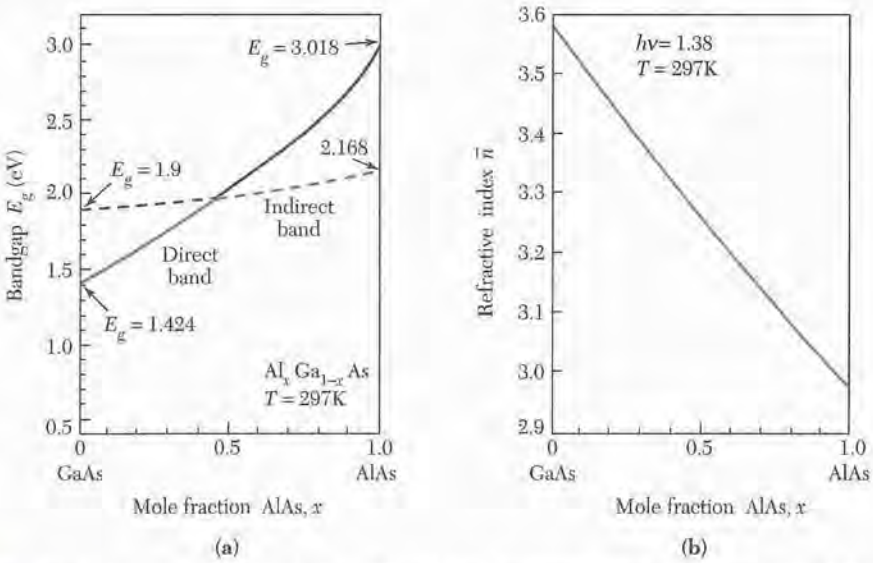


Fig. 19 (*a*) Compositional dependence of the $Al_xGa_{1-x}As$ energy gap.[1] (*b*) Compositional dependence of the refractive index at 1.38 eV.

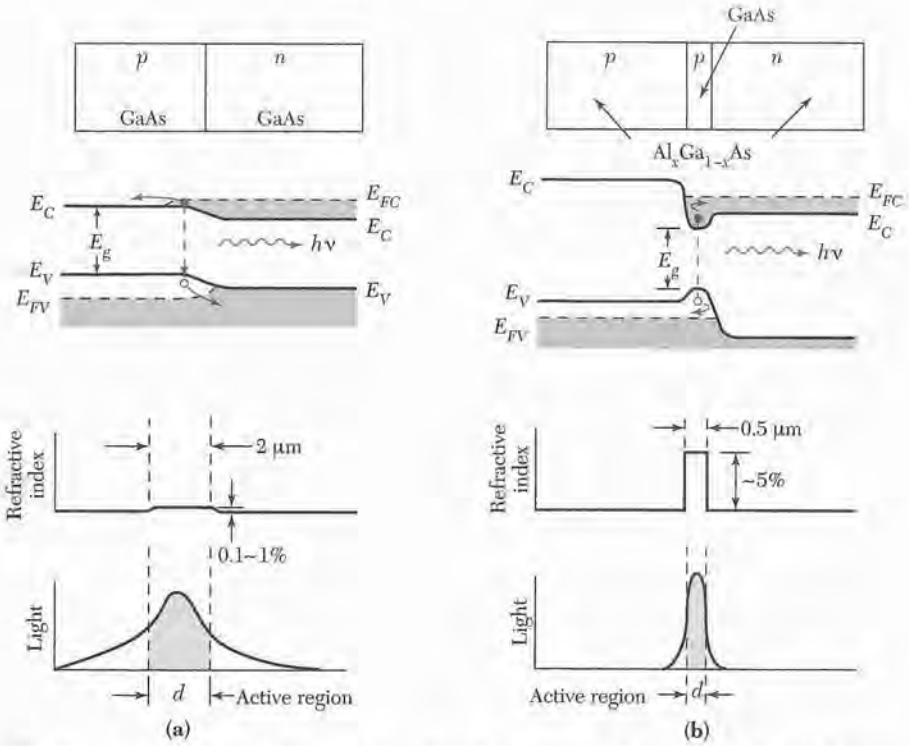


Fig. 20 Comparison of some characteristics of (*a*) homojunction laser and (*b*) double-heterojunction (DH) laser. Second from the top row shows energy band diagrams under forward bias. The refractive index change for a homojunction laser is less than 1%. The refractive index change for DH laser is about 5%. The confinement of light is shown in the bottom row.[14]

transition, the minimum energy required is the bandgap energy E_g. Therefore, from the band diagram in Fig. 20, we can write the condition necessary for population inversion: $(E_{FC} - E_{FV}) > E_g$.

Carrier and Optical Confinement

As can be seen in the DH laser, the carriers are confined on both sides of the active region by the heterojunction barriers, whereas in the homojunction laser the carriers can move away from the active region, where radiative recombination occurs.

In the DH laser the optical field is also confined within the active region by the abrupt reduction of the refractive index outside the active region. The optical confinement can be explained by Fig. 21, which shows a three-layer dielectric wave guide with refractive indices $\overline{n}_1$, $\overline{n}_2$ and $\overline{n}_3$, where an active layer is sandwiched between two confining layers (Fig. 21*a*). Under the condition $\overline{n}_2 > \overline{n}_1 \geq \overline{n}_3$, the ray angle θ_{12} at the layer 1/layer 2 interface in Fig. 21*b* exceeds the critical angle given by Eq. 10. A similar situation occurs for θ_{23} at the layer 2/layer 3 interface. Therefore, when the refractive index in the active layer is larger than the index of its surrounding layers, the propagation of the optical radiation is guided (confined) in a direction parallel to the layer interfaces. We can define a *confinement factor* Γ, which is the ratio of the light intensity within the active layer to the

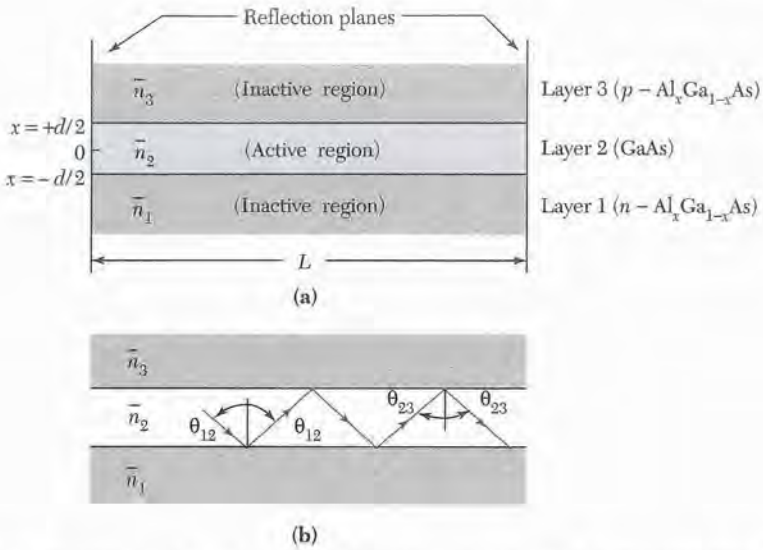


Fig. 21 (*a*) Representation of a three-layer dielectric waveguide. (*b*) Ray trajectories of the guided wave.

sum of light intensity both within and outside the active layer. The confinement factor is given as

$$\boxed{\Gamma \cong 1 - \exp\left(-C\Delta\bar{n}d\right),} \tag{14}$$

where C is a constant, $\Delta\bar{n}$ is the difference in the refractive index, and d is the thickness of the active layer. It is clear that the larger $\Delta\bar{n}$ and d are, the higher Γ will be.

Optical Cavity and Feedback

We have considered the condition necessary to produce laser action: population inversion. Photons released by stimulated emission are likely to cause further stimulations as long as there is population inversion. This is the phenomenon of optical gain. The gain obtained in a single travel of an optical wave down a laser cavity is small. To increase gain, multiple passes of a wave must occur. This is achieved using mirrors placed at either end of the cavity, shown as the reflection planes at the left side and right side in Fig. 21*a*. For a semiconductor laser, the cleaved ends of the crystal forming the device can act as the mirrors. For a GaAs device, cleaving along (110) plane creates two parallel identical mirrors. Sometimes the back mirror of the laser is metallized to enhance the reflectivity. The reflectivity R at each mirror can be calculated as

$$\boxed{R = \left(\frac{\bar{n} - 1}{\bar{n} + 1}\right)^2,} \tag{15}$$

where $\bar{n}$ is the refractive index in the semiconductor corresponding to the wavelength λ ($\bar{n}$ is generally a function of λ).

EXAMPLE 3

Calculate the R for GaAs ($\bar{n} = 3.6$).

SOLUTION From Eq. 15,

$$R = \left(\frac{3.6-1}{3.6+1}\right)^2 = 0.32,$$

that is, 32% of the light will be reflected at the cleaved surface.

If an integral number of half-wavelength fit between the two end planes, reinforced and coherent light will be reflected back and forth within the cavity. Therefore, for stimulated emission, the length L of the cavity must satisfy the condition

$$m\left(\frac{\lambda}{2\bar{n}}\right) = L \tag{16}$$

or

$$m\lambda = 2\bar{n}L, \tag{16a}$$

where m is an integral number. Obviously, many values of λ can satisfy this condition (Fig. 22*a*), but only those within the spontaneous emission spectrum will be produced (Fig. 22*b*). In addition, optical losses in the path traveled by the wave mean that only the strongest lines will survive, leading to a set of lasing modes, as given in Fig. 22*c*. The separation $\Delta\lambda$ between the allowed modes in the longitudinal direction is the difference in the wavelengths corresponding to m and $m+1$. Differentiating Eq. 16*a* with respect to λ, we obtain

$$\Delta\lambda = \frac{\lambda^2 \Delta m}{2\bar{n}L[1-(\lambda/\bar{n})(d\bar{n}/d\lambda)]}. \tag{17}$$

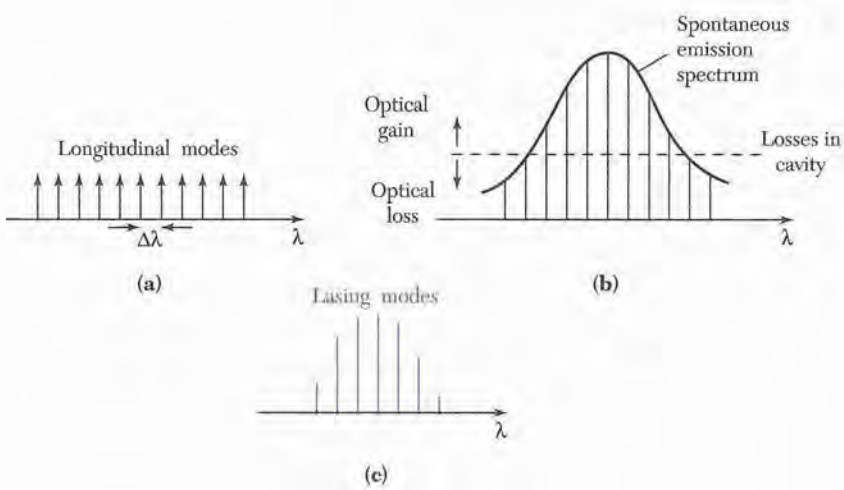


Fig. 22 (*a*) Resonant modes of a laser cavity. (*b*) Spontaneous emission spectrum. (*c*) Optical-gain wavelengths.

Although $\overline{n}$ is a function of λ, over the very small change in wavelength between adjacent modes $d\overline{n}/d\lambda$ is very small, and hence to a good approximation the mode spacing $\Delta\lambda$ is given by

$$\boxed{|\Delta\lambda| \cong \frac{\lambda^2}{2\overline{n}L}.} \tag{18}$$

EXAMPLE 4

Calculate the mode spacing for a typical GaAs laser with $\lambda = 0.94$ μm, $\overline{n} = 3.6$, and $L = 300$ μm.

SOLUTION From Eq. 18,

$$\Delta\lambda \cong \frac{\left(0.94\times10^{-6}\right)^2}{2\times3.6\times300\times10^{-6}} = 4\times10^{-10}\,\text{m} = 4\ \text{Å}.$$

9.3.3 Basic Laser Structure

Figure 23 shows three laser structures.[14,15] The first structure (Fig. 23*a*), is a basic *p-n* junction laser and is called a *homojunction laser* because it has the same semiconductor material (e.g., GaAs) on both sides of the junction. A pair of parallel planes (or facets) are cleaved or polished perpendicular to the ⟨110⟩-axis. Under appropriate biasing conditions laser light will be emitted from these planes (only the front emission is shown in Fig. 23). The two remaining sides of the diode are roughened to eliminate lasing in the directions other than the main ones. This structure is called a *Fabry–Perot cavity*, with a typical cavity length L of about 300 μm. The Fabry–Perot cavity configuration is used extensively for modern semiconductor lasers.

Figure 23*b* shows a *double-heterostructure* (DH) *laser*, in which a thin layer of a semiconductor (e.g., GaAs) is sandwiched between layers of a different semiconductor (e.g., $Al_xGa_{1-x}As$). The laser structures shown in Figs. 23*a* and *b* are broad-area lasers because the entire area along the junction plane can emit radiation. Figure 23*c* shows a DH laser with a stripe geometry. The oxide layer isolates all but the stripe contact; consequently the lasing area is restricted to a narrow region under the contact. The stripe widths S are typically 5–30 μm. The advantages of the stripe geometry are reduced operating current, elimination of multiple-emission areas along the junction, and improved reliability that is the result of removing most of the junction perimeter.

Threshold Current Density

One of the most important parameters for laser operation is the threshold current density J_{th}, that is, the minimum current density required for lasing to occur. Figure 24 compares J_{th} versus operating temperature for a homojunction laser and a DH laser.[14] Note that as the temperature increases, J_{th} for the DH laser increases much more slowly than J_{th} for the homojunction laser. Because of the low values of J_{th} for DH lasers at 300 K, DH lasers can be operated continuously at room temperature. This characteristic has led to the increased use of semiconductor lasers, especially in optical-fiber communication systems.

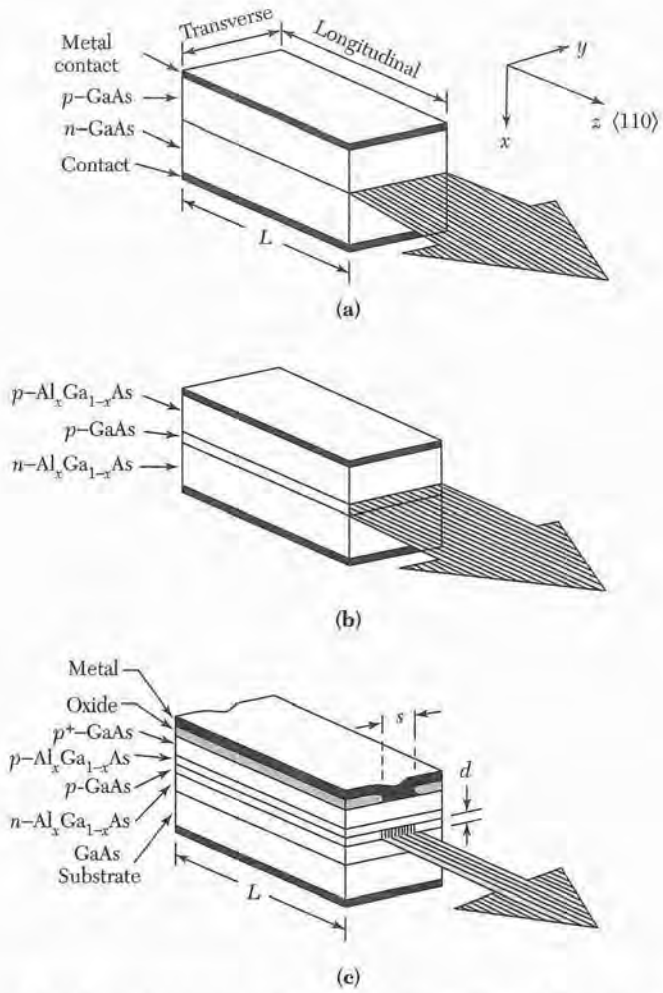


Fig. 23 Semiconductor laser structure in the Fabry–Perot-cavity configuration. (*a*) Homojunction laser. (*b*) Double-heterojunction (DH) laser. (*c*) Stripe-geometry DH laser.[14,15]

In a semiconductor laser, the gain g, the incremental optical energy flux per unit length, depends on the current density. The gain g can be expressed as a function of a nominal current density J_{nom}, which is defined for unity quantum efficiency (i.e., number of carriers generated per photon, $\eta = 1$) as the current density required to uniformly excite 1 μm thick an active layer. The actual current density is then given by

$$J(\mathrm{A/cm^2}) = \frac{J_{nom}\, d}{\eta}, \tag{19}$$

where d is the thickness of the active layer in μm. Figure 25 shows the calculated gain for a typical gallium arsenide DH laser.[16] The gain increases linearly with J_{nom} for $50 \le g \le 400\ \mathrm{cm^{-1}}$. The linear dashed line can be written as

$$g = \left(g_0/J_0\right)\left(J_{nom} - J_0\right), \tag{20}$$

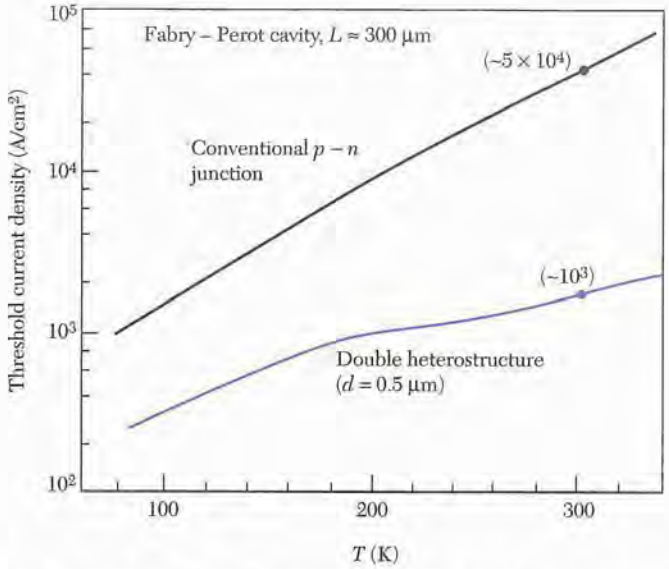


Fig. 24 Threshold current density versus temperature for the two laser structures shown[14] in Fig. 20.

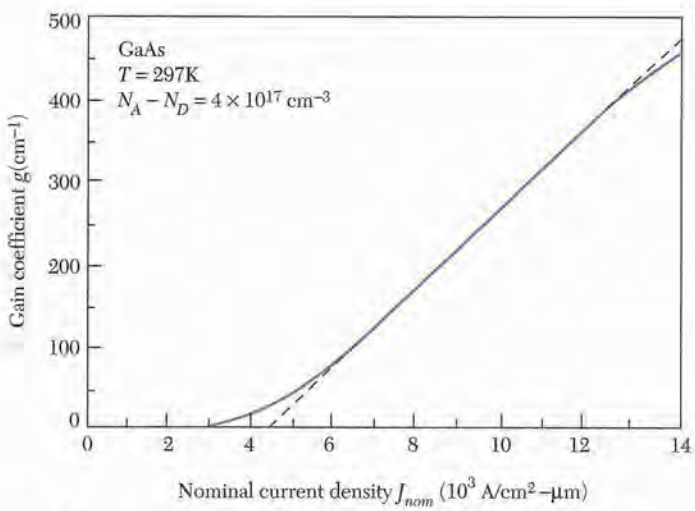


Fig. 25 Variation of gain coefficient versus nominal current density. Dashed line represents a linear dependence.[16]

where $g_0/J_0 = 5 \times 10^{-2}$ cm-μm/A and $J_0 = 4.5 \times 10^3$ A/ cm-μm.

As discussed previously, at low currents there is spontaneous emission in all directions. As the current increases, the gain increases (Fig. 25) until the threshold for lasing is reached, that is, until the gain satisfies the condition that a light wave makes a complete traversal of the cavity without attenuation:

$$R \exp\left[\left(\Gamma g - \alpha\right)L\right] = 1 \tag{21}$$

or

$$\Gamma g\left(\text{threshold gain}\right) = \alpha + \frac{1}{L}\ln\left(\frac{1}{R}\right), \tag{22}$$

where Γ is the confinement factor, α is the loss per unit length from absorption and other scattering mechanism, L is the length of the cavity shown in Fig. 23, and R is the reflectance of the ends of the cavity assuming that R for both ends is equal. Equations 19, 20, and 22 may be combined to give the threshold current density as

$$\boxed{J_{th}(\text{A/cm}^2) = \frac{J_0 d}{\eta} + \left(\frac{J_0 d}{g_0 \eta \Gamma}\right)\left[\alpha + \frac{1}{L}\ln\left(\frac{1}{R}\right)\right].} \tag{23}$$

The term $(J_0 d/g_0 \eta\Gamma)$ is often called $1/\beta$, where β is known as the gain factor. To reduce J_{th}, we can increase η, Γ, L, and R and reduce d and α.

EXAMPLE 5

Find the threshold current for a laser diode using the following data: front and rear mirror reflectivities are 0.44 and 0.99, respectively. The cavity length and width are 300 μm and 5 μm, respectively, $\alpha = 100\ \text{cm}^{-1}$, $\beta = 0.1\ \text{cm}^{-3}\text{A}^{-1}$, $g_0 = 100\ \text{cm}^{-1}$, and $\Gamma = 0.9$.

SOLUTION With a known gain factor, the term $J_0 d/\eta$ in Eq. 23 can be expressed as $g_0\Gamma/\beta$.

Due to different reflectivities of the two mirrors, Eq. 23 is modified to

$$J_{th}(\text{A/cm}^2) = \frac{g_0\Gamma}{\beta} + \frac{1}{\beta}\left[\alpha + \frac{1}{2L}\ln\left(\frac{1}{R_1 R_2}\right)\right] \tag{23a}$$

Thus, $$J_{th} = \frac{100\times 0.9}{0.1} + 10\times\left[100 + \frac{1}{2\times 300\times 10^{-4}}\ln\left(\frac{1}{0.44\times 0.99}\right)\right] = 2036\ \text{A/cm}^2,$$

and so $I_{th} = 2036\times 300\times 10^{-4}\times 5\times 10^{-4} = 30$ mA.

Temperature Effect

Figure 26 shows the temperature dependence of the threshold current I_{th} for a cw (continuous wave) stripe geometry $Al_xGa_{1-x}As$-GaAs DH laser.[17] Figure 26*a* shows cw light outputs versus injection current at various temperatures between 25° and 115°C. Note the excellent linearity in the light-current characteristics. The threshold current at a given temperature is the extrapolated value for zero output power. Figure 26*b* shows a plot of threshold currents as a function of temperature. The threshold current increases exponentially with temperature as

$$I_{th} \sim \exp\left(\frac{T}{T_0}\right), \tag{24}$$

where T is the temperature in °C and T_0 is 110°C for this laser.

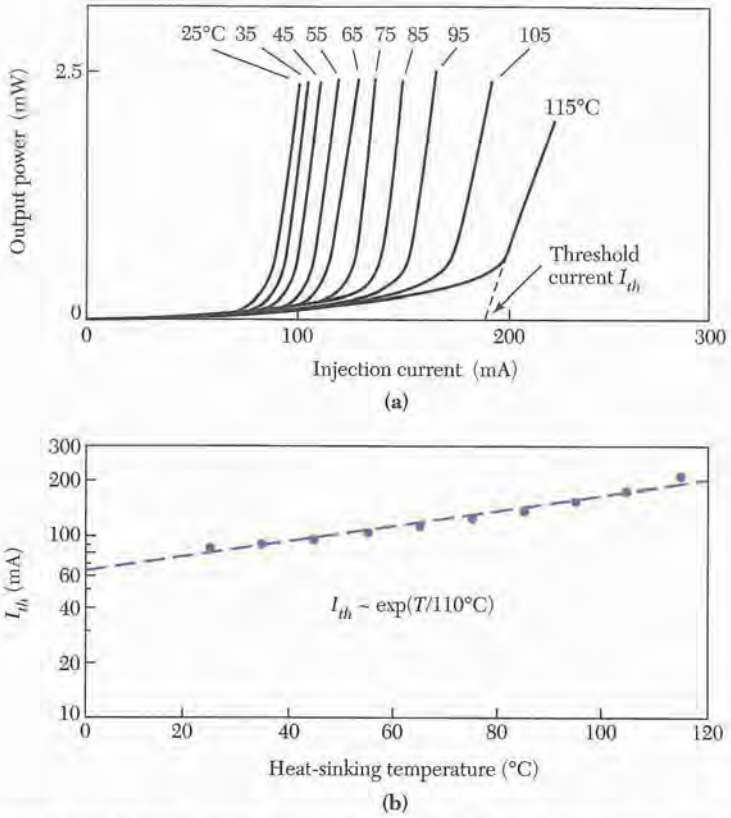


Fig. 26 (a) Light output versus diode current for a GaAs/AlGaAs heterostructure laser. (b) Temperature dependence of the continuous wave (cw)-current threshold. [17]

EXAMPLE 6

Calculate the temperature at which the room-temperature value of the threshold current doubles for the laser shown in Fig. 26.

SOLUTION

$$\frac{J_{th}}{2J_{th}} = \frac{\exp(27/110)}{\exp(T/110)},$$

therefore $T = 27 + 110 \times \ln 2 = 27 + 76 = 103°C$.

Modulation Frequency and Longitudinal Modes

For optical fiber communications, the optical source must be able to be modulated at high frequencies. Unlike LEDs, whose output power decreases with increasing modulation bandwidth (Eq. 12), the output power of typical GaAs or GaInAsP laser remains at a constant level (e.g., 10 mW per facet) well into GHz range.

For a stripe-geometry GaInAs-AlGaAs DH laser at a current above the threshold, many emission lines exist that are approximately evenly spaced with a separation of $\Delta\lambda$

(e.g., $\Delta\lambda = 4$ Å in Ex. 4). These emission lines belong to the longitudinal modes given in Eq. 17. Because of these longitudinal modes, the stripe geometry laser is not a spectrally pure light source. For optical-fiber communication systems, an ideal light source is one that has a single frequency. This is because light pulses of different frequencies travel through optical fiber at different speeds, thus causing pulse spread.

9.3.4 Distributed Feedback Lasers

Because of the multimodes in stripe-geometry lasers, these devices are only useful for telecommunication systems operated at relatively low rates (i.e., below 1 Gbit/s). For advanced optical fiber systems, *single-frequency* lasers are necessary. The fundamental approach is to take a laser cavity that allows only one mode to resonate and to provide a constructive interference mechanism that picks out a single frequency. Two laser configurations use this approach—the distributed Bragg reflector laser and the distributed feedback laser.[18]

Figure 27*a* shows the cross section of a distributed Bragg reflector (DBR) laser. The region which conducts electric current is called the pumped region. A wavelength-selection grating is placed outside the pumped region. Because of efficient coupling between the active region and the passive grating structure, the reflection is enhanced at the wavelength λ_B, known as the Bragg wavelength, which is related to the period of the grating Λ by

$$\lambda_B = \frac{2\bar{n}\Lambda}{l}, \tag{25}$$

where $\bar{n}$ is the effective refractive index of the mode and l is the integer order of the grating. The mode at the Bragg wavelength that has the lowest loss, and thus, the lowest threshold gain will have the predominant output.

Figure 27*b* shows the distributed feedback (DFB) laser, which has a corrugated grating structure within the active region. The grating region has a periodically varying index of refraction that enhances the wavelength closest to the Bragg wavelength, thus achieving single-frequency operation. Because of the small temperature dependence of the refractive index, the lasing wavelength of the DFB laser has a very small temperature coefficient

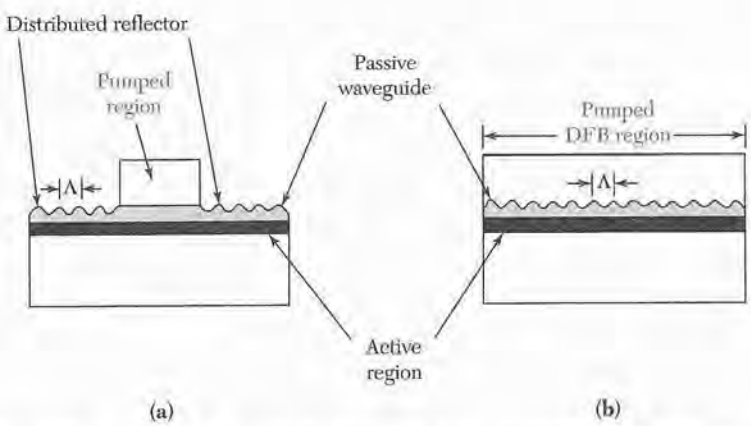


Fig. 27 Two methods of obtaining a single-frequency laser. (*a*) Distributed Bragg reflector (DBR) laser, and (*b*) a distributed feedback (DFB) laser.

(~0.5 Å/°C), while the temperature coefficient for a corresponding stripe-geometry laser is much larger (~3 Å/°C), because it follows the temperature dependence of the bandgap. DBR and DFB lasers are also useful as optical sources in integrated optics, which uses miniature optical waveguide components and circuits made by planar technology on rigid substrates.

9.3.5 Quantum-Well Lasers

The structure of a quantum-well (QW) laser[18,19] is similar to that of a DH laser except the thickness of the active layer in a QW laser is very small, about 10–20 nm. Figure 28*a* shows the band diagram of a QW laser where the central GaAs region ($L_y \cong 20$ nm) is sandwiched between two larger bandgap AlGaAs layers. The length L_y is comparable to the de Broglie wavelength ($\lambda = h/p$, where h is the Planck constant and p is the momentum of the charge carrier), and the carriers are confined in a finite potential well in the y-direction.

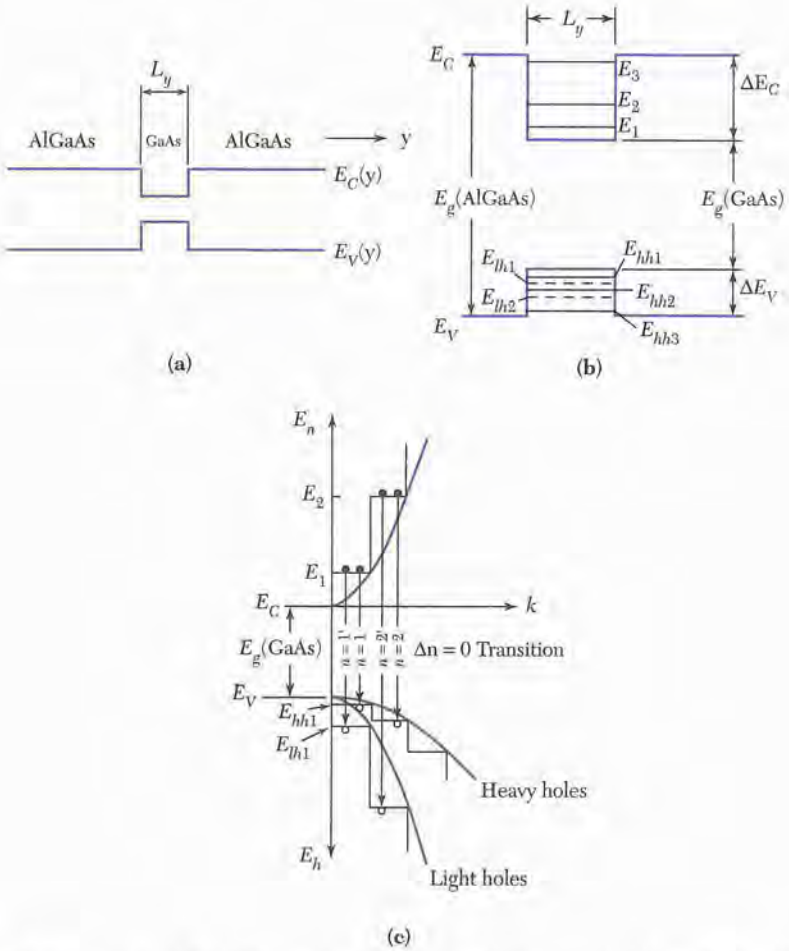


Fig. 28 The quantum-well (QW) laser: (*a*) single GaAs QW surrounded by AlGaAs, (*b*) discrete energy levels within the well, and (*c*) density of states for electrons and holes within the well.

The energy of the charge particle can be separated into a confinement component in the y-direction and two unconfined components in the x- and z-direction of the QW layer.

$$E(n, k_x, k_z) = E_n + \frac{\hbar^2}{2m^*}(k_x^2 + k_z^2), \tag{26}$$

where E_n is the nth eigenvalue of the confined particle, m^* is the effective mass, and k_x and k_z are the wave number in the x- and z-direction, respectively. Figure 28*b* shows the energy levels in the quantum well. The values of E_n are shown as E_1, E_2, E_3 for electrons, E_{hh1}, E_{hh2}, E_{hh3}, for heavy holes,[§] and E_{lh1}, E_{lh2} for light holes.[18] The usual parabolic forms for the conduction and valence band density of states have been replaced by a "staircase" representation of discrete levels (Fig. 28*c*), each corresponding to a constant density of states per unit area given by

$$\frac{dN}{dE} = \frac{m^*}{\pi\hbar^2}, \tag{27}$$

Since the density of states is constant, rather than gradually increasing from zero, as in a conventional laser, there is a group of electrons of nearly the same energy available to recombine with a group of holes of nearly the same energy, for example, the level E_1 in the conduction with the level E_{hh1} in the valence band. QW lasers offer significant improvement in laser performance, such as reduction in the threshold current, high output power, and high speed, compared with the conventional DH lasers. QW lasers made in GaAs/ AlGaAs material systems have threshold current densities as low as 65 A/cm^2 and submilliampere threshold currents. These lasers operate at emission wavelengths around 0.9 μm.

For longer-wavelength operation, GaInAs/GaInAsP multiple–quantum-well (MQW) lasers for the 1.3 μm and 1.5 μm wavelength regions have been developed. Figure 29*a* shows a schematic diagram of a separate-confinement-heterostructure (SCH) MQW laser where four QWs of GaInAs with GaInAsP barrier layers are sandwiched between the InP cladding layers to form a waveguide with a step index change.[20] These alloy compositions are chosen so that they are lattice matched to the InP substrate. The active region is composed of four 8 nm thick, undoped GaInAs QWs (with E_g of 0.75 eV) separated by 30 nm thick undoped GaInAsP layers (with E_g of 0.95 eV). Figure 29*b* shows the corresponding band diagram of the active region. The n- and p-cladding InP layers are doped with sulfur (10^{18} cm^{-3}) and zinc (10^{17}cm^{-3}), respectively.

A graded-index SCH (GRIN-SCH) is shown in Fig. 29*c*, in which a GRIN of the waveguide is accomplished by several small stepwise increases of the bandgap energies of multiple cladding layers. The GRIN-SCH structure confines both the carriers and the optical field more effectively than the SCH structure and, consequently, leads to an even lower-threshold current density. With the MQW structure, a variety of advanced lasers and photonic integrated circuits becomes possible for future system applications.

[§] In GaAs, the effective mass for heavy holes is 0.62 m_0, whereas that for light holes is 0.074 m_0.

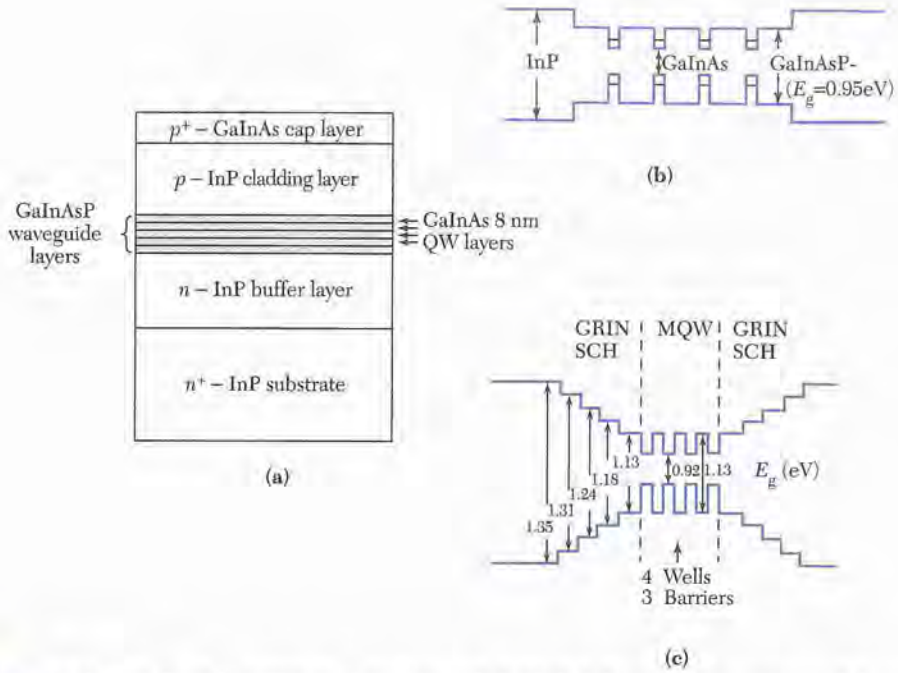


Fig. 29 (*a*) Schematic of the cross section of an GaInAs/GaInAsP multiple-quantum-well laser structure. (*b*) Schematic of the bandgaps of the SCH-MQW layers shown in (*a*). (*c*) GRIN-SCH-MQW structure with thin layers of increasing bandgaps to approximate the graded-index change. [20]

9.4 PHOTODETECTOR

Photodetectors are semiconductor devices that can convert optical signals into electrical signals. The operation of a photodetector involves three steps: carrier generation by incident light, carrier transport and/or multiplication by whatever current gain mechanism may be present, and interaction of current with the external circuit to provide the output signal.

Photodetectors have a broad range of applications, including infrared sensors in optoisolators and detectors for optical-fiber communications. For these applications, the photodetectors must have high sensitivity at the operating wavelengths, high response speed, and low noise. In addition, the photodetector should be compact, use low biasing voltages or currents, and be reliable under the required operating conditions.

9.4.1 Photoconductor

A photoconductor consists simply of a slab of semiconductor with ohmic contacts at both ends of the slab (Fig. 30). When incident light falls on the surface of the photoconductor, electron-hole pairs are generated either by band-to-band transition (intrinsic) or by transitions involving forbidden-gap energy levels (extrinsic), resulting in an increase in conductivity.

For the intrinsic photoconductor, the conductivity is given by

$$\sigma = q\left(\mu_n n + \mu_p p\right), \tag{28}$$

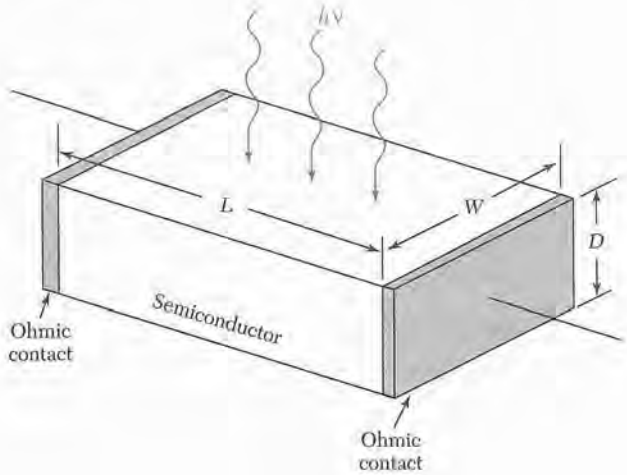


Fig. 30 Schematic diagram of a photoconductor that consists of a slab of semiconductor and two contacts at the ends.

and the increase in conductivity under illumination is mainly due to the increase in the number of carriers. The long-wavelength cutoff for an intrinsic photoconductor is given by Eq. 9. For the extrinsic photoconductor, photoexcitation may occur between the band edge and an energy level in the energy gap. In this case, the long-wavelength cutoff is determined by the depth of the forbidden-gap energy level.

Consider the operation of a photoconductor under illumination. At time zero, the number of carriers generated in a unit volume by a given photon flux is n_0. At a later time t, the number of carriers $n(t)$ in the same volume decays by recombination as

$$n = n_0 \exp\left(\frac{-t}{\tau}\right), \tag{29}$$

where τ is the carrier lifetime. From Eq. 29 the recombination rate is

$$\left|\frac{dn}{dt}\right| = \frac{1}{\tau} n_0 \exp\left(-\frac{t}{\tau}\right) = \frac{n}{\tau}. \tag{30}$$

If we assume a steady flow of photon flux impinging uniformly on the surface of a photoconductor (Fig. 30) with an area $A = WL$, the total number of photons arriving at the surface is $(P_{opt}/h\nu)$ per unit time, where P_{opt} is the incident optical power and $h\nu$ is the photon energy. At steady state, the carrier-generation rate G must be equal to the recombination rate n/τ. If the detector thickness D is much larger than the light penetration depth $1/\alpha$ the total steady-state carrier-generation rate per unit volume is

$$G = \frac{n}{\tau} = \frac{\eta\left(P_{opt}/h\nu\right)}{WLD}, \tag{31}$$

where η is the quantum efficiency, the number of carriers generated per photon, and n is the carrier density, the number of carriers per unit volume. The photocurrent flowing between the electrodes is

$$I_p = \left(\sigma\mathscr{E}\right) WD = \left(q\mu_n n\mathscr{E}\right) WD = \left(qnv_d\right)WD, \tag{32}$$

where $\mathscr{E}$ is the electric field inside the photocoductor and v_d is the carrier drift velocity. Substituting n in Eq. 31 into Eq. 32 gives

$$I_p = q\left(\eta \frac{P_{opt}}{h\nu}\right)\cdot\left(\frac{\mu_n \tau \mathscr{E}}{L}\right). \tag{33}$$

If we define the primary photocurrent as

$$I_{ph} \equiv q\left(\eta \frac{P_{opt}}{h\nu}\right), \tag{34}$$

the photocurrent gain from Eq. 33 is

$$\boxed{\text{Gain} \equiv \frac{I_p}{I_{ph}} = \frac{\mu_n \tau \mathscr{E}}{L} = \frac{\tau}{t_r},} \tag{35}$$

where $t_r \equiv L/v_d = L/\mu_n \mathscr{E}$ is the carrier transit time. The gain depends on the ratio of carrier lifetime to the transit time.

EXAMPLE 7

Calculate the photocurrent and gain when 5×10^{12} photons/s arriving at the surface of a photoconductor of $\eta = 0.8$. The minority carrier lifetime is 0.5 ns, and the device has $\mu_n = 2500$ cm²/V-s, $\mathscr{E} = 5000$ V/cm, and $L = 10$ μm.

SOLUTION From Eq. 33,

$$I_p = q\left(0.8 \times 5 \times 10^{12} \text{ photons/s}\right)\cdot\left(\frac{2500 \text{ cm}^2/\text{V-s} \cdot 5 \times 10^{-10}\text{s} \cdot 5000 \text{ V/cm}}{10 \times 10^{-4}\text{cm}}\right),$$

$$= 4 \times 10^{-6}\text{A} = 4 \ \mu\text{A},$$

and from Eq. 35,

$$\text{Gain} = \frac{\mu_n \tau \mathscr{E}}{L} = \frac{2500 \cdot 5 \times 10^{-10} \cdot 5000}{10 \times 10^{-4}} = 6.25.$$

For a sample with long minority-carrier lifetime and short electrode spacing, the gain can be substantially greater than unity. Gains as high as 10^6 can be obtained from some photoconductors. The response time of a photoconductor is determined by the transit time t_r. To achieve short transit time, small electrode spacing and a high electric field must be used. The response times of photoconductors cover a wide range, from 10^{-3} to 10^{-10} seconds. They are extensively used for infrared detection especially for wavelengths greater than a few microns.

9.4.2 Photodiode

A photodiode is basically a p-n junction or a metal-semiconductor contact operated under reverse bias. When an optical signal impinges on the photodiode, the depletion region serves to separate the photogenerated electron-hole pairs and an electric current flows

in the external circuit. For high-frequency operation, the depletion region must be kept thin to reduce the transit time. On the other hand, to increase the quantum efficiency, the depletion layer must be sufficiently thick to allow a large fraction of the incident light to be absorbed. Thus, there is a trade-off between the response speed and quantum efficiency.

Quantum Efficiency

The quantum efficiency as mentioned previously is the number of electron-hole pairs generated for each incident photon:

$$\boxed{\eta = \left(\frac{I_p}{q}\right) \cdot \left(\frac{P_{opt}}{h\nu}\right)^{-1},} \tag{36}$$

where I_p is the photogenerated current from the absorption of incident optical power P_{opt} at a wavelength λ (corresponding to a photon energy $h\nu$). One of the key factors that determine η is the absorption coefficient α (Fig. 5). Since α is a strong function of the wavelength, the wavelength range in which appreciable photocurrent can be generated is limited. The long-wavelength cutoff λ_c is established by the bandgap, Eq. 9, and is about 1.8 μm for germanium and 1.1 μm for silicon. For wavelengths longer than λ_c, the values of α are too small to give appreciable band-to-band absorption. The short-wavelength cutoff of the photoresponse comes about because the values of α are very large (~10^5 cm^{-1}), and hence the radiation is mostly absorbed very near the surface where recombination time is short. Therefore, the photocarriers can recombine before they can be collected in the p–n junction.

Figure 31 shows typical plots of quantum efficiency versus wavelength for some high-speed photodiodes.[21,22] Note that in the ultraviolet and visible region, metal-semiconductor photodiodes show good quantum efficiencies. In the near-infrared region, silicon photodiodes (with an antireflection coating) can reach 100% quantum efficiency near the 0.8- to 0.9-μm region. In the 1.0- to 1.6-μm region, germanium photodiodes and Group

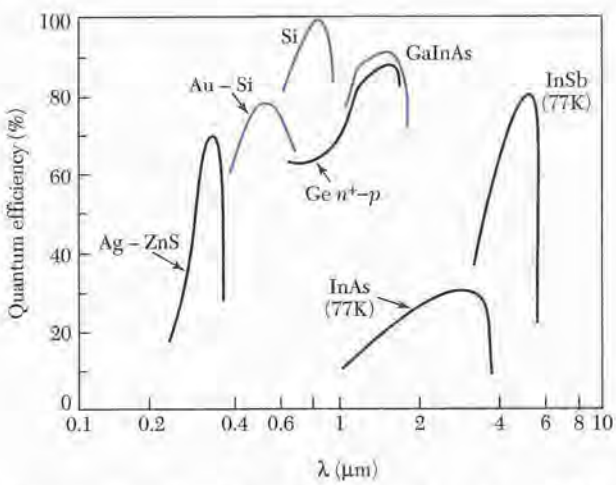


Fig. 31 Quantum efficiency versus wavelength for various photodetectors.[21,22]

III-V photodiodes (e.g., GaInAs) have shown high quantum efficiencies. For even longer wavelengths, photodiodes are cooled (e.g., to 77 K) for high-efficiency operation.

Response Speed

The response speed is limited by three factors: (1) diffusion of carriers, (2) drift time in the depletion region, and (3) capacitance of the depletion region. Carriers generated outside the depletion region must diffuse to the junction resulting in considerable time delay. To minimize the diffusion effect, the junction should be formed very close to the surface. The greatest amount of light will be absorbed when the depletion region is wide. However, the depletion layer must not be too wide or transit time effects will limit the frequency response. It also should not be too thin, or excessive capacitance C will result in a large RC time constant, where R is the load resistance. The optimal compromise is the width at which the depletion layer transit time is approximately one half the modulation period. For example, for a modulation frequency of 2 GHz, the optimal depletion layer thickness in silicon (with a saturation velocity of 10^7 cm/s) is about 25 μm.

p-i-n Photodiode

The *p-i-n* photodiode is one of the most common photodetectors because the depletion region thickness (the intrinsic layer) can be tailored to optimize the quantum efficiency and frequency response. Figure 32*a* shows a cross section of a *p-i-n* photodiode that has an antireflection coating to increase quantum efficiency.

Figures 32*b* and 32*c* show the energy band diagram of the *p-i-n* diode under reverse-bias condition and its optical absorption characteristics. Light absorption in the semiconductor produces electron-hole pairs. Pairs produced in the depletion region or within a diffusion length of it will eventually be separated by the electric field as shown in Fig. 32*b*, whereby a current flows in the external circuit as carriers drift across the depletion layer.

EXAMPLE 8

On reaching the surface of the semiconductor, the incident optical power P_0 will have its level reduced to $P_0(1-R)$ on entering the material, where R is the reflection coefficient. On passing though the semiconductor the light will be absorbed, and so at any depth x the amount of residual optical power $P(x)$ is given by $P(x) = P_0(1-R)\exp(-\alpha x)$. For $\alpha = 10^4$ cm^{-1} and $R = 0.1$, calculate the depth at which half the incident optical power has been absorbed in a material.

SOLUTION

$$x = \frac{-1}{\alpha}\ln\left[\frac{P(x)}{P_0(1-R)}\right] = -10^{-4}\cdot\ln\left(\frac{1}{2\times 0.9}\right)\ \text{cm}$$

$$= 0.59\ \mu\text{m}.$$

Metal-Semiconductor Photodiode

The construction of a high-speed metal-semiconductor photodiode is shown in Fig. 33. To avoid large reflection and absorption losses when the diode is illuminated through the metal contact, the metal film must be very thin (~10 nm) and an antireflection coating must be used. Metal-semiconductor photodiodes are particularly useful in the ultraviolet- and visible-light regions. In these regions the absorption coefficients, α, in most of the common semiconductors are very high, of the order of 10^4 cm^{-1} or more, which corresponds to an effective absorption length $1/\alpha$ of 1.0 μm or less. It is possible to choose

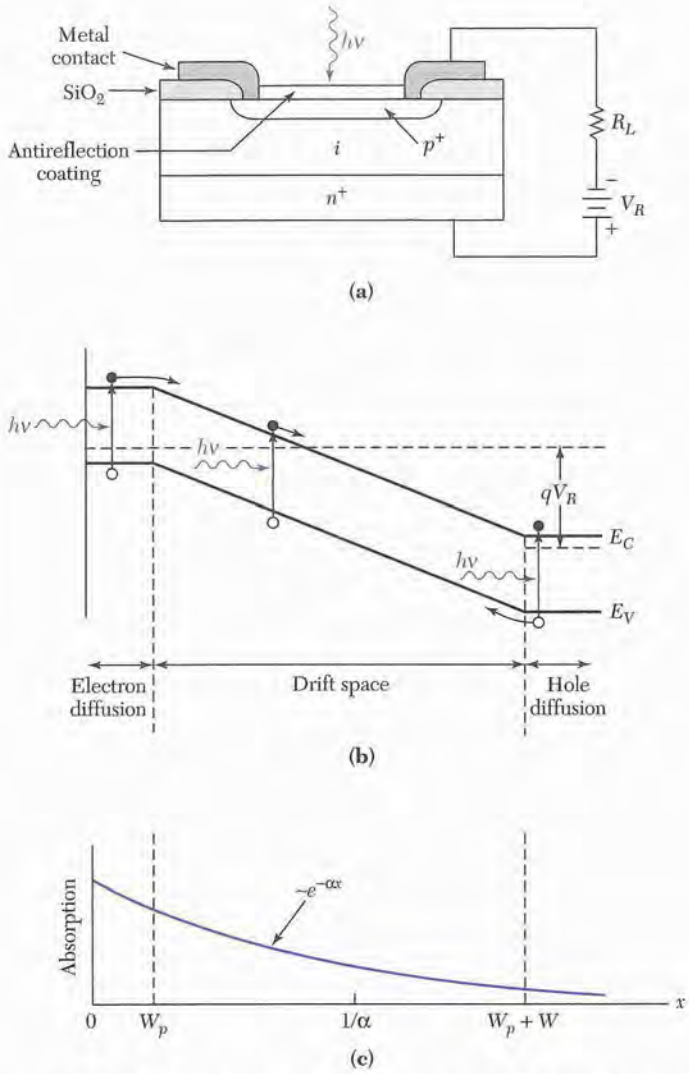


Fig. 32 Operation of a *p-i-n* photodiode (*a*) Cross-section view of *p-i-n* photodiode. (*b*) Energy band diagram under reverse bias. (*c*) Carrier absorption characteristics.

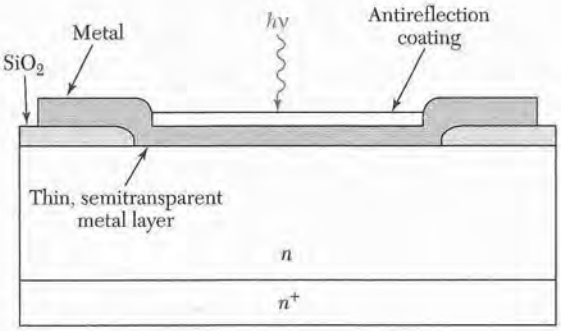


Fig. 33 Metal-semiconductor photodiode.

a metal and an antireflection coating so that a large fraction of the incident radiation will be absorbed near the surface of the semiconductor. As an example, for a gold-silicon photodetector having 10 nm gold and 50 nm zinc sulfide as the antireflection coating, more than 95% of the incident light with $\lambda = 0.6328$ μm (helium-neon laser wavelength, red light) will be transmitted into the silicon substrate.

Heterojunction Photodiode

Another photodiode structure is a heterojunction device formed by depositing a large-bandgap semiconductor epitaxially on a smaller-bandgap semiconductor. One advantage of a heterojunction photodiode is that the quantum efficiency does not depend critically on the distance of the junction from the surface, because the large-bandgap material can be used as a window for the transmission of optical power. In addition, the heterojunction can provide unique material combinations so that the quantum efficiency and response speed can be optimized for a given optical-signal wavelength.

To obtain a heterojunction with low leakage current, the lattice constants of the two semiconductors must be closely matched. Ternary III-V compounds $Al_xGa_{1-x}As$ epitaxially grown on GaAs can form heterojunctions with perfectly matched lattices. These heterojunction are important for photonic devices operated in the wavelength range from 0.65 to 0.85 μm. At longer wavelengths (1 to 1.6 μm), ternary compounds such as $Ga_{0.47}In_{0.53}As$ (with $E_g = 0.75$ eV) and quaternary compounds such as $Ga_{0.27}In_{0.73}As_{0.63}P_{0.37}$ (with $E_g = 0.95$ eV) can be used. These compounds have a nearly perfect lattice match to an InP substrate. The quantum efficiency is greater than 70% over the wavelength range from 1 to 1.6 μm as shown in Fig. 31 (GaInAs curve).

9.4.3 Avalanche Photodiode

An avalanche photodiode (APD) is operated under a reverse-bias voltage that is sufficient to enable avalanche multiplication. The multiplication results in internal current gain and the device can respond to light modulated at frequencies as high as microwave frequencies.

One important consideration in the design of an APD is the need to minimize avalanche noise. The avalanche noise comes about from the random nature of the avalanche multiplication process in which every electron-hole pair generated at a given distance in the depletion region does not experience the same multiplication. The avalanche noise depends on the ratio of the ionization coefficients α_p/α_n; the smaller the ratio, the smaller is the avalanche noise. This is because when $\alpha_p = \alpha_n$, each incident photocarrier results in three carriers in the multiplicating region: the primary carrier and its secondary hole and electron. A fluctuation that changes the number of carriers by one represents a large percentage change, and the noise will be large. On the other hand, if one of the ionization coefficients approaches zero (e.g., $\alpha_p \to 0$), each incident photocarrier can result in a large number of carriers in the multiplication region. In this case, a fluctuation of one carrier is a relatively insignificant perturbation. To minimize the avalanche noise, we should use semiconductors with a large difference in α_p and α_n. The noise factor is given by

$$F = M\left(\frac{\alpha_p}{\alpha_n}\right) + \left(2 - \frac{1}{M}\right)\left(1 - \frac{\alpha_p}{\alpha_n}\right) \tag{37}$$

where M is the multiplication factor. We can see from Eq. 37 that when $\alpha_p = \alpha_n$, the noise factor has a maximum value of M; while for $\alpha_p/\alpha_n = 0$ and for a large M, the minimum noise factor is 2.

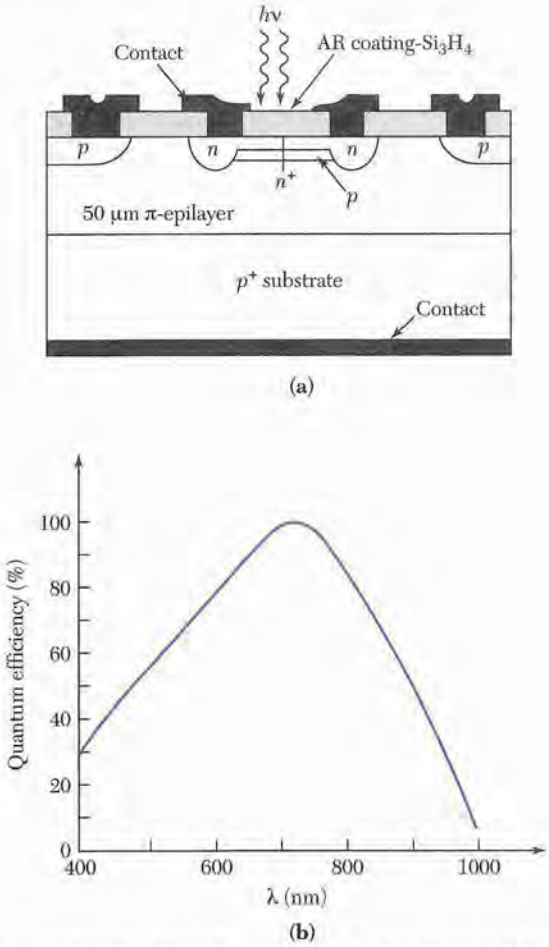


Fig. 34 A typical silicon avalanche photodiode: (*a*) device structure and (*b*) quantum efficiency.

Figure 34*a* shows the structure of a typical silicon APD having a n^+-p-π-p^+ doping profile (π is a lightly doped p-region). The quantum efficiency is near 100 % at a wavelength of about 0.75 μm for a device having a SiO_2-Si_3N_4 antireflection coating (Fig. 34*b*). Because the ratio of α_p/α_n is about 0.04, the noise factor obtained from Eq. 37 is 2.3 for $M = 10$.

9.5 SOLAR CELL

Solar cells are useful for both space and terrestrial applications. Solar cells furnish the long-duration power supply for satellites. The solar cell is an important candidate for an alternative terrestrial energy source because it can convert sunlight directly to electricity with good conversion efficiency, can provide nearly permanent power at low operating cost, and is virtually nonpolluting.[23,24]

9.5.1 Solar Radiation

The radiative energy output from the sun derives from a nuclear fusion reaction. In every second, about 6×10^{11} kg hydrogen is converted to helium, with a net mass loss of about

4×10^3 kg. The mass loss is converted through the Einstein relation ($E = mc^2$) to 4×10^{20} J. This energy is emitted primarily as electromagnetic radiation in the ultraviolet to infrared region (0.2 to 3 μm). The total mass of the sun is now about 2×10^{30} kg, and a reasonably stable life with a nearly constant radiative-energy output of over 10 billion (10^{10}) years is projected.

The intensity of solar radiation outside the earth's atmosphere, at the average distance of its orbit around the sun, is defined as the solar constant and has a value of 1367 W/m^2. Terrestrially, the sunlight is attenuated by clouds and by atmospheric scattering and absorption. The attenuation depends primary on the length of the light's path through the atmosphere, or the mass of air through which it passes. The *"air mass"* is defined as $1/\cos\phi$, where ϕ is the angle between the vertical and the sun's position.

EXAMPLE 9

The air mass can most easily be estimated from the length of the shadow, s, of a vertical structure of height, h, as $\sqrt{1 + (s/h)^2}$. If s = 1.118 m and h = 1.00 m, find the air mass.

SOLUTION

$$\sqrt{1+(1.118/1.0)^2} = \sqrt{2.25} = 1.5.$$

We have an air mass 1.5 (AM 1.5). The corresponding $\cos\phi$ is 1/1.5 = 0.667 and the angle ϕ between the vertical and the sun's position is $\cos^{-1}(0.667) = 48^\circ$. The maximum sunlight intensity occurs when the sun is straight overhead (i.e., AM 1.0 with $\phi = 0^\circ$)

Figure 35 shows two curves related to solar spectral irradiance (power per unit area per unit wavelength).[25] The upper curve, which represents the solar spectrum outside the Earth's atmosphere, is the air mass zero condition (AM0). The AM0 spectrum is relevant for satellite and space vehicle applications. Terrestrial solar-cell performance is

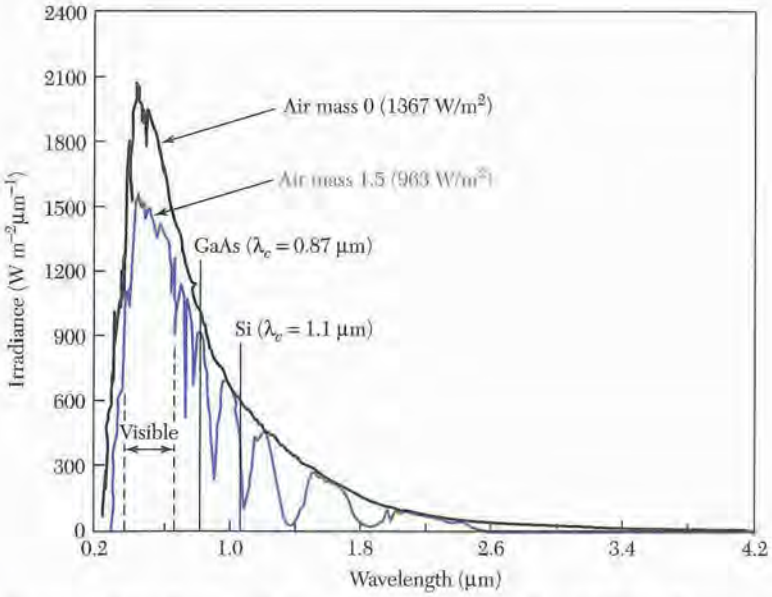


Fig. 35 Solar spectral irradiance [25] at air mass 0 and air mass 1.5 and the cutoff wavelength of GaAs and Si.

specified with reference to the air mass 1.5 (AM1.5) spectrum. This spectrum represents the sunlight at the Earth's surface when the sun is at an angle of 48° from the vertical. At this angle the incident power is about 963 W/m^2.

9.5.2 *p–n* Junction Solar Cell

A schematic representation of a *p–n* junction solar cell is shown in Fig. 36. It consists of a shallow *p–n* junction formed on the surface, a front ohmic contact stripe and fingers, a back ohmic contact that covers the entire back surface, and an antireflection coating on the front surface.

When the cell is exposed to the solar spectrum, a photon that has an energy less than the bandgap E_g makes no contribution to the cell output. A photon that has energy greater than E_g contributes an energy E_g to the cell output. Energy greater than E_g is wasted as heat. To derive the conversion efficiency, we shall consider the energy band diagram of a *p–n* junction under solar radiation shown in Fig. 37*a*. The equivalent circuit is shown in Fig. 37*b*, where a constant-current source is in parallel with the junction. The source I_L results from the excitation of excess carriers by solar radiation, I_s is the diode saturation current, and R_L is the load resistance.

The ideal *I-V* characteristics of such a device are given by

$$I = I_s\left(e^{qV/kT} - 1\right) - I_L \tag{38}$$

and

$$J_s = \frac{I_s}{A} = qN_C N_V\left(\frac{1}{N_A}\sqrt{\frac{D_n}{\tau_n}} + \frac{1}{N_D}\sqrt{\frac{D_p}{\tau_p}}\right)\cdot e^{-E_g/kT}, \tag{38a}$$

where A is the device area. A plot of Eq. 38 is given in Fig. 38*a* for I_L = 100 mA, I_s = 1 nA, cell area A = 4 cm^2, and T = 300 K. The curve passes through the fourth quadrant

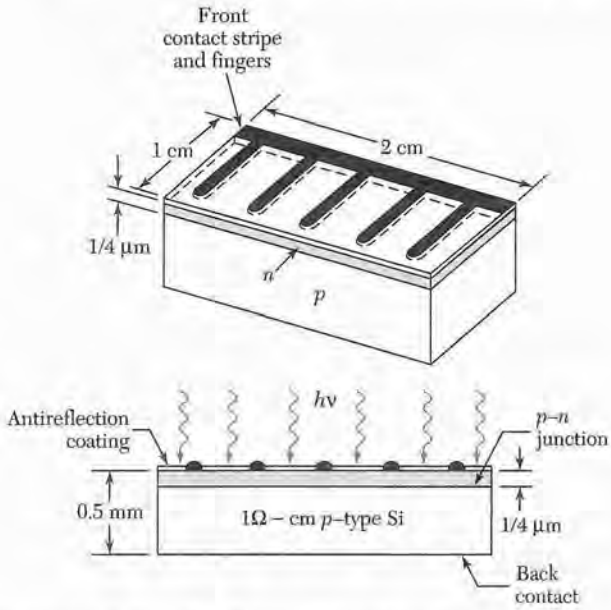


Fig. 36 Schematic representation of a silicon *p–n* junction solar cell. [23]

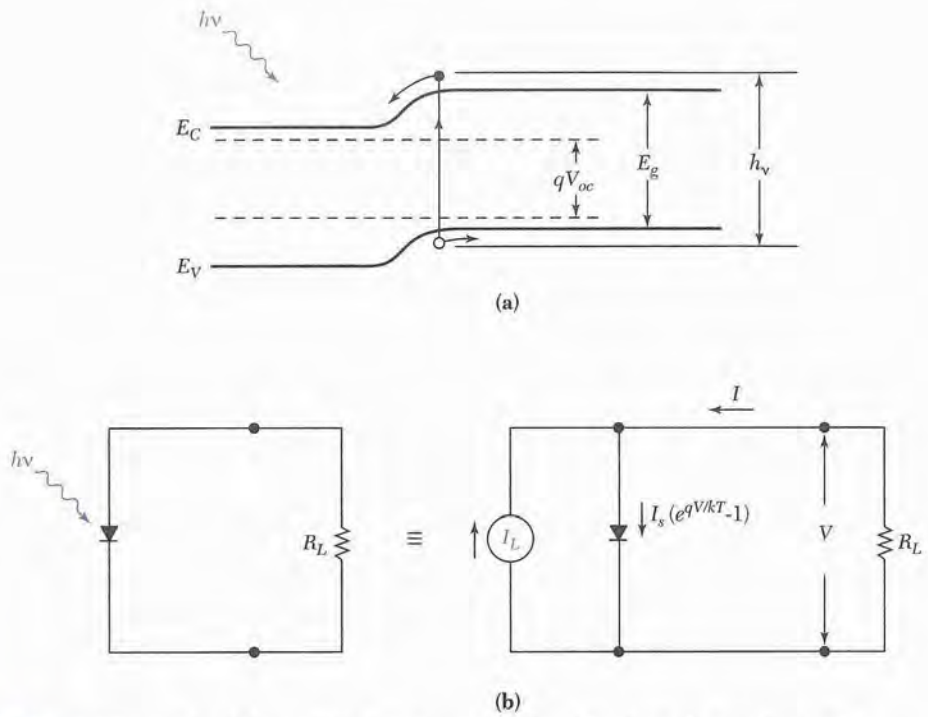


Fig. 37 (*a*) Energy band diagram of a *p–n* junction solar cell under solar irradiation. (*b*) Idealized equivalent circuit of a solar cell.

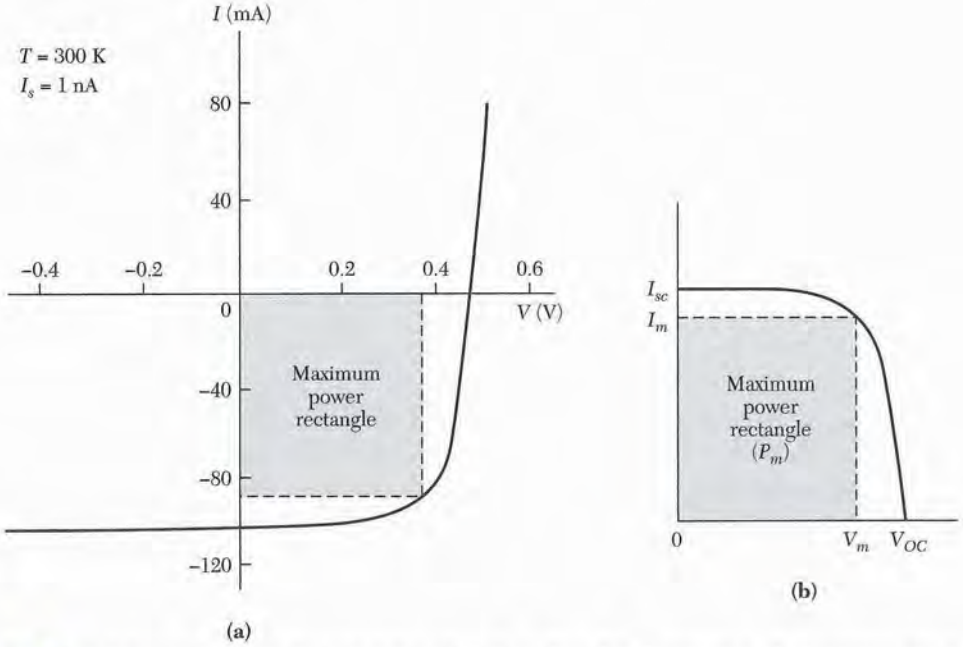


Fig. 38 (*a*) Current-voltage characteristics of a solar cell under illumination. (*b*) Inversion of (*a*) about the voltage axis.

and ,therefore, power can be extracted from the device. The *I-V* curve is more generally represented by Fig. 38*b*, which is an inversion of Fig. 38*a* about the voltage axis. By choosing a proper load, close to 80% of the product $I_{SC}V_{OC}$ can be extracted, where I_{SC} is the short-circuit current equal to I_L and V_{OC} is the open-circuit voltage of the cell; the shaded area in the figure is the maximum-power rectangle. Also defined in Fig. 38*b* are the quantities I_m and V_m that correspond to the current and voltage, respectively, for the maximum power output $P_m (= I_m \times V_m)$.

From Eq. 38 we obtain for the open-circuit voltage ($I = 0$)

$$\boxed{V_{OC} = \frac{kT}{q}\ln\left(\frac{I_L}{I_s}+1\right) \cong \frac{kT}{q}\ln\left(\frac{I_L}{I_s}\right).} \tag{39}$$

Hence, for a given I_L, V_{OC} increases logarithmically with decreasing saturation current I_s. The output power is given by

$$P = IV = I_s V\left(e^{qV/kT}-1\right) - I_L V. \tag{40}$$

The condition for maximum power is obtained when $dP/dV = 0$, or

$$V_m = \frac{kT}{q}\ln\left[\frac{1+\left(I_L/I_s\right)}{1+\left(qV_m/kT\right)}\right] \cong V_{OC} - \frac{kT}{q}\ln\left(1+\frac{qV_m}{kT}\right), \tag{41a}$$

$$I_m = I_s\left(\frac{qV_m}{kT}\right)e^{qV_m/kT} \cong I_L\left(1-\frac{1}{qV_m/kT}\right). \tag{41b}$$

The maximum output power P_m is then

$$\boxed{P_m = I_m V_m \cong I_L\left[V_{OC} - \frac{kT}{q}\ln\left(1+\frac{qV_m}{kT}\right) - \frac{kT}{q}\right].} \tag{42}$$

EXAMPLE 10

Calculate the open-circuit voltage and the output power at a voltage of 0.35 V for a solar cell shown in Fig. 38*a*.

SOLUTION From Eq. 39

$$V_{oc} = (0.026\ \text{V})\ln\left(\frac{100\times10^{-3}\,\text{A}}{1\times10^{-9}\,\text{A}}\right) = 0.48\ \text{V}.$$

The output power at 0.35 V is given by Eq. 40 (note that I_s and I_L are reverse currents so we need negative signs for them):

$$P = \left(-10^{-9}\,\text{A}\right)\cdot(0.35\ \text{V})\left(e^{0.35/0.026}-1\right) - \left(-0.1\ \text{A}\right)\cdot(0.35\ \text{V}) = 3.48\times10^{-2}\ \text{W}.$$

9.5.3 Conversion Efficiency

Ideal efficiency

The power conversion efficiency of a solar cell is given by

$$\eta = \frac{I_m V_m}{P_{in}} = \frac{I_L\left[V_{OC} - \frac{kT}{q}\ln\left(1 + \frac{qV_m}{kT}\right) - \frac{kT}{q}\right]}{P_{in}} \tag{43}$$

or

$$\boxed{\eta = \frac{FF \cdot I_L V_{OC}}{P_{in}},} \tag{43a}$$

where P_{in} is the incident power and FF is the fill factor defined as

$$FF \equiv \frac{I_m V_m}{I_L V_{OC}} = 1 - \frac{kT}{qV_{OC}}\ln\left(1 + \frac{qV_m}{kT}\right) - \frac{kT}{qV_{OC}}. \tag{44}$$

The fill factor is the ratio of the maximum power rectangle (Fig. 38*b*) to the rectangle of $I_{SC} \times V_{OC}$. To maximize the efficiency, we should maximize all three items in the numerator of Eq. 43*a*.

The ideal efficiency can be obtained from the ideal *I-V* characteristics defined by Eq. 38. For a given semiconductor, the saturation current density is obtained from Eq. 38*a*. For a given air mass condition (e.g., AM 1.5), the short-circuit current I_L is the product q and the number of the available photons with energy $h\nu \geq E_g$ in the solar spectrum. Once I_s and I_L are known, the output power P and the maximum power P_m can be obtained from Eqs. 40 to 42. The input power P_{in} is the integration of all the photons in the solar spectrum (Fig. 35). Under AM 1.5 condition, the efficiency P_m/P_{in} has a broad maximum[24,26] of about 29% and does not depend critically on E_g. Therefore, semiconductors with bandgap between 1 and 2 eV can all be considered solar cell materials.

Spectrum Splitting

The simplest way to exceed the efficiency limit is by "spectrum splitting". By splitting sunlight into narrow wavelength bands and directing each band to a cell that has a bandgap optimally chosen to convert just this band, as shown in Fig. 39*a*, efficiency above 60% is possible in principle.[27] Fortunately, simply stacking cell on top of one another with the highest bandgap cell uppermost, as in Fig. 39*b*, automatically achieves an identical spectral-splitting effect, making this "tandem" cell approach a reasonably practical way of increasing cell efficiency.

Series Resistance and Recombination Current

Many factors degrade the ideal efficiency. One of the major factors is the series resistance R_s from the ohmic loss in the front surface. The equivalent circuit is shown in Fig.40. From the ideal diode current given by Eq. 38, the *I-V* characteristics are found to be

$$\ln\left(\frac{I + I_L}{I_s} + 1\right) = \frac{q}{kT}\left(V - IR_s\right). \tag{45}$$

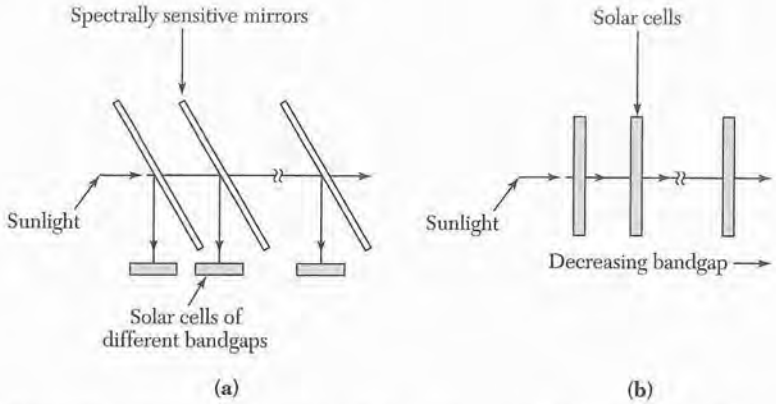


Fig. 39 Multigap cell concepts. (*a*) Spectrum splitting approach. (*b*) Tandem-cell approach.[27]

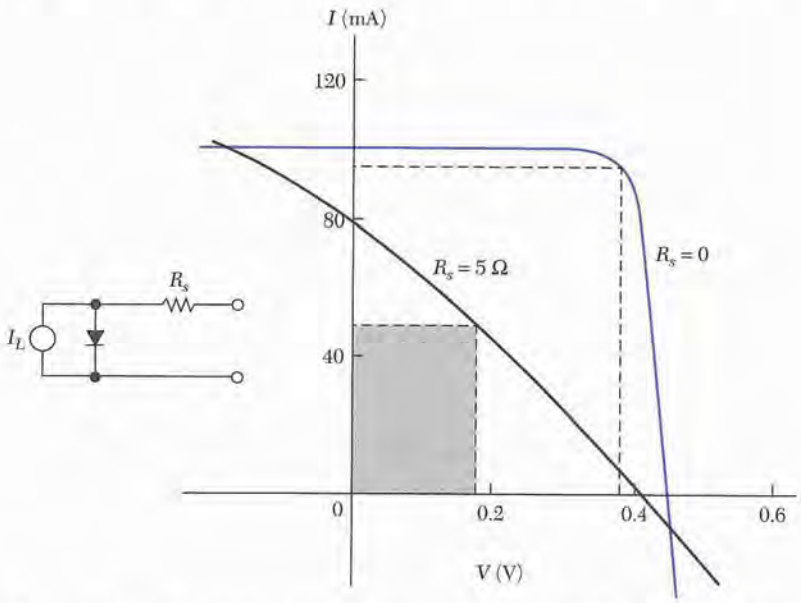


Fig. 40 Current-voltage characteristics and the equivalent circuit of solar cells that have resistances.

Plots of this equation are shown in Fig. 40, with $R_s = 0$ and 5 Ω and where the other parameter I_s, I_L, and T are the same as those in Fig. 38. It can be seen that a series resistance of only 5Ω reduces the available power to less than 30% of the maximum power with $R_s = 0$. The output current and output power are

$$I = I_s\left\{\exp\left[\frac{q(V - IR_s)}{kT}\right] - 1\right\} - I_L, \tag{46}$$

$$P = I\left[\frac{kT}{q}\ln\left(\frac{I + I_L}{I_s}\right) + IR_s\right]. \tag{47}$$

The series resistance depends on the junction depth, the impurity concentrations of p-type and n-type regions, and the arrangement of the front-surface ohmic contacts. For

a typical silicon solar cell with the geometry shown in Fig. 36, the series resistance is about 0.7 Ω for n^+-p cells and 0.4 Ω for p^+-n cells. The difference in resistance is mainly the result of the lower resistivity in n-type substrates.

Another factor is the recombination current in the depletion region. For single-level centers, the recombination current can be expressed as

$$I_{rec} = I'_s\left[\exp\left(\frac{qV}{2kT}\right) - 1\right] \tag{48}$$

and

$$\frac{I'_s}{A} = \frac{qn_iW}{\sqrt{\tau_p\tau_n}}, \tag{48a}$$

where I'_s is the saturation current. The energy conversion equation can be put into a closed form yielding equations similar to Eqs. 39 through 42 with the exception that I_s is replaced by I'_s and the exponential factor is divided by 2. The efficiency for the recombination current case is found to be much less than the ideal current due to the degradation of both V_{OC} and the fill factor. For silicon solar cells at 300 K, the recombination current can cause a 25% reduction in efficiency.

9.5.4 Silicon and Compound-Semiconductor Solar Cells

Silicon is the most important semiconductor for solar cells. It is nontoxic and is second only to oxygen in the earth's crust. Therefore, silicon poses minimal environmental or resource depletion risks if used on a large scale. It also has a well established technological base because of its use in microelectronics.

III-V compound semiconductors and their alloy systems provide wide choices of bandgaps with closely matched lattice constants. These compounds are ideal for producing tandem solar cells. For example, AlGa/GaAs, GaInP/GaAs, and GaInAs/InP material systems have been developed for solar cells in satellite and space vehicle applications.

PERL Cell

The silicon *p*assivated *e*mitter *r*ear *l*ocally-diffused (PERL) cell[28] is shown in Fig.41. The cell has inverted pyramids on the top that are formed by using anisotropic etches to expose the slowly etching (111) crystallographic planes. The pyramids reduce reflections of light incident on the top surface, since light incident perpendicularly to the cell will strike one of the inclined (111) planes obliquely, and will be refracted obliquely into the cell. The rear contact is separated from the silicon by an intervening oxide layer. This gives much better rear reflection than an aluminum layer. To date, the PERL cell shows the highest conversion efficiency of 24%.

Amorphous Si Solar Cell

The basic cell structure for a series interconnected *a*-Si solar cells is shown[39] in Fig. 42. A layer of SiO_2 followed by a transparent conducting layer of a large bandgap, degenerately doped semiconductor such as SnO_2 is deposited onto a glass substrate and patterned using a laser. The substrate is then coated by a *p-i-n* junction stack of amorphous silicon by the decomposition of silane in a radio-frequency plasma-discharge system. After deposition, the *a*-Si layers are patterned in a second laser system. A layer of aluminum is sputtered onto the rest of the silicon and this layer is also patterned by laser. This technique forms a series of interconnected cells, as shown in Fig. 42. The cell has the lowest manufacturing cost and has a modest efficiency of 5%.

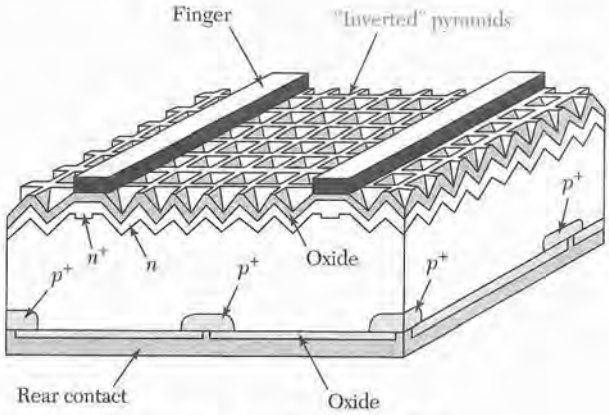


Fig. 41 Passivated emitter rear locally diffused (PERL) cell.[24]

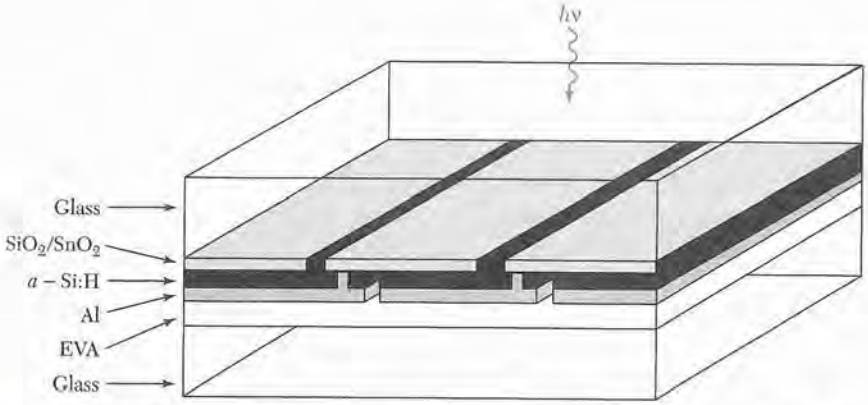


Fig. 42 Series-interconnected a-Si solar cells deposited on a glass substrate with a rear glass cover bonded using ethylene vinyl acetate (EVA).[29]

Tandem Solar Cell

Figure 43 shows the structure of a monolithic tandem solar cell.[24] A p-type germanium is used as the substrate, which has a lattice constant very close to that of GaAs and $Ga_{0.51}In_{0.49}P$. A p-GaInP layer is then grown to reduce the minority-carrier concentration near the rear contact. The bottom junction is the GaAs p-n junction (E_g = 1.42 eV), and the top junction is the GaInP junction (E_g = 1.9 eV). A tunneling p^+-n^+ GaAs junction is placed between the top and bottom junctions to connect the cells. Tandem solar cells with efficiency as high as 30% have been obtained.

9.5.5 Optical Concentration

Sunlight can be focused by using mirrors and lenses. Optical concentration offers an attractive and flexible approach to reducing high cell costs by substituting a concentrator area for much of the cell area. It also offers other advantages, such as a 20% increase in efficiency for a concentration of 1000 suns (an intensity of 963×10^3 W/m^2).

Figure 44 shows the measured results of a typical silicon solar cell mounted in a concentrated system.[30] Note that device performances improve as the concentration increases

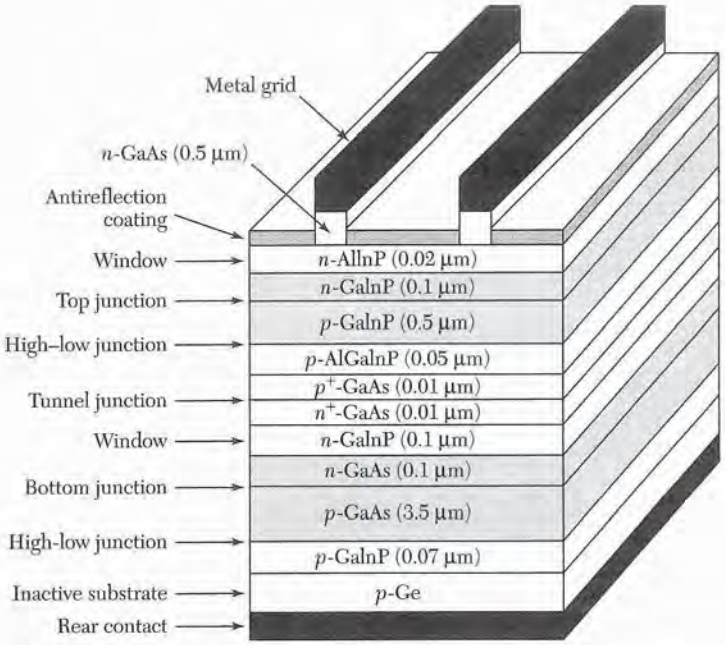


Fig. 43 Monolithic tandem solar cell.[24]

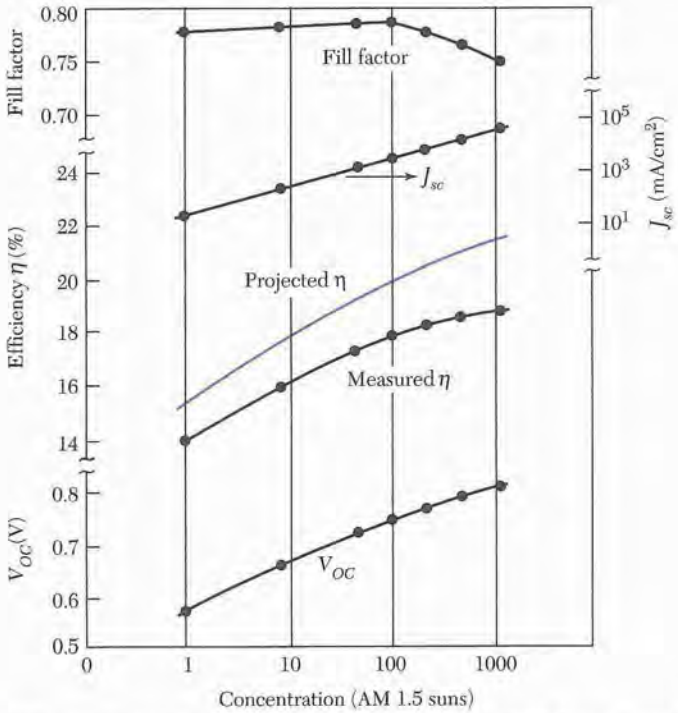


Fig. 44 Efficiency, open-circuit voltage, short-circuit current, and fill factor versus solar concentration.[30]

from one sun toward 1000 suns. The short-circuit current density increases linearly with concentration. The open-circuit voltage increases at a rate of 0.1 V per decade, while the fill factor varies slightly. The efficiency, which is the product of the foregoing three factors divided by the input power, increases at a rate of about 2% per decade. With a proper antireflection coating, we project an efficiency increase of 30% at 1000 suns. Therefore, one cell operated under 1000-sun concentration can produce the same power output as 1300 cells under one sun. Potentially, the optical concentration approach can replace expensive solar cells with less expensive concentrator materials and a related tracking and heat removal system to minimize the overall system cost.

SUMMARY

The operation of the four groups of photonic devices — LEDs, laser diodes, photodetectors, and solar cells depends upon the emission or absorption of photons. Photons are emitted due to the recombination of charge carriers, and photons are absorbed to create charge carriers.

LEDs are *p–n* junctions that can emit spontaneous radiation due to recombinations of electrons and holes in a forward-biased junction. Visible LEDs can emit radiation with photon energies in the range of 1.8 to 2.8 eV, corresponding to wavelengths from 0.7 to 0.4 um. They are used extensively for display and various electronic instruments. By combining LEDs of different colors (i.e., red, green, and blue), we can form white LEDs which are useful for general illumination. Organic semiconductors can also be used for display applications. OLED is particularly useful for multi-color large-area flat-panel displays. Infrared LEDs can emit radiation with $h\nu < 1.8$ eV. They are used for opto-isolator and short-distance optical-fiber communication.

The laser diode is also a *p–n* junction operated under forward bias condition. However, the diode structure must provide the confinement of the carriers and the optical field so that the stimulated emission condition can be established. Laser diodes have evolved from the homojunction, to the distributed feedback configuration, and to quantum-well structures. The main objectives are to lower the threshold current density for lasing and to have all emitted photons at a single frequency. The laser diode is a key device for long distance, optical fiber communication systems. It is also extensively used for video recording, high-speed printing, and optical reading.

Photodetectors include photoconductors, photodiodes and avalanche photodiodes. They can convert optical signals into electrical signals. When photons are absorbed, electron-holes pairs are generated in the device, which are subsequently separated by an electrical field to produce a photo-current flowing between the electrodes. Photodetectors are used for optical sensing and detection in opto-isolators and optical-fiber communication systems.

A solar cell is similar to a photodiode and has the same operational principle. However, the solar cell differs from a photodiode in that it is a large-area device and covers a wide range of optical spectrum (solar radiation). Solar cells furnish the long-duration power supply for satellites. The solar cell is a major candidate for a terrestrial energy source because it can convert sunlight directly to electricity with good efficiency and is environmentally benign. Currently, the important solar cells are the highly efficient silicon PERL cell (24%), the GaInP/GaAs tandem cell (30%), and the low cost, series-interconnected *a*-Si solar cell (5 %).

REFERENCE

1. H. C. Casey, Jr. and M. B. Panish, *Heterostructure Lasers*. Academic, New York, 1978.
2. H. Melchior, "Demodulation and Photodetection Techniques," in F. T. Arecchi and E. O. Schulz-Dubois, Eds., *Laser Handbook*, Vol. 1, North-Holland, Amsterdam, 1972.
3. M. G. Craford, "Recent Developments in LED Technology," *IEEE Trans. Electron Devices*, **ED-24**, 935 (1977).
4. W. O. Groves, A. H. Herzog, and M. G. Craford, "The Effect of Nitrogen Doping on GaAsP Electroluminescent Diodes," *Appl. Phys. Lett.*, **19**, 184 (1971).
5. S. Gage, et al., *Optoelectronic Application Manual*, McGraw-Hill, New York, 1977.
6. S. Nakamura and G. Fasol, *The Blue Laser Diode*, Wiley, New York, 1997.
7. R. H. Saul, T. P. Lee, and C. A. Burrus, "Light-Emitting Diode Device Design," in R. K. Willardon and A. C. Bear, Eds., *Semiconductor and Semimetals*, Academic, New York, 1984.
8. A. A. Bergh and P. J. Dean, *Light Emitting Diodes*, Clarendon, Oxford, 1976.
9. N. Bailey, " The Future of Organic Light-Emitting Diodes," *Inf. Disp.*, **16**, 12 (2000).
10. C. H. Chen, J. Shi, and E. W. Tang, " Recent Development, in Molecular Organic Electroluminescent Materials," *Macromal. Symp.* **125**, 1(1997).
11. S. E. Miller and A. G. Chynoweth, Eds., *Optical Fiber Communications*, Academic, New York, 1979.
12. W. T. Tsang, "High Speed Photonic Devices," in S. M. Sze, Ed., *High Speed Semiconductor Devices*, Wiley, New York, 1990.
13. O. Madelung, Ed., *Semiconductor-Group IV Elements and III-V Compounds*, Springer-Verlag, Berlin, 1991.
14. M. B. Panish, I. Hayashi, and S. Sumski, "Double-Heterostructure Injection Lasers with Room Temperature Threshold As Low As 2300 A/cm^2," *Appl. Phys. Lett.*, **16**, 326(1970).
15. T. E. Bell, "Single-Frequency Semiconductor Lases," *IEEE Spectrum*, **20**, 38(1983).
16. F. Stern, "Calculated Spectral Dependence of Gain in Excited GaAs," *J. Appl. Phys.*, **47**, 5328(1976).
17. W. T. Tsang, R. A. Logan, and J. P. Van der Ziel, "Low-Current-Threshold Stripe-Buried-Heterostructure Laser with Self-Aligned Current Injection Stripes," *Appl. Phys. Lett.*, **34**, 644 (1979).
18. N. Holonyak, et al., "Quantum Well Heterostructure Laser," *IEEE J. Quant. Electron.*, **QE-16**, 170 (1980).
19. T. P. Lee, "High Speed Photonic Devices," in S. M. Sze, Ed., *Modern Semiconductor Device Physics*, Wiley Interscience, New York, 1998.
20. K. Kasukawa, Y. Imajo, and T. Makino, "1.3 μm GaInAsP/InP Buried Heterostructure Graded Index Separate Confinement Multiple Quantums Well Lasers Epitaxially Grown by MOCVD," *Electron. Lett.*, **25**,104(1989).
21. S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., Wiley, New York, 1981, Ch. 12-14.
22. S. R. Forrest, "Photodiodes for Long-Wavelength Communication Systems," *Laser Focus*, **18**, 81 (1982).
23. D. M. Chapin, C. S. Fuller, and G. L. Pearson, "A New Silicon p-n Junction Photocell for Converting Solar Radiation into Electrical Power," *J. Appl. Phys.*, **25**, 676 (1954).
24. M. S. Green, "Solar Cells" in S. M. Sze, Ed., *Modern Semiconductor Device Physics*, Wiley Interscience, New York, 1998.
25. R. Hulstrom, R. Bird, and C. Riordan, "Spectral Solar Irradiance Data Sets for Selected Terrestrial Conditions," *Solar Cells*, **15**, 365 (1985).
26. C. H. Henry, "Limiting Efficiency of Ideal Single and Multiple Energy Gap Terrestrial Solar Cells," *J. Appl. Phys.*, **51**, 4494 (1980).

27. A. Luque, Ed, *Physical Limitation to Photovoltaic Energy Conversion*, IOP Press, Philadelphia, 1990.

28. M. A. Green, *Silicon Solar Cells: Advanced Principles and Practice*, Bridge Printery, Sydney, 1995.

29. J. Macneil, et. al., "Recent Improvements in Very Large Area α-*Si* PV Module Manufacturing," in *Proc., 10th Euro. Photovolt. Sol. Energy Conf.*, Lisbon, 1188, 1991.

30. R. I. Frank, J. L. Goodich, and R. Kaplow, "A Novel Silicon High-Intensity Photovoltaic Cell," *Conf. Rec. 14th IEEE Photovolt. Conf.*, IEEE, New York, p. 1350 (1980).

PROBLEMS (* DENOTES DIFFICULT PROBLEMS)

FOR SECTION 9.1 RADIATIVE TRANSITIONS AND OPTICAL ABSORPTION

*1. A GaAs sample is illuminated with a light having a wavelength of 0.6 μm. The incident power is 15 mW. If one-third of the incident power is reflected and another third exits from the other end of the sample, what is the thickness of the sample? Find the thermal energy dissipated per second to the lattice.

FOR SECTION 9.2 LIGHT EMITTING DIODE

2. The efficiency for electrical-to-optical conversion in a LED is given by $4\overline{n}_1\overline{n}_2(1 - \cos\theta_c)/(\overline{n}_1 + \overline{n}_2)^2$, where $\overline{n}_1$ and $\overline{n}_2$ are the refractive index of air and the semiconductor, respectively, and θ_c is the critical angle. Find the efficiency of an $Al_{0.3}Ga_{0.7}As$ LED operated at 0.898 μm.

FOR SECTION 9.3 SEMICONDUCTOR LASER

3. An InGaAsP Fabry–Perot laser operating at a wavelength of 1.33 μm has a cavity length of 300 μm. The index of refraction of InGaAsP is 3.39. (a) What is the mirror loss expressed in cm^{-1}? (b) If one of the laser facets is coated to produce 90% reflectivity, how much threshold current reduction (as a percentage) can be expected, assuming $\alpha = 10\ cm^{-1}$.

4. Calculate the confinement factor for a GaAs laser with an active region thickness 1 μm, refractive index 3.6, and critical angle at the active-to-nonactive boundary of 84°. Assume the C constant to be $8\times10^7\ m^{-1}$. Repeat the calculation for a GaAs/AlGaAs DH laser, where all the factors remain unchanged except for the critical angle, which is now 78°.

5. Derive Eq. 17 for the separation $\Delta\lambda$ between the allowed modes in the longitudinal direction. For a GaAs laser diode operated at $\lambda = 0.89$ μm, with $\overline{n}_1 = 3.58$, $L = 300$ μm, and $d\overline{n}_1/d\lambda = 2.5\ \mu m^{-1}$, find $\Delta\lambda$.

6. Calculate the gain coefficient for the two cases of Prob. 4 if the cavity length is 100 μm, the absorption coefficient is $10^4\ m^{-1}$, and the end mirrors are cleaved. How much shorter can the cavity be and still produce the same gain if one end-mirror is metallized to produce a reflectivity of 0.99?

7. Calculate the threshold current for the two cases of Prob. 4 if one end-mirror's reflectivity is 0.99. The cavity width is 5 μm and the gain factor for case 1 is $0.1\ cm^{-3}A^{-1}$.

*8. For a DFB laser with a cavity length of 300 μm, a material reflective index of 3.4 and an oscillating wavelength of 1.33 μm, find the Bragg wavelength and grating periodicity. The oscillating wavelength λ_0 is given by $\lambda_0 = \lambda_B \pm \dfrac{\left(m + \frac{1}{2}\right)\lambda_B^{\ 2}}{2\pi L}$, where m is an integer.

9. For high-temperature laser operation, it is important to have a low-temperature coefficient of the threshold current $\xi = (dI_{th}/dT)/dI_{th}$. What is the coefficient ξ for the laser shown in Fig. 26? If $T_0 = 50°C$, is this laser better or worse for high-temperature operation?

FOR SECTION 9.4 PHOTODETECTOR

10. A photoconductor with dimensions $L = 6$ mm, $W = 2$ mm, and $D = 1$ mm (Fig. 30) is placed under uniform radiation. The absorption of the light increases the current by 2.83 mA. A voltage of 10 V is applied across the device. As the radiation is suddenly cut off, the current falls, initially at a rate of 23.6 A/s. The electron and hole mobility are 3600 and 1700 cm^2/V-s, respectively. Find (a) the equilibrium density of electron-hole pairs generated under radiation, (b) the minority-carrier lifetime, and (c) the excess density of electrons and holes remaining 1 ms after the radiation is cut off.

11. Calculate the gain and current generated when 1 μW of optical power with $h\nu = 3$ eV is shone onto a photoconductor of $\eta = 0.85$ and a minority carrier lifetime of 0.6 ns. The material has an electron mobility of 3000 cm^2/V-s, the electric field is 5000 V/cm, and $L = 10$ μm.

*12. Show that the quantum efficiency η of a p-i-n photodetector is related to the responsivity $R(= I_p/P_{opt})$ at a wavelength λ (μm) by the equation $R = \dfrac{\eta\lambda}{1.24}$.

*13. A silicon n^+-p-π-p^- avalanche photodiode operated at 0.8 μm has a p-layer of 3 μm and a n-layer 9 μm thick. The biasing voltage must be high enough to cause avalanche breakdown in the p-region and velocity saturation in the π-region. Find the minimum required biasing voltage and the corresponding doping concentration of the p-region. Estimate the transit time of the device.

FOR SECTION 9.5 SOLAR CELL

14. A p-n junction photodiode can be operated under photovoltaic conditions similar to that of a solar cell. The current-voltage characteristics of a photodiode under illumination are also similar (Fig. 37). State three major differences between a photodiode and a solar cell.

*15. Consider a silicon p-n junction solar cell of area 2 cm^2. If the dopings of the solar cell are $N_A = 1.7\times 10^{16}$ cm^{-3} and $N_D = 5\times10^{19}$ cm^{-3}, and given $\tau_n = 10$ μs, $\tau_p = 0.5$ μs, $D_n = 9.3$ cm^2/s, $D_p = 2.5$ cm^2/s, and $I_L = 95$ mA, (a) calculate and plot the I-V characteristics of the solar cell, (b) calculate the open-circuit voltage, and (c) determine the maximum output power of the solar cell, all at room temperature.

*16. At 300 K, an ideal solar cell has a short circuit current of 3 A and an open-circuit voltage of 0.6 V. Calculate and sketch its power output as a function of operating voltage and find its fill factor from this power output.

17. For the solar cell shown in Fig. 40, find the relative maximum power output for a R_s of 0 and 5 Ω.

18. For the solar cell operated under solar-concentration conditions (Fig. 44 with the measured η), how many such solar cells operated under one-sun conditions are needed to produce the same power output as one cell operated under a 10-sun, 100-sun, or 1000-sun concentration?

CHAPTER 10

Crystal Growth and Epitaxy

- 10.1 SILICON CRYSTAL GROWTH FROM THE MELT
- 10.2 SILICON FLOAT-ZONE PROCESS
- 10.3 GaAs CRYSTAL-GROWTH TECHNIQUES
- 10.4 MATERIAL CHARACTERIZATION
- 10.5 EPITAXIAL-GROWTH TECHNIQUES
- 10.6 STRUCTURES AND DEFECTS IN EPITAXIAL LAYERS
- SUMMARY

As discussed in Chapter 2, the two most important semiconductors for discrete devices and integrated circuits are silicon and gallium arsenide. In this chapter we describe the common techniques for growing single crystals of these two semiconductors. The basic process flow from starting materials to polished wafers is shown in Fig. 1. The starting

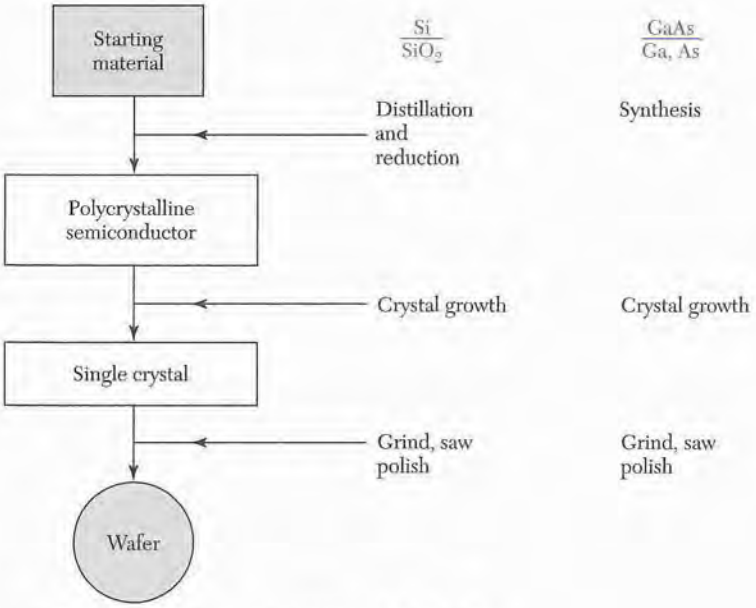


Fig. 1 Process flow from starting material to polished wafer

materials, silicon dioxide for a silicon wafer and gallium and arsenic for a gallium arsenide wafer, are chemically processed to form a high-purity polycrystalline semiconductor from which single crystals are grown. The single-crystal ingots are shaped to define the diameter of the material and sawed into wafers. These wafers are etched and polished to provide smooth, specular surfaces on which devices will be made.

A technology closely related to crystal growth involves the growth of single-crystal semiconductor layers on a single-crystal semiconductor substrate. This is called *epitaxy*, from the Greek words epi (meaning "on") and taxis (meaning "arrangement"). The epitaxial layer and the substrate materials may be the same, giving rise to *homoepitaxial*. For example, an n-type silicon can be grown epitaxially on an n^+-silicon substrate. On the other hand, if the epitaxial layer and the substrate are chemically and often crystallographically different, we have *heteroepitaxy*, such as the epitaxial growth of $Al_xGa_{1-x}As$ on GaAs.

Specifically, we cover the following topics:

- Basic techniques to grow silicon and GaAs single-crystal ingots.
- Wafer-shaping steps from ingots to polished wafers.
- Wafer characterization in term of its electrical and mechanical properties.
- Basic techniques of epitaxy, that is, to grow a single-crystal layer on a single-crystal substrate.
- Structures and defects of lattice-matched and strained-layer epitaxial growth.

10.1 SILICON CRYSTAL GROWTH FROM THE MELT

The basic technique for silicon crystal growth from the melt, which is material in liquid form, is the Czochralski technique. A substantial percentage (> 90%) of the silicon crystals for the semiconductor industry is prepared by the Czochralski technique, and virtually all the silicon used for fabricating integrated circuits is prepared by this technique.

10.1.1 Starting Material

The starting material for silicon is a relatively pure form of sand (SiO_2) called quartzite. This is placed in a furnace with various forms of carbon (coal, coke, and wood chips). Although a number of reactions take place in the furnace, the overall reaction is

$$\text{SiC (solid)} + \text{SiO}_2\text{ (solid)} \rightarrow \text{Si (solid)} + \text{SiO (gas)} + \text{CO (gas)}. \tag{1}$$

This process produces metallurgical-grade silicon with a purity of about 98%. Next, the silicon is pulverized and treated with hydrogen chloride (HCl) to form trichlorosilane ($SiHCl_3$):

$$\text{Si (solid)} + 3\text{HCl (gas)} \xrightarrow{300^\circ\text{C}} \text{SiHCl}_3\text{ (gas)} + \text{H}_2\text{ (gas)}. \tag{2}$$

The trichlorosilane is a liquid at room temperature (boiling point 32°C). Fractional distillation of the liquid removes the unwanted impurities. The purified $SiHCl_3$ is then used in a hydrogen reduction reaction to prepare the electronic-grade silicon (EGS):

$$\text{SiHCl}_3\text{ (gas)} + \text{H}_2\text{ (gas)} \rightarrow \text{Si (solid)} + 3\text{HCl (gas)}. \tag{3}$$

This reaction takes place in a reactor containing a resistance-heated silicon rod, which serves as the nucleation point for the deposition of silicon. The EGS, a polycrystalline material of high purity, is the raw material used to prepare device-quality, single-crystal silicon. Pure EGS generally has impurity concentrations in the parts-per-billion range.[1]

10.1.2 The Czochralski Technique

The Czochralski technique uses an apparatus called a crystal puller, shown in Chapter 2. A simplified version is shown in Fig. 2. The puller has three main components: (a) a furnace, which includes a fused-silicon (SiO_2) crucible, a graphite susceptor, a rotation mechanism (clockwise as shown), a heating element, and a power supply; (b) a crystal-pulling mechanism, which includes a seed holder and a rotation mechanism (counter–clockwise); and (c) an ambient control, which includes a gas source (such as argon), a flow control, and a exhaust system. In addition, the puller has an overall microprocessor-based control system to control process parameters such as temperature, crystal diameter, pull rate, and rotation speeds, as well as to permit programmed process steps. Also, various sensors and feedback loops allow the control system to respond automatically, reducing operator intervention.

In the crystal-growing process, polycrystalline silicon (EGS) is placed in the crucible and the furnace is heated above the melting temperature of silicon. A suitably oriented seed crystal (e.g., <111>) is suspended over the crucible in a seed holder. The seed is inserted into the melt. Part of it melts, but the tip of the remaining seed crystal still touches the liquid surface. It is then slowly withdrawn. Progressive freezing at the solid-liquid

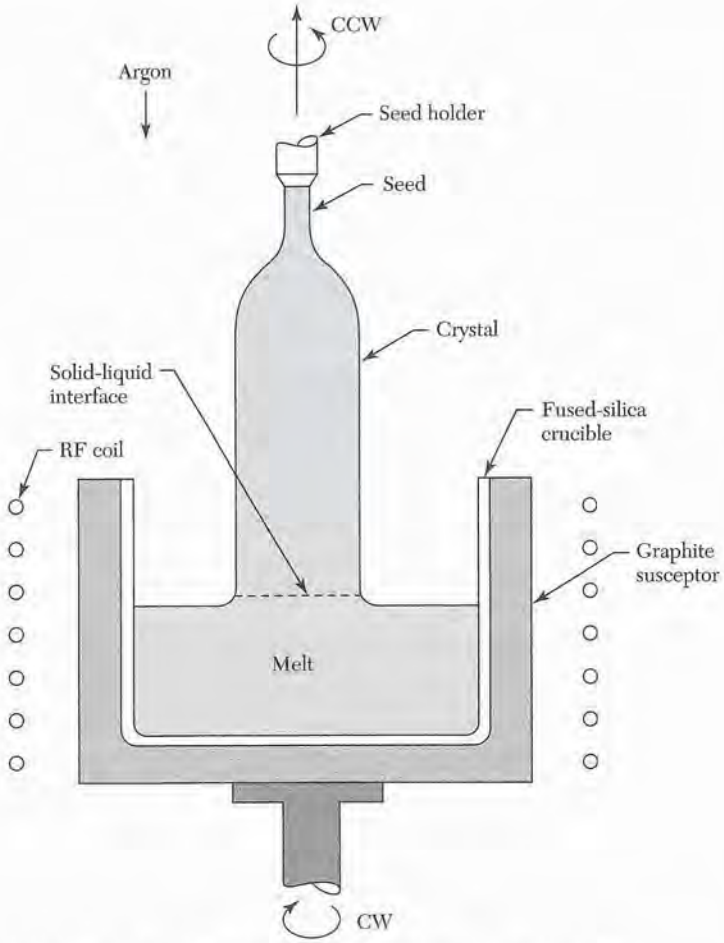


Fig. 2 Czochralski crystal puller. CW, clockwise; CCW, counter clockwise.

interface yields a large, single crystal. A typical pull rate is a few millimeters per minute. For large-diameter silicon ingots, an external magnetic field is applied to the basic Czochialski puller. The purpose of the external magnetic field is to control the concentration of defects, impurities, and oxygen contents.[2] Figure 3 shows a 300 mm (12 in.) and a 400 mm (16 in.) Czochralski grown silicon ingots.

10.1.3 Distribution of Dopant

In crystal growth, a known amount of dopant is added to the melt to obtain the desired doping concentration in the grown crystal. For silicon, boron and phosphorus are the most common dopants for p- and n-type materials, respectively.

As a crystal is pulled from the melt, the doping concentration incorporated into the crystal (solid) is usually different from the doping concentration of the melt (liquid) at the interface. The ratio of these two concentrations is defined as the *equilibrium segregation coefficient* k_0:

$$\boxed{k_0 \equiv \frac{C_s}{C_l},} \tag{4}$$

where C_s and C_l are, respectively, the equilibrium concentrations of the dopant in the solid and liquid near the interface. Table 1 lists values of k_0 for the commonly used dopants for silicon. Note that most values are below 1, which means that during growth the dopants

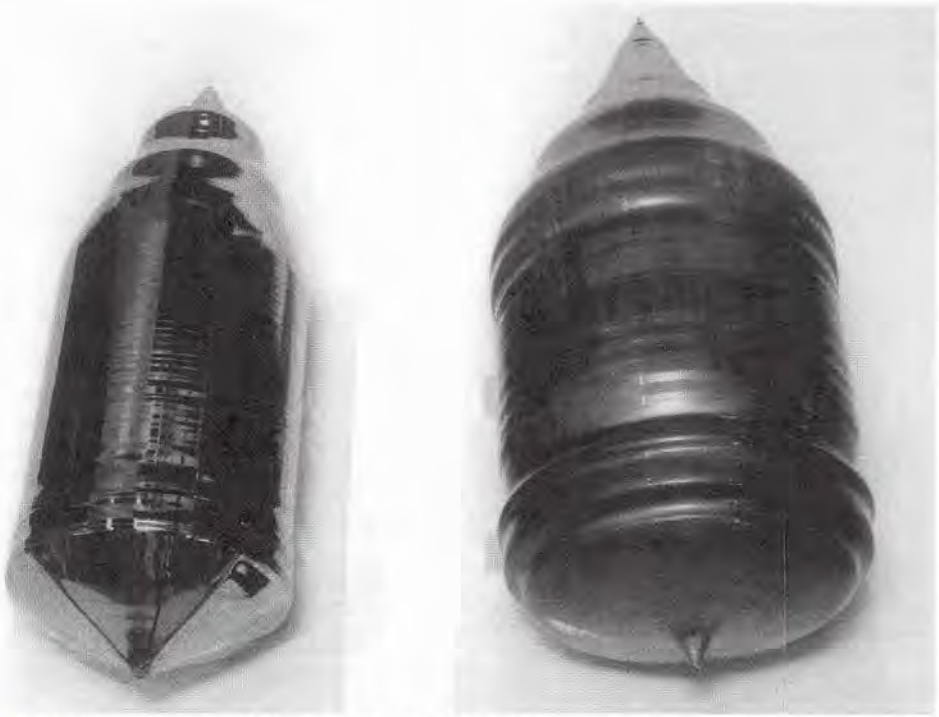


Fig. 3 300 mm (12 in.) and 400 mm (16 in.) Czochralski-grown silicon ingots. (Photo courtesy of Shin-Etsu Handotai Co., Tokyo.)

TABLE 1 Equilibrium Segregation Coefficients for Dopants in Si

Dopant	k_0	Type	Dopant	k_0	Type
B	8×10^{-1}	p	As	3.0×10^{-1}	n
Al	2×10^{-3}	p	Sb	2.3×10^{-2}	n
Ga	8×10^{-3}	p	Te	2.0×10^{-4}	n
In	4×10^{-4}	p	Li	1.0×10^{-2}	n
O	1.25	n	Cu	4.0×10^{-4}	—[a]
C	7×10^{-2}	n	Au	2.5×10^{-5}	—[a]
P	0.35	n			

[a]Deep-lying impurity level.

are rejected into the melt. Consequently, the melt becomes progressively enriched with the dopant as the crystal grows.

Consider a crystal being grown from a melt having an initial weight M_0 with an initial doping concentration C_0 in the melt (i.e., the weight of the dopant per 1 g of melt). At a given point of growth when a crystal of weight M has been grown, the amount of dopant remaining in the melt (by weight) is S. For an incremental amount of the crystal with weight dM, the corresponding reduction of the dopant ($-dS$) from the melt is $C_s\, dM$, where C_s is the doping concentration in the crystal (by weight):

$$-dS = C_s\, dM. \tag{5}$$

Now, the remaining weight of the melt is $M_0 - M$, and the doping concentration in the liquid (by weight), C_l, is given by

$$C_l = \frac{S}{M_0 - M}. \tag{6}$$

Combining Eqs. 5 and 6 and substituting $C_s/C_l = k_0$ yields

$$\frac{dS}{S} = -k_0\left(\frac{dM}{M_0 - M}\right). \tag{7}$$

Given the initial weight of the dopant, C_0M_0, we can integrate Eq. 7 :

$$\int_{C_0M_0}^{S} \frac{dS}{S} = k_0\int_0^M \frac{-dM}{M_0 - M}. \tag{8}$$

Solving Eq. 8 and combining with Eq. 6 gives

$$\boxed{C_s = k_0C_0\left(1 - \frac{M}{M_0}\right)^{k_0-1}}. \tag{9}$$

Figure 4 illustrates the doping distribution as a function of the fraction solidified (M/M_0) for several segregation coefficients.[3,4] As crystal growth progresses, the composition initially at k_0C_0 will increase continually for $k_0 < 1$ and decrease continually for $k_0 > 1$. When $k_0 \cong 1$, a uniform impurity distribution can be obtained.

EXAMPLE 1

A silicon ingot, which should contain 10^{16} boron atoms/cm^3, is to be grown by the Czochralski technique. What concentration of boron atoms should be in the melt to give the required concentra-

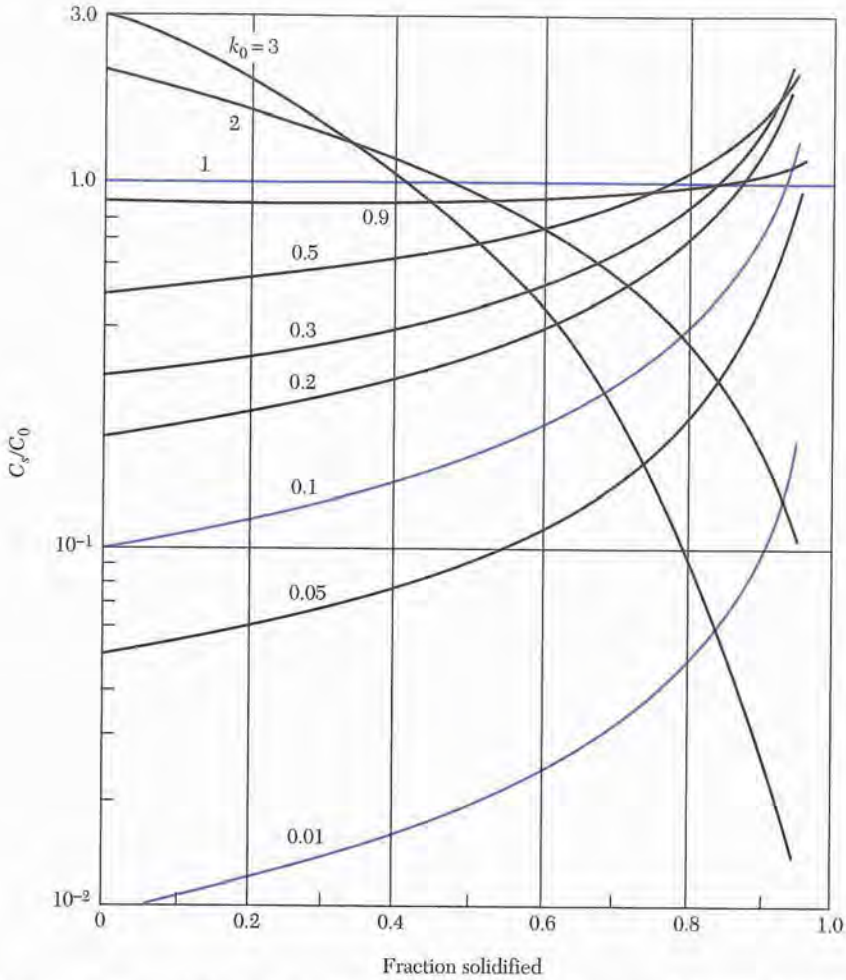


Fig. 4 Curves for growth from the melt showing the doping concentration in a solid as a function of the fraction solidified.[4]

tion in the ingot? If the initial load of silicon in the crucible is 60 kg, how many grams of boron (atomic weight 10.8) should be added? The density of molten silicon is 2.53 g/cm^3.

SOLUTION Table 1 shows that the segregation coefficient k_0 for boron is 0.8. We assume that $C_s = k_0 C_l$ throughout the growth. Thus, the initial concentration of boron in the melt should be

$$\frac{10^{16}}{0.8} = 1.25 \times 10^{16} \text{ boron atoms / cm}^3.$$

Since the amount of boron concentration is so small, the volume of melt can be calculated from the weight of silicon. Therefore, the volume of 60 kg of silicon is

$$\frac{60 \times 10^3}{2.53} = 2.37 \times 10^4 \text{ cm}^3.$$

The total number of boron atoms in the melt is

1.25×10^{16} atoms/cm^3 $\times$ 2.37×10^4 cm^3 = 2.96×10^{20} boron atoms,

so that

$$\frac{2.96 \times 10^{20} \text{ atoms} \times 10.8 \text{ g / mol}}{6.02 \times 10^{23} \text{ atoms / mol}} = 5.31 \times 10^{-3} \text{ g of boron}$$
$$= 5.31 \text{ mg of boron.}$$

Note the small amount of boron needed to dope such a large load of silicon.

10.1.4 Effective Segregation Coefficient

While the crystal is growing, dopants are constantly being rejected into the melt (for $k_0 < 1$). If the rejection rate is higher than the rate of which the dopant can be transported away by diffusion or stirring, then a concentration gradient will develop at the interface, as illustrated in Fig. 5. The segregation coefficient (given in Section 10.1.3) is $k_0 = C_s / C_l(0)$. We can define an effective segregation coefficient k_e, which is the ratio of C_s and the impurity concentration far away from the interface:

$$k_e \equiv \frac{C_s}{C_l}. \tag{10}$$

Consider a small, virtually stagnant layer of melt with width δ in which the only flow is that required to replace the crystal being withdrawn from the melt. Outside this stagnant layer, the doping concentration has a constant value C_l. Inside the layer, the doping concentration can be described by the continuity equation (Eq. 58) derived in Chapter 3. At steady state, the only significant terms are the second and third terms on the right-hand side (we replace n_p by C and $\mu_n \mathscr{E}$ by v):

$$0 = v\frac{dC}{dx} + D\frac{d^2C}{dx^2}, \tag{11}$$

where D is the dopant diffusion coefficient in the melt, v is the crystal growth velocity, and C is the doping concentration in the melt.

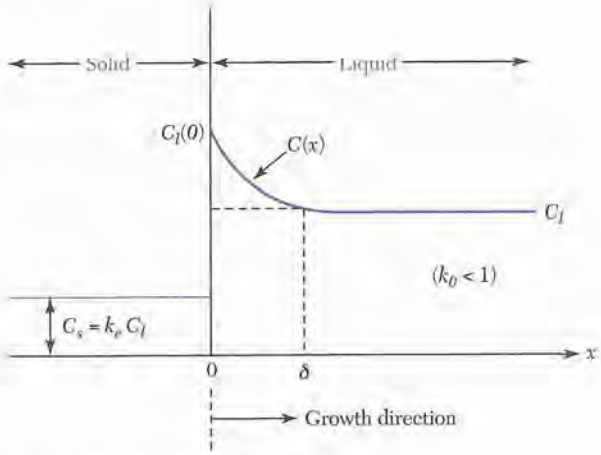


Fig. 5 Doping distribution near the solid-melt interface.

The solution of Eq. 11 is

$$C = A_1 e^{-vx/D} + A_2 \tag{12}$$

where A_1 and A_2 are constants to be determined by the boundary conditions. The first boundary condition is that $C = C_l(0)$ at $x = 0$. The second boundary condition is the conservation of the total number of dopants; that is, the sum of the dopant fluxes at the interface must be zero. By considering the diffusion of dopant atoms in the melt (neglecting diffusion in the solid), we have

$$D\left(\frac{dC}{dx}\right)_{x=0} + \left[C_l(0) - C_s\right] v = 0. \tag{13}$$

Substituting these boundary conditions into Eq. 12 and noting that $C = C_l$ at $x = \delta$ gives

$$e^{-v\delta/D} = \frac{C_l - C_s}{C_l(0) - C_s}. \tag{14}$$

Therefore,

$$\boxed{k_e \equiv \frac{C_s}{C_l} = \frac{k_0}{k_0 + (1 - k_0)e^{-v\delta/D}}.} \tag{15}$$

The doping distribution in the crystal is given by the same expression as in Eq. 9, except that k_0 is replaced by k_e. Values of k_e are larger than those of k_0 and can approach 1 for large values of the growth parameter $v\delta/D$. Uniform doping distribution ($k_e \to 1$) in the crystal can be obtained by employing a high pull rate and a low rotation speed (since δ is inversely proportional to the rotation speed). Another approach to achieve uniform doping is to add ultrapure polycrystalline silicon continuously to the melt so that the initial doping concentration is maintained.

10.2 SILICON FLOAT–ZONE PROCESS

The float–zone process can be used to grow silicon that has lower contaminations than that normally obtained from the Czochralski technique. A schematic setup of the float zone process is shown in Fig. 6*a*. A high-purity polycrystalline rod with a seed crystal at the bottom is held in a vertical position and rotated. The rod is enclosed in a quartz envelope within which an inert atmosphere (argon) is maintained. During the operation, a small zone (a few centimeters in length) of the crystal is kept molten by a radio-frequency heater, which is moved from the seed upward so that this *floating zone* traverses the length of the rod. The molten silicon is retained by surface tension between the melting and growing solid-silicon faces. As the floating zone moves upward, a single-crystal silicon freezes at the zone's retreating end and grows as an extension of the seed crystal. Materials with higher resistivities can be obtained from the float-zone process than from the Czochralski process because it can be used to purify the crystal more easily. Furthermore, since no crucible is used in the float-zone process, there is no contamination from the crucible (as with Czochralski growth). At the present time, float-zone crystals are used mainly for high-power, high-voltage devices, where high-resistivity materials are required.

To evaluate the doping distribution of a float-zone process, consider a simplified model, as shown in Fig. 6*b*. The initial, uniform doping concentration in the rod is C_0 (by weight).

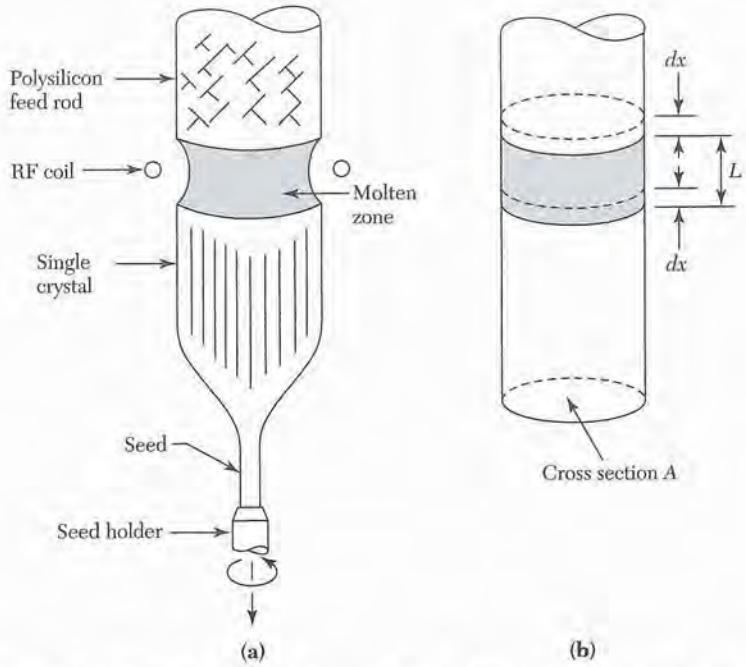


Fig. 6 Float-zone process. (*a*) Schematic setup. (*b*) Simple model for doping evaluation.

L is the length of the molten zone at a distance x along the rod, A the cross-sectional area of the rod, ρ_d the specific density of silicon, and S the amount of dopant present in the molten zone. As the zone traverses a distance dx, the amount of dopant added to it at its advancing end is $C_0\rho_d A\,dx$, whereas the amount of dopant removed from it at the retreating end is $k_e(S\,dx/L)$, where k_e is the effective segregation coefficient. Thus,

$$dS = C_0\rho_d A dx - \frac{k_e S}{L}dx = \left(C_0\rho_d A - \frac{k_e S}{L}\right)dx, \tag{16}$$

so that

$$\int_0^x dx = \int_{S_0}^{S} \frac{dS}{C_0\rho_d A - (k_e S / L)}, \tag{16a}$$

where $S_0 = C_0\rho_d AL$ is the amount of dopant in the zone when it was first formed at the front end of the rod. From Eq. 16*a* we obtain

$$\exp\left(\frac{k_e x}{L}\right) = \frac{C_0\rho_d A - (k_e S_0 / L)}{C_0\rho_d A - (k_e S / L)} \tag{17}$$

or

$$S = \frac{C_0 A\rho_d L}{k_e}\,[1-(1-k_e)^{-k_e x/L}]. \tag{17a}$$

Since C_s (the doping concentration in the crystal at the retreating end) is given by $C_s = k_e(S/A\rho_d L)$, then

$$\boxed{C_s = C_0[1-(1-k_e)^{-k_e x/L}].} \tag{18}$$

Figure 7 shows the doping concentration versus the solidified zone length for various values of k_e.

These two crystal growth techniques can also be used to remove impurities. A comparison of Fig. 7 with Fig. 4 shows that a single pass in the float-zone process does not produce as much purification as a single Czochralski growth. For example, for $k_0 = k_e = 0.1$, C_s/C_0 is smaller over most of the solidified ingot made by the Czochralski growth. However, multiple float-zone passes can be performed on a rod much more easily than a crystal can be grown, the end region cropped off, and regrown from the melt. Figure 8 shows the impurity distribution for an element with $k_e = 0.1$ after a number of successive passes of the zone along the length of the rod.[4] Note that there is a substantial reduction of impurity concentration in the rod after each pass. Therefore, the float-zone process is ideally suited for crystal purification. This process is also called the zone-refining technique, which can provide a very-high-purity level of the raw material.

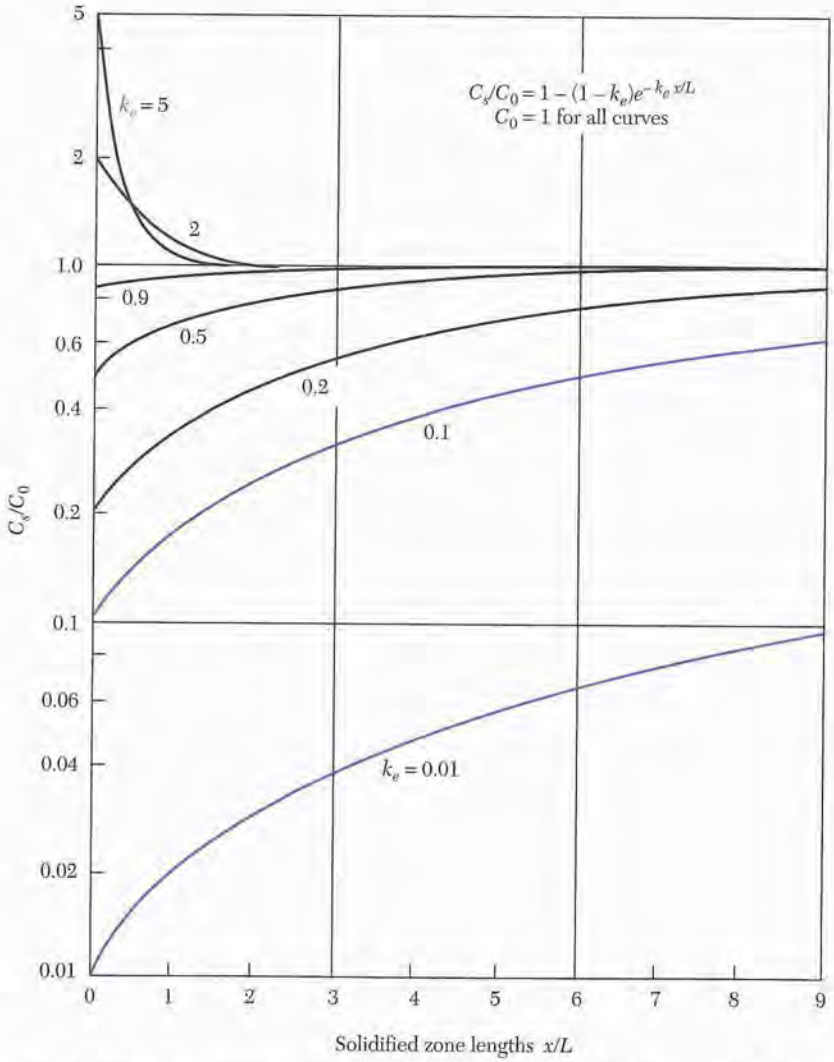


Fig. 7 Curves for the float-zone process showing doping concentration in the solid as a function of solidified zone lengths.[4]

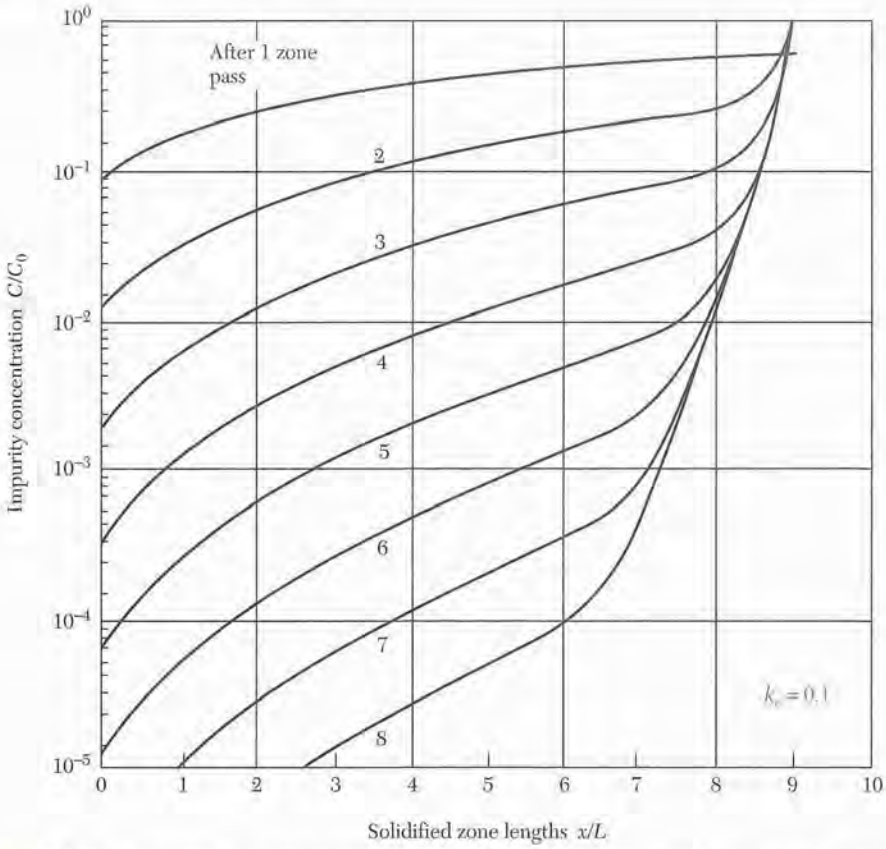


Fig. 8 Relative impurity concentration versus zone length for a number of passes. L denotes the zone length.[4]

If it is desirable to dope the rod rather than purify it, consider the case in which all the dopants are introduced in the first zone ($S_0 = C_1 A\rho_d L$) and the initial concentration C_0 is negligibly small. Equation 17 gives

$$S_0 = S \exp\left(\frac{k_e x}{L}\right). \tag{19}$$

Since $C_s = k_e(S/A\rho_d L)$, we obtained from Eq. 19

$$C_s = k_e C_1 e^{-k_e x/L}, \tag{20}$$

Therefore, if $k_e x/L$ is small, C_s will remain nearly constant with distance except at the end that is last to solidify.

For certain switching devices, such as high-voltage thyristors, discussed in Chapter 4, large chip areas are used, frequently an entire wafer for a single device. This size imposes stringent requirements on the uniformity of the starting material. To obtain homogeneous distribution of dopants, we use a float-zone silicon slice that has an average doping concentration well below the required amount. The slice is then irradiated with thermal neutrons. This process, called *neutron irradiation*, gives rise to fractional transmutation of silicon into phosphorus and dopes the silicon n-type:

$$\mathrm{Si}_{14}^{30} + \text{neutron} \longrightarrow \mathrm{Si}_{14}^{31} + \gamma\,\text{ray} \xrightarrow{2.62\ \text{hr}} \mathrm{P}_{15}^{31} + \beta\,\text{ray}. \tag{21}$$

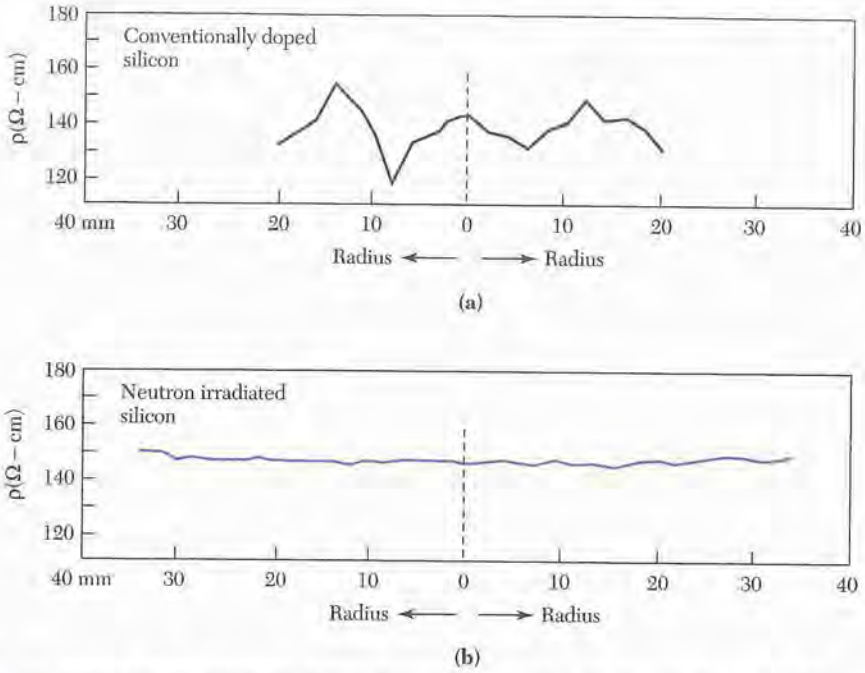


Fig. 9 (*a*) Typical lateral resistivity distribution in a conventionally doped silicon. (*b*) Silicon doped by neutron irradiation.[5]

The half-life of the intermediate element Si_{14}^{31} is 2.62 hours. Because the penetration depth of neutrons in silicon is about 100 cm, doping is very uniform throughout the slice. Figure 9 compares the lateral resistivity distributions in conventionally doped silicon and in silicon doped by neutron irradiation.[5] Note that the resistivity variations for the neutron-irradiated silicon are much smaller than that for the conventionally doped silicon.

10.3 GaAs CRYSTAL-GROWTH TECHNIQUES

10.3.1 Starting Materials

The starting materials for the synthesis of polycrystalline gallium arsenide are the elemental, chemically pure gallium and arsenic. Because gallium arsenide is a combination of two materials, its behavior is different from that of a single material such as silicon. The behavior of a combination can be described by a *phase diagram*. A phase is a state (e.g., solid, liquid, or gaseous) in which a material may exist. A phase diagram shows the relationship between the two components, gallium and arsenic, as a function of temperature.

Figure 10 shows the phase diagram of the gallium-arsenic system. The abscissa represents various compositions of the two components in terms of atomic percent (lower scale) or weight percent (upper scale).[6,7] Consider a melt that is initially of composition x (e.g., 85 atomic percent arsenic shown in Fig. 10). When the temperature is lowered, its composition will remain fixed until the *liquidus* line is reached. At the point (T_1, x), material of 50 atomic percent arsenic (i.e., gallium arsenide) will begin to solidify.

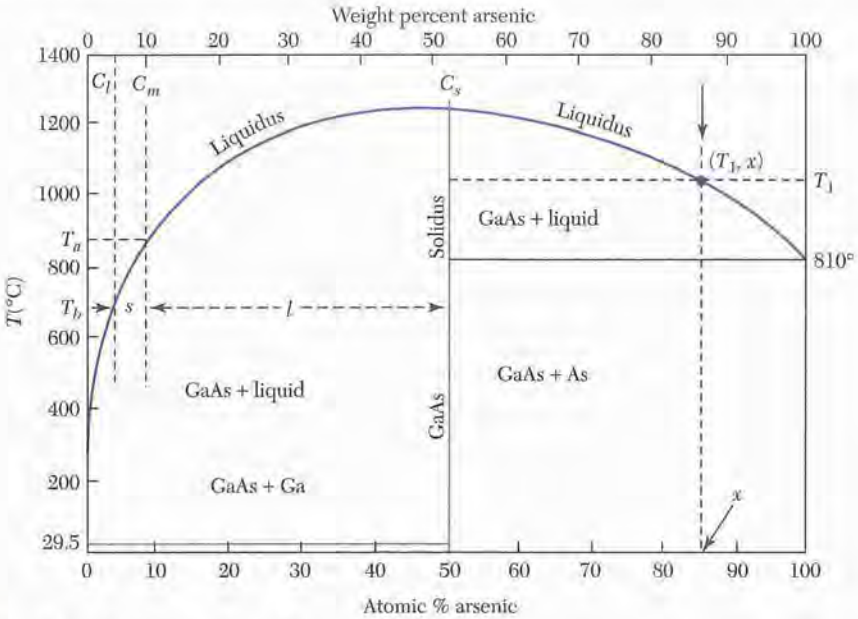


Fig. 10 Phase diagram for the gallium-arsenic system.[6]

EXAMPLE 2

In Fig.10, consider a melt of initial composition C_m (weight percent scale) that is cooled from T_a (on the liquidus line) to T_b. Find the fraction of the melt that will be solidified.

SOLUTION At T_b, M_l is the weight of the liquid, M_s the weight of the solid (i.e., GaAs), and C_l and C_s the concentrations of dopant in the liquid and the solid, respectively. Therefore, the weight of arsenic in the liquid and solid are M_lC_l and M_sC_s, respectively. Because the total arsenic weight is $(M_l + M_s)C_m$, we have

$$M_lC_l + M_sC_s = (M_l + M_s)C_m$$

or

$$\frac{M_s}{M_l} = \frac{\text{weight of GaAs at } T_b}{\text{weight of liquid at } T_b} = \frac{C_m - C_l}{C_s - C_m} = \frac{s}{l},$$

where s and l are the lengths of the two lines measured from C_m to the liquidus and solidus line, respectively. As can be seen from Fig. 10, about 10% of the melt is solidified.

Unlike silicon, which has a relatively low vapor pressure at its melting point (~10^{-6} atm at 1412°C), arsenic has much higher vapor pressures at the melting point of gallium arsenide (1240°C). In its vapor phase, arsenic has As_2 and As_4 as its major species. Figure 11 shows the vapor pressures of gallium and arsenic along the liquidus curve.[8] Also shown for comparison is the vapor pressure of silicon. The vapor pressure curves for gallium arsenide are double valued. The dashed curves are for arsenic-rich gallium arsenide melt (right side of liquidus line in Fig. 10), and the solid curves are for gallium-rich gallium arsenide melt (left side of liquidus line in Fig. 10). Because there is a larger amount of arsenic in an arsenic-rich melt than in a gallium-rich melt, more arsenic (As_2 and As_4) will be vaporized from the arsenic-rich melt, thus resulting in a higher vapor pressure. A similar argument can explain the higher vapor pressure of gallium in a gallium-rich melt. Note that long before the melting point is reached, the surface layers of liquid gal-

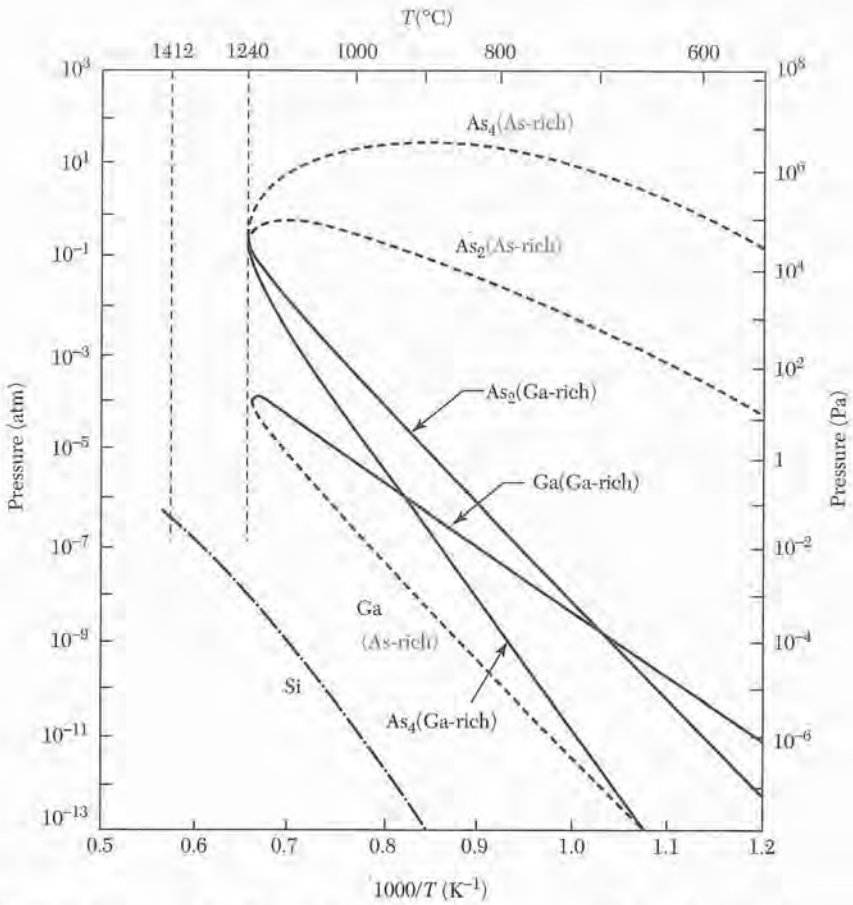


Fig. 11 Partial pressure of gallium and arsenic over gallium arsenide as a function of temperature.[8] Also shown is the partial pressure of silicon.

lium arsenide may decompose into gallium and arsenic. Since the vapor pressure of gallium and arsenic is different, there is a preferential loss of the more volatile arsenic species, and the liquid becomes gallium rich.

To synthesize gallium arsenide, an evacuated, sealed quartz tube system with a two-temperature furnace is commonly used. The high-purity arsenic is placed in a graphite boat and heated to 610°–620°C, whereas the high-purity gallium is placed in another graphite boat and heated to slightly above the gallium arsenide melting temperature (1240°–1260°C). Under these conditions, an overpressure of arsenic is established (a) to cause the transport of arsenic vapor to the gallium melt, converting it into gallium arsenide, and (b) to prevent decomposition of the gallium arsenide while it is being formed in the furnace. When the melt cools, a high-purity polycrystalline gallium arsenide results. This serves as the raw material to grow single-crystal gallium arsenide. [7]

10.3.2 Crystal-Growth Techniques

There are two techniques for GaAs crystal growth: the Czochralski technique and the Bridgman technique. Most gallium arsenide is grown by the Bridgman technique. However, the Czochralski technique is more popular for the growth of larger-diameter GaAs ingots.

For Czochralski growth of gallium arsenide, the basic puller is identical to that for silicon. However, to prevent decomposition of the melt during crystal growth, a liquid encapsulation method is employed. The liquid encapsulant is a molten boron trioxide (B_2O_3) layer about 1 cm thick. Molten boron trioxide is inert to the gallium arsenide surface and serves as a cap to cover the melt. This cap prevents decomposition of the gallium arsenide as long as the pressure on its surface is higher than 1 atm (760 Torr). Because boron trioxide can dissolve silicon dioxide, the fused-silica crucible is replaced with a graphite crucible.

To obtain the desired doping concentration in the grown crystal of GaAs, cadmium and zinc are commonly used for *p*-type materials, whereas selenium, silicon, and tellurium are used for *n*-type materials. For semiinsulating GaAs, the material is undoped. The equilibrium segregation coefficients for dopants in GaAs are listed in Table 2. Similar to those in Si, most of the segregation coefficients are less than 1. The expressions derived previously for Si are equally applicable to GaAs (Eqs. 4 to 15).

Figure 12 shows a Bridgman system in which a two-zone furnace is used for growing single-crystal gallium arsenide. The left-hand zone is held at a temperature (~610°C) to maintain the required overpressure of arsenic, whereas the right-hand zone is held just above the melting point of gallium arsenide (1240°C). The sealed tube is made of quartz and the boat is made of graphite. In operation, the boat is loaded with a charge of polycrystalline gallium arsenide, with the arsenic kept at the other end of the tube.

As the furnace is moved toward the right, the melt cools at one end. Usually, there is a seed placed at the left end of the boat to establish a specific crystal orientation. The gradual freezing (solidification) of the melt allows a single crystal to propagate at the liquid-solid interface. Eventually, a single crystal of gallium arsenide is grown. The impurity distribution can be described essentially by Eqs. 9 and 15, where the growth rate is given by the traversing speed of the furnace.

TABLE 2 Equilibrium Segregation Coefficients for Dopants in GaAs

Dopant	k_0	Type
Be	3	*p*
Mg	0.1	*p*
Zn	4×10^{-1}	*p*
C	0.8	*n/p*
Si	1.85×10^{-1}	*n/p*
Ge	2.8×10^{-2}	*n/p*
S	0.5	*n*
Se	5.0×10^{-1}	*n*
Sn	5.2×10^{-2}	*n*
Te	6.8×10^{-2}	*n*
Cr	1.03×10^{-4}	Semiinsulating
Fe	1.0×10^{-3}	Semiinsulating

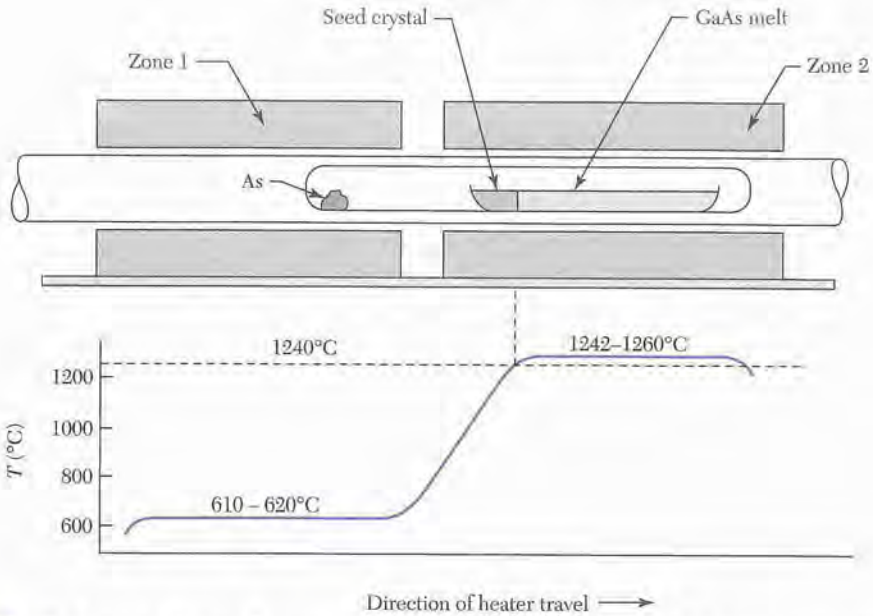


Fig. 12 Bridgman technique for growing single-crystal gallium arsenide and a temperature profile of the furnace.

10.4 MATERIAL CHARACTERIZATION

10.4.1 Wafer Shaping

After a crystal is grown, the first shaping operation is to remove the seed and the other end of the ingot, which is last to solidify.[1] The next operation is to grind the surface so that the diameter of the material is defined. After that, one or more flat regions are ground along the length of the ingot. These regions, or *flats*, mark the specific crystal orientation of the ingot and the conductivity type of the material. The largest flat, the *primary flat*, allows a mechanical locator in automatic processing equipment to position the wafer and to orient the devices relative to the crystal. Other smaller flats, called *secondary flats*, are ground to identify the orientation and conductivity type of the crystal, as shown in Fig. 13. For crystals with diameters equal or larger than 200 mm, no flats are ground. Instead, a small groove is ground along the length of the ingot.

The ingot is ready to be sliced by diamond saw into wafers. Slicing determines four wafer parameters: *surface orientation* (e.g., <111> or <100>), *thickness* (e.g., 0.5–0.7 mm, depending on wafer diameter); *taper*, which is the wafer thickness variations from one end to another; and *bow*, which is the surface curvature of the wafer, measured from the center of the wafer to its edge.

After slicing, both sides of the wafer are lapped using a mixture of Al_2O_3 and glycerine to produce a typical flatness uniformity within 2 μm. The lapping operation usually leaves the surface and edges of the wafer damaged and contaminated. The damaged and contaminated regions can be removed by chemical etching (see Chapter 12). The final step of wafer shaping is polishing. Its purpose is to provide a smooth, specular surface where device features can be defined by lithographic processes (see Chapter 12). Figure 14 shows 200 mm (8 in.) and 400 mm (16 in.) polished silicon wafers in cassettes.

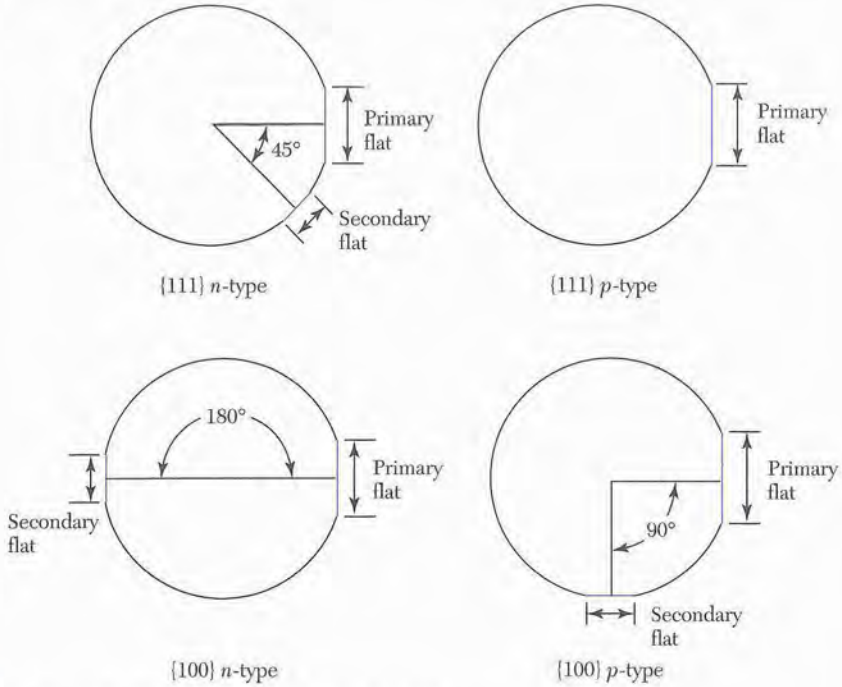


Fig. 13 Identifying flats on a semiconductor wafer.

Fig. 14 200 mm (8 in.) and 400 mm (16 in.) polished silicon wafers in cassettes. (Photo courtesy of Shin-Etsu Handotai Co., Tokyo.)

Table 3 shows the specifications for 125, 150, 200, and 300 mm diameter polished silicon wafers from the Semiconductor Equipment and Materials Institute (SEMI). As mentioned previously, for large crystals (≥ 200 mm diameter) no flats are ground; instead, a groove is made on the edge of the wafer for positioning and orientation purpose.

Gallium arsenide is a more fragile material than silicon. Although the basic shaping operation of gallium arsenide is essentially the same as that for silicon, greater care must be exercised in gallium arsenide wafer preparation. The state of gallium arsenide tech-

TABLE 3 Specification for Polished Monocyrstalline Silicon Wafers

Parameter	125 mm	150 mm	200 mm	300 mm
Diameter (mm)	125±1	150±1	200±1	300±1
Thickness (mm)	0.6–0.65	0.65–0.7	0.715–0.735	0.755–0.775
Primary flat length (mm)	40–45	55–60	NA[a]	NA
Secondary flat length (mm)	25–30	35–40	NA	NA
Bow (μm)	70	60	30	< 30
Total thickness variation (μm)	65	50	10	< 10
Surface orientation	(100) ± 1°	Same	Same	Same
	(111) ± 1°	Same	Same	Same

[a]NA: not available.

nology is relatively primitive compared with that of silicon. However, the technology of group III-V compounds has advanced partly because of the advances in silicon technology.

10.4.2 Crystal Characterization

Crystal Defects

A real crystal (such as a silicon wafer) differs from the ideal crystal in important ways. It is finite; thus, surface atoms are incompletely bonded. Furthermore, it has defects, which strongly influence the electrical, mechanical, and optical properties of the semiconductor. There are four categories of defects: point defects, line defects, area defects, and volume defects.

Figure 15 shows several forms of *point defects*.[1,9] Any foreign atom incorporated into the lattice at either a substitutional site [i.e., at a regular lattice site (Fig. 15*a*)] or interstitial site [i.e., between regular lattice sites (Fig.15*b*)] is a point defect. A missing atom in the lattice creates a vacancy, also considered a point defect (Fig. 15*c*). A host atom that is situated between regular lattice sites and adjacent to a vacancy is called a *Frenkel defect* (Fig. 15*d*). Point defects are particularly important subjects in the kinetics of diffusion and oxidation processes. These topics are considered in Chapters 11 and 13.

The next class of defects is the *line defect*, also called a dislocation.[10] There are two types of dislocations: the edge and screw types. Figure 16*a* is a schematic representation of an edge dislocation in a cubic lattice. There is an extra plane of atoms *AB* inserted into the lattice. The line of the dislocation would be perpendicular to the plane of the page. The screw dislocation may be considered as being produced by cutting the crystal partway through and pushing the upper part one lattice spacing over, as show in Fig. 16*b*. Line defects in devices are undesirable because they act as precipitation sites for metallic impurities, which may degrade device performance.

Area defects represent a large area discontinuity in the lattice. Typical defects are twins and grain boundaries. Twinning represents a change in the crystal orientation across a plane. A grain boundary is a transition between crystals having no particular orientational relationship to one another. Such defects appear during crystal growth. Another area defect is the stacking fault.[9] In this defect, the stacking sequence of atomic layer is interrupted. In Fig. 17 the sequence of atoms in a stack is *ABCABC* . . . . When a part of layer *C* is missing, this is called the intrinsic stacking fault, Fig. 17*a*. If an extra plane *A* is inserted between layers *B* and *C*, this is an extrinsic stacking fault (Fig. 17*b*). Such defects may appear during crystal growth. Crystals having these area defects are not usable for integrated-circuit manufacture and are discarded.

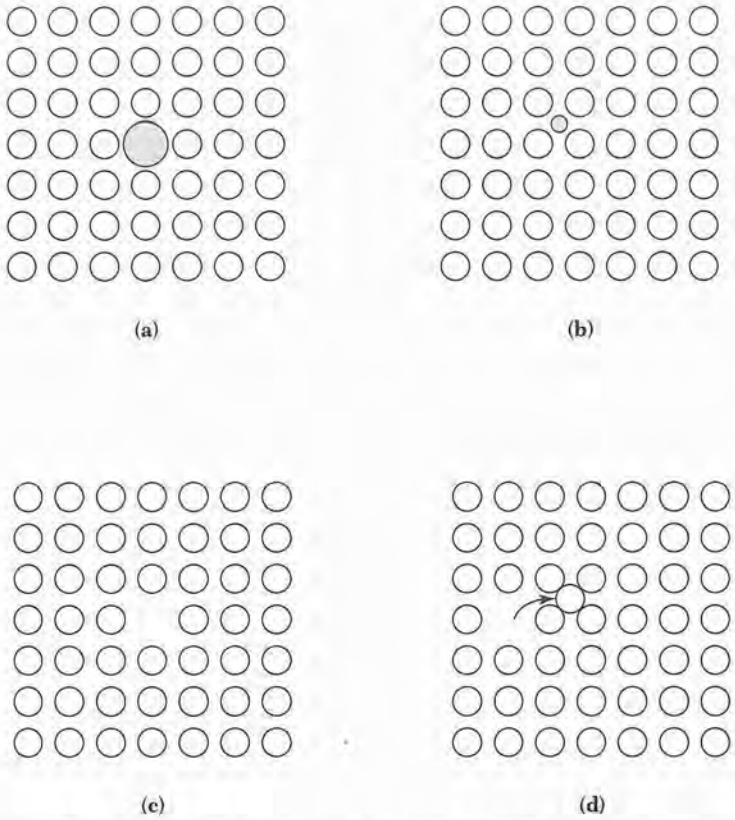


Fig. 15 Point defects. (*a*) Substitutional impurity. (*b*) Interstitial impurity. (*c*) Lattice vacancy. (*d*) Frenkel-type defect.[9]

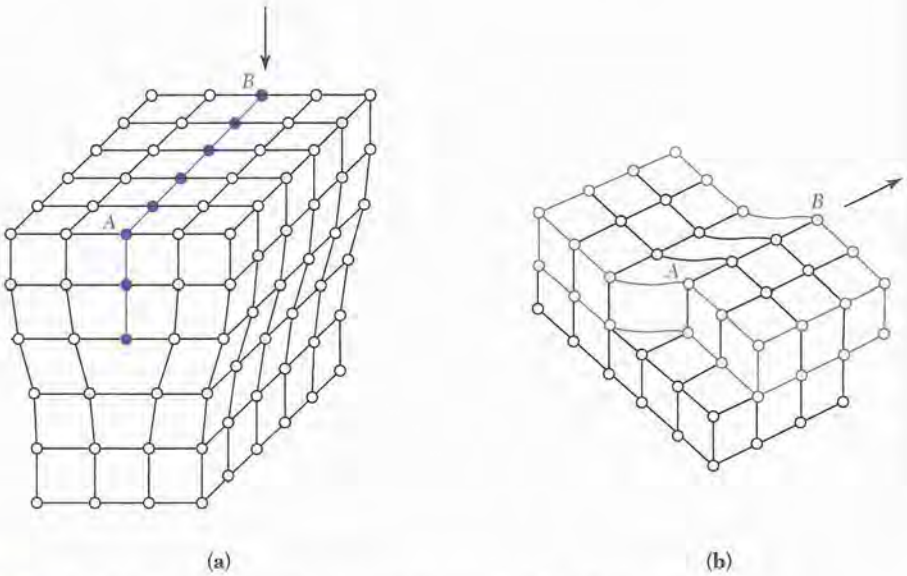


Fig. 16 (*a*) Edge and (*b*) screw dislocation formation in cubic crystals.[10]

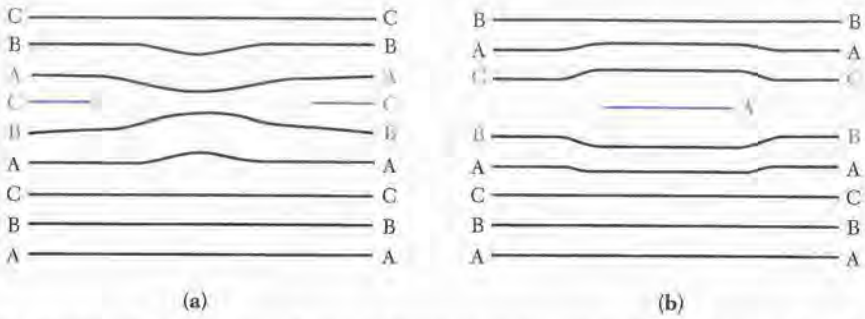


Fig. 17 Stacking fault in semiconductor. (*a*) Intrinsic stacking fault. (*b*) Extrinsic stacking fault.[9]

Precipitates of impurities or dopant atoms make up the fourth class of defects, the volume defects. These defects arise because of the inherent solubility of the impurity in the host lattice. There is a specific concentration of impurity that the host lattice can accept in a solid solution of itself and the impurity. Figure 18 shows solubility versus temperature for a variety of elements in silicon.[11] The solubility of most impurities decreases with decreasing temperature. Thus, at a given temperature, if an impurity is introduced to the maximum concentration allowed by its solubility and the crystal is then cooled to a lower temperature, the crystal can only achieve an equilibrium state by precipitating the impurity atoms in excess of the solubility level. However, the volume mismatch between the host lattice and the precipitates results in dislocations.

Material Properties

Table 4 compares silicon characteristics and the requirements for ultralarge-scale integration[§] (ULSI).[12,13] The semiconductor material properties listed in Table 4 can be measured by various methods. The resistivity is measured by the four-point probe method discussed in Section 3.1, and the minority-carrier lifetime can be measured by the photoconductivity method considered in Section 3.3. The trace impurities such as oxygen and carbon in silicon can be analyzed by the secondary-ion–mass spectroscope (SIMS) techniques to be described in Chapter. 13. Note that although the current capabilities can meet most of the wafer specifications listed in Table 3, many improvements are needed to satisfy the stringent requirements for ULSI technology.[13]

The oxygen and carbon concentrations are substantially higher in Czochralski crystals than in float-zone crystals because of to the dissolution of oxygen from the silica crucible and transport of carbon to the melt from the graphite susceptor during crystal growth. Typical carbon concentrations range from 10^{16} to about 10^{17} atoms/cm^3 and carbon atoms in silicon occupy substitutional lattice sites. The presence of carbon is undesirable because it aids the formation of defects. Typical oxygen concentrations range from 10^{17} to 10^{18} atoms/cm^3. Oxygen, however, has both deleterious and beneficial effects. It can act as a donor, distorting the resistivity of the crystal caused by intentional doping. On the other hand, oxygen in an interstitial lattice site can increase the yield strength of silicon.

In addition, the precipitates of oxygen due to the solubility effect can be used for *gettering*. Gettering is a general term meaning a process that removes harmful impurities or defects from the region in a wafer where devices are fabricated. When the wafer is subjected to high-temperature treatment (e.g., 1050°C in N_2), oxygen evaporates from

[§] The number of components in an ultralarge-scale integrated circuit is more than 10^7.

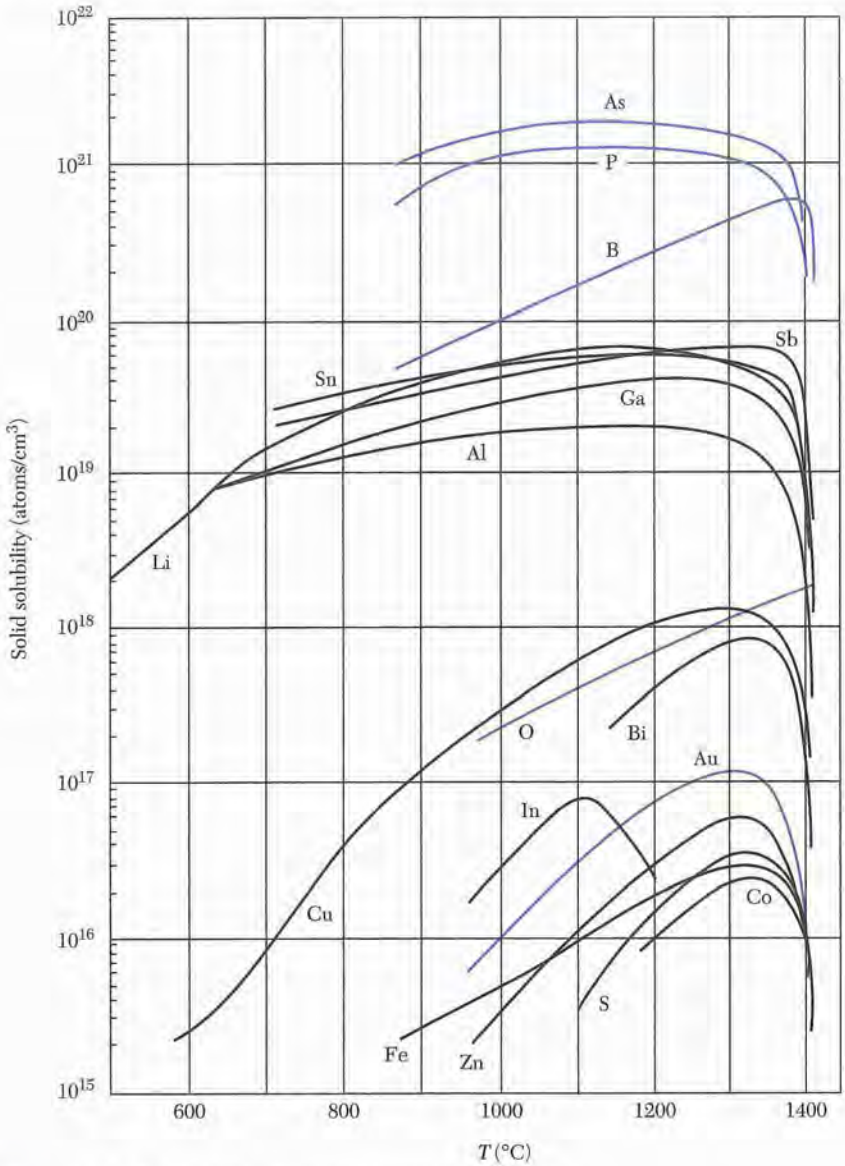


Fig. 18 Solid solubilities of impurity elements in silicon.[11]

the surface. This lowers the oxygen content near the surface. The treatment creates a defect-free (or *denuded*) zone for device fabrication, as shown in the inset[1] of Fig. 19. Additional thermal cycles can be used to promote the formation of oxygen precipitates in the interior of the wafer for gettering of impurities. The depth of the defect-free zone depends on the time and temperature of the thermal cycle and on the diffusivity of oxygen in silicon. Measured results for the denuded zone are shown[1] in Fig. 19. It is possible to obtain Czochralski crystals of silicon that are virtually free of dislocations.

Commercial melt-grown materials of gallium arsenide are heavily contaminated by the crucible. However, for photonic applications, most requirements call for heavily doped materials (between 10^{17} and 10^{18} cm^{-3}). For integrated circuits or for discrete MESFET (metal-semiconductor field-effect transistor) devices, undoped gallium arsenide can be

TABLE 4 Comparison of Silicon Material Characteristics and Requirements for ULSI

Property[a]	Characteristics: Czochralski	Characteristics: Float zone	Requirements for ULSI
Resistivity (phosphorus) *n*-type (ohm-cm)	1–50	1–300 and up	5-50 and up
Resistivity (antimony) *n*-type (ohm-cm)	0.005–10	—	0.001–0.02
Resistivity (boron) *p*-type (ohm-cm)	0.005–50	1–300	5–50 and up
Resistivity gradient (four-point probe) (%)	5–10	20	< 1
Minority carrier lifetime (μs)	30–300	50–500	300–1000
Oxygen (ppma)	5–25	Not detected	Uniform and controlled
Carbon (ppma)	1–5	0.1–1	< 0.1
Dislocation (before processing) (per cm^2)	≤ 500	≤ 500	≤ 1
Diameter (mm)	Up to 200	Up to 100	Up to 300
Slice bow (μm)	≤ 25	≤ 25	< 5
Slice taper (μm)	≤ 15	≤ 15	< 5
Surface flatness (μm)	≤ 5	≤ 5	< 1
Heavy-metal impurities (ppba)	≤ 1	≤ 0.01	< 0.001

[a] ppma, parts per million atoms; ppba, parts per billion atoms.

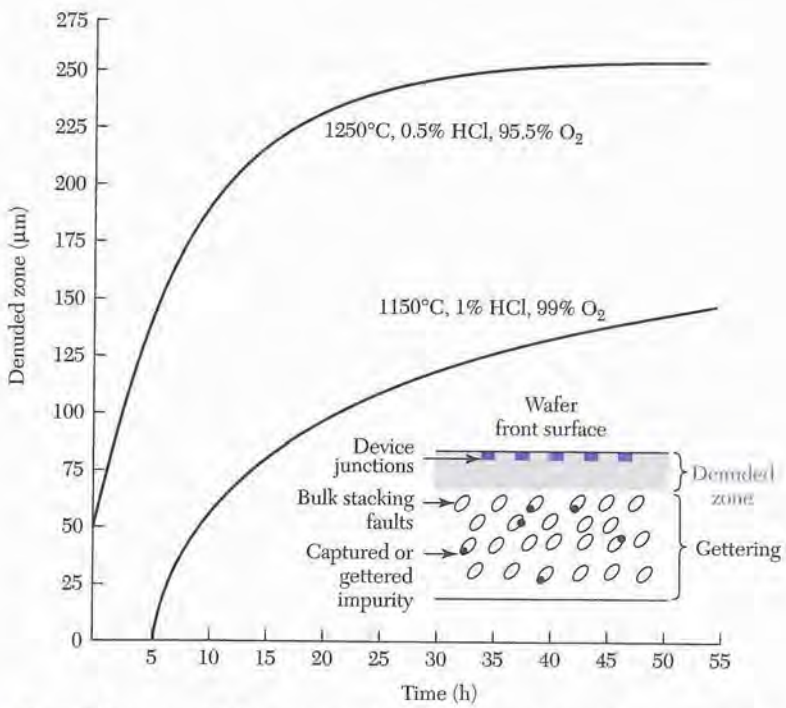


Fig. 19 Denuded zone width for two sets of processing conditions. Inset shows a schematic of the denuded zone and gettering sites in a wafer cross section.[1]

used as the starting material with a resistivity of 10^9 Ω-cm. Oxygen is an undesirable impurity in GaAs because it can form a deep donor level, which contributes to a trapping charge in the bulk of the substrate and increases its resistivity. Oxygen contamination can be minimized by using graphite crucibles for melt growth. The dislocation content for Czochralski-grown gallium arsenide crystals is about two orders of magnitude higher than that for silicon. For Bridgman GaAs crystals, the dislocation density is about an order of magnitude lower than that for Czochralski-grown GaAs crystals.

10.5 EPITAXIAL-GROWTH TECHNIQUES

In an epitaxial process, the substrate wafer acts as the seed crystal. Epitaxial processes are differentiated from the melt-growth processes described in previous sections in that the epitaxial layer can be grown at a temperature substantially below the melting point, typically 30–50% lower. The common techniques for epitaxial growth are chemical-vapor deposition (CVD) and molecular-beam epitaxy (MBE).

10.5.1 Chemical-Vapor Deposition

CVD is also known as vapor-phase epitaxy (VPE). CVD is a process whereby an epitaxial layer is formed by a chemical reaction between gaseous compounds. CVD can be performed at atmospheric pressure (APCVD) or at low pressure (LPCVD).

Figure 20 shows three common susceptors for epitaxial growth. Note that the geometric shape of the susceptor provides the name for the reactor: horizontal, pancake, and barrel susceptors—all made from graphite blocks. Susceptors in the epitaxial reactors are

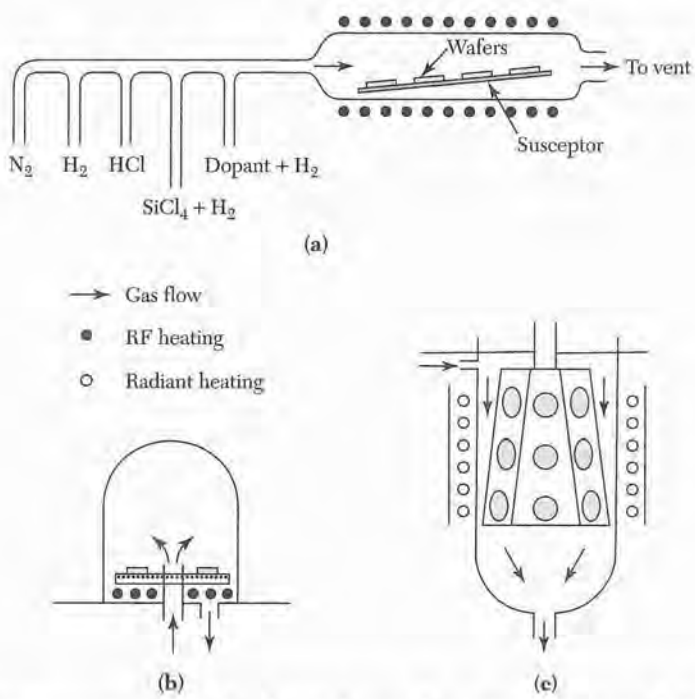


Fig. 20 Three common susceptors for chemical vapor deposition: (*a*) horizontal, (*b*) pancake, and (*c*) barrel susceptor.

analogous to the crucible in the crystal growing furnaces. Not only do they mechanically support the wafer, but in induction-heated reactors they also serve as the source of thermal energy for the reaction. The mechanism of CVD involves a number of steps: (a) the reactants such as the gases and dopants are transported to the substrate region, (b) they are transferred to the substrate surface where they are adsorbed, (c) a chemical reaction occurs, catalyzed at the surface, followed by growth of the epitaxial layer, (d) the gaseous products are desorbed into the main gas stream, and (e) the reaction products are transported out of the reaction chamber.

CVD for Silicon

Four silicon sources have been used for VPE growth. They are silicon tetrachloride ($SiCl_4$), dichlorosilane (SiH_2Cl_2), trichlorosilane ($SiHCl_3$), and silane (SiH_4). Silicon tetrachloride has been the most studied and has the widest industrial use. The typical reaction temperatures is 1200°C. Other silicon sources are used because of lower reaction temperatures. The substitution of a hydrogen atom for each chlorine atom from silicon tetrachloride permits about a 50°C reduction in the reaction temperature. The overall reaction of silicon tetrachloride that results in the growth of silicon layers is

$$SiCl_4\,(gas) + 2H_2\,(gas) \leftrightarrows Si\,(solid) + 4HCl\,(gas). \tag{22}$$

An additional competing reaction is taking place along with that given in Eq. 22:

$$SiCl_4\,(gas) + Si\,(solid) \leftrightarrows 2SiCl_2\,(gas). \tag{23}$$

As a result, if the silicon tetrachloride concentration is too high, etching rather than growth of silicon will take place. Figure 21 shows the effect of the concentration of silicon tetrachloride in the gas on the reaction, where the *mole fraction* is defined as the ratio of the number of molecules of a given species to the total number of molecules.[14] Note that

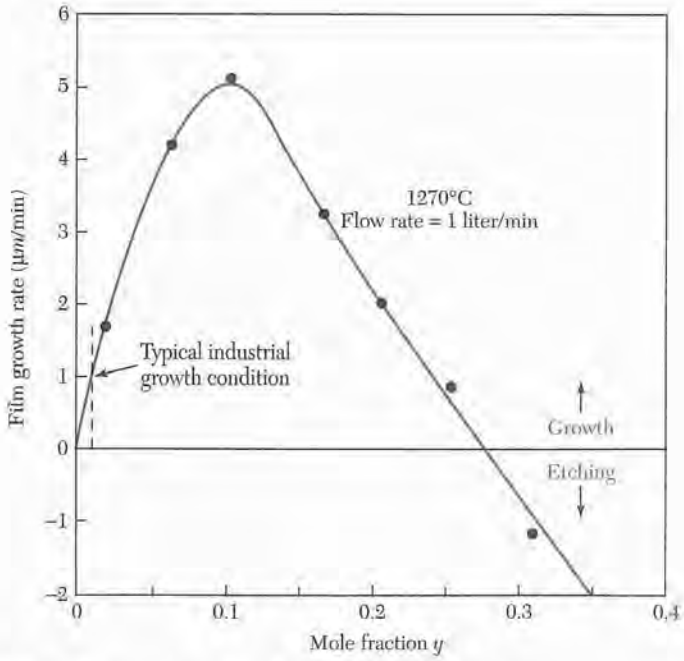


Fig. 21 Effect of $SiCl_4$ concentration on silicon epitaxial growth.[14]

initially the growth rate increases linearly with an increasing concentration of silicon tetrachloride. As the concentration of silicon tetrachloride is increased, a maximum growth rate is reached. Beyond that, the growth rate starts to decrease and eventually etching of the silicon will occur. Silicon is usually grown in the low-concentration region, as indicated in Fig. 21.

The reaction of Eq. 22 is reversible, that is, it can take place in either direction. If the carrier gas entering the reactor contains hydrochloric acid, removal or etching will take place. Actually, this etching operation is used for in-situ cleaning of the silicon wafer prior to epitaxial growth.

The dopant is introduced at the same time as the silicon tetrachloride during epitaxial growth (Fig. 20*a*). Gaseous diborane (B_2H_6) is used as the *p*-type dopant, whereas phosphine (PH_3) and arsine (AsH_3) are used as *n*-type dopants. Gas mixtures are ordinarily used with hydrogen as the diluent to allow reasonable control of flow rates for the desired doping concentration. The dopant chemistry for arsine is illustrated in Fig. 22, which shows arsine being adsorbed on the surface, decomposing, and being incorporated into the growing layer. Figure 22 also shows the growth mechanisms at the surface, which are based on the surface adsorption of host atoms (silicon) as well as the dopant atom (e.g., arsenic) and the movement of these atoms toward the ledge sites.[15] To give these adsorbed atoms sufficient mobility for finding their proper positions within the crystal lattice, epitaxial growth needs relatively high temperatures.

CVD for GaAs

For gallium arsenide, the basic setup is similar to that shown in Fig. 20*a*. Since gallium arsenide decomposes into gallium and arsenic upon evaporation, its direct transport in the vapor phase is not possible. One approach is the use of As_4 for the arsenic component and gallium chloride ($GaCl_3$) for the gallium component. The overall reaction leading to epitaxial growth of gallium arsenide is

$$As_4 + 4GaCl_3 + 6H_2 \rightarrow 4GaAs + 12HCl. \tag{24}$$

The As_4 is generated by thermal decomposition of arsine (AsH_3):

$$4AsH_3 \rightarrow As_4 + 6H_2, \tag{24a}$$

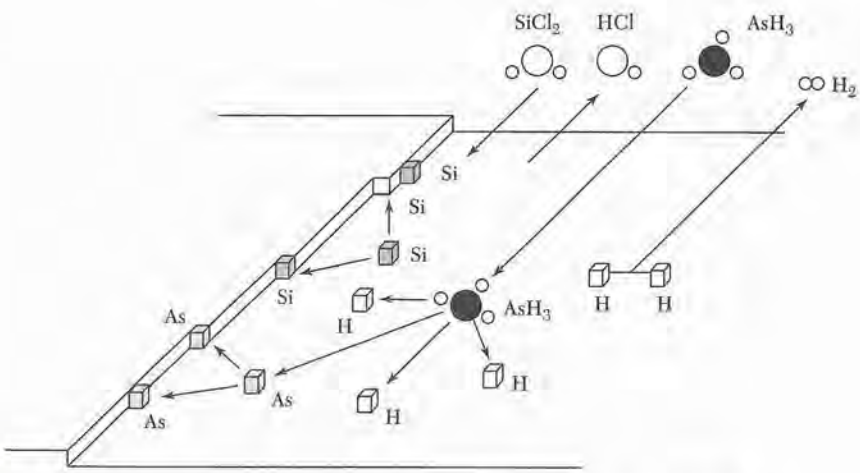


Fig. 22 Schematic representation of arsenic doping and the growing processes.[15]

and the gallium chloride is generated by the reaction

$$6HCl + 2Ga \rightarrow 2GaCl_3 + 3H_2. \tag{24b}$$

The reactants are introduced into a reactor with a carrier gas (e.g., H_2). The gallium arsenide wafers are typically held within the 650°–850°C temperature range. There must be sufficient arsenic overpressure to prevent thermal decomposition of the substrate and the growing layer.

Metalorganic CVD

Metalorganic CVD (MOCVD) is also a VPE process based on pyrolytic reactions. Unlike the conventional CVD, MOCVD is distinguished by the chemical nature of the precursor. It is important for those elements that do not form stable hydrides or halides but that form stable metalorganic compounds with reasonable vapor pressure. MOCVD has been extensibly applied in the heteroepitaxial growth of III-V and II-VI compounds.

To grow GaAs, we can use metalorganic compounds such as trimethylgallium $Ga(CH_3)_3$ for the gallium component and aresine AsH_3 for the arsenic component. Both chemicals can be transported in vapor form into the reactor. The overall reaction is

$$AsH_3 + Ga(CH_3)_3 \rightarrow GaAs + 3CH_4. \tag{25}$$

For Al-containing compounds, such as AlAs, we can use trimethylaluminum $Al(CH_3)_3$. During epitaxy, the GaAs is doped by introducting dopants in vapor form. Diethylzine $Zn(C_2H_5)_2$ and diethylcadmium $Cd(C_2H_5)_2$ are typical p-type dopants and silane SiH_4 is an n-type dopant for III-V compounds. The hydrides of sulfur and selenium or tetramethyltin are also used for n-type dopants; and chromyl chloride is used to dope chromium into GaAs to form semiinsulating layers. Since these compounds are highly poisonous and often spontaneously inflammable in air, rigorous safety precautions are necessary in the MOCVD process.

A schematic of an MOCVD reactor is shown[16] in Fig. 23. Typically, the metalorganic compound is transported to the quartz reaction vessel by hydrogen carrier gas, where it is mixed with AsH_3 in the case of GaAs growth. The chemical reaction is induced by heating the gases to 600°–800°C above a substrate placed on a graphite susceptor using radiofrequency heating. A pyrolytic reaction forms the GaAs layer. The advantages of using metalorganics are that they are volatile at moderately low temperatures and there are no troublesome liquid Ga or In sources in the reactor.

10.5.2 Molecular-Beam Epitaxy

MBE[17] is an epitaxial process involving the reaction of one or more thermal beams of atoms or molecules with a crystalline surface under ultrahigh-vacuum conditions (~10^{-8} Pa).[§] MBE can achieve precise control in both chemical compositions and doping profiles. Single-crystal multilayer structures with dimensions on the order of atomic layers can be made using MBE. Thus, the MBE method enables the precise fabrication of semiconductor heterostructures having thin layers from a fraction of a micron down to a monolayer. In general, MBE growth rates are quite low, and for GaAs, a value of 1 μm/hr is typical.

[§] The international unit for pressure is the Pascal (Pa); 1 Pa = 1 N/m². However, various other units have been used. The conversion of these units are as follows: 1 atm = 760 mm Hg = 760 Torr = 1.013×10^5 Pa.

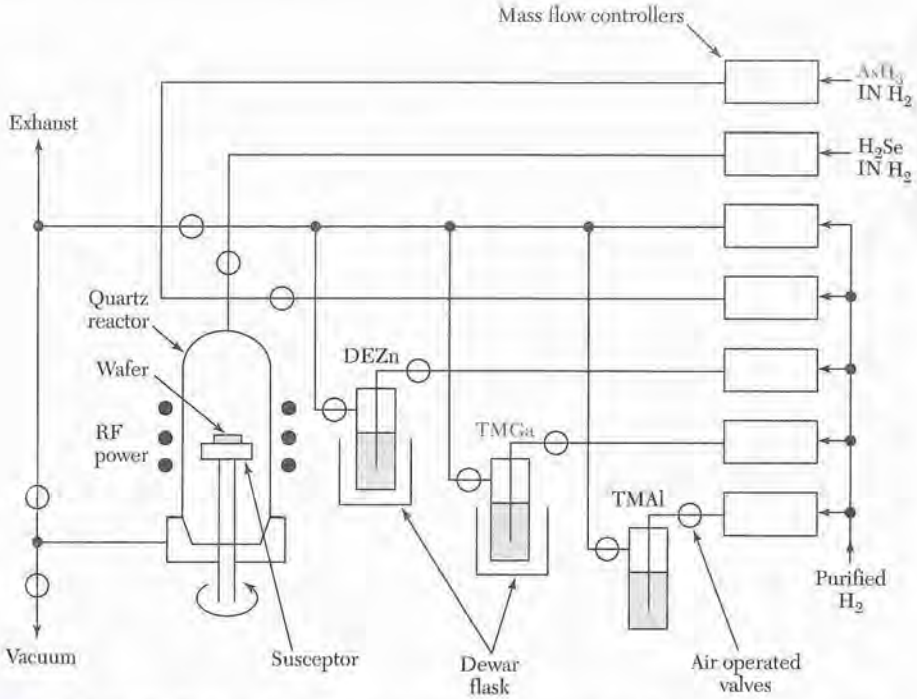


Fig. 23 Schematic diagram of a vertical atmospheric-pressure metalorganic chemical-vapor deposition (MOCVD) reactor.[16] DEZn is diethylozinc $Zn(C_2H_5)_2$, TMGa is trimethylgallium $Ga(CH_3)_3$, and TMAl is trimethylaluminum $Al(CH_3)_3$.

A schematic of MBE system is shown in Fig. 24 for gallium arsenide and related III-V compounds such as $Al_xGa_{1-x}As$. The system represents the ultimate in film deposition control, cleanliness, and in-situ chemical characterization capability. Separate effusion ovens made of pyrolytic boron nitride are used for Ga, As, and the dopants. All the effusion ovens are housed in an ultrahigh-vacuum chamber ($\sim 10^{-8}$ Pa). The temperature of each oven is adjusted to give the desired evaporation rate. The substrate holder rotates continuously to achieve uniform epitaxial layers (e.g., ±1% in doping variations and ±0.5% in thickness variations).

To grow GaAs, an overpressure of As is maintained, since the sticking coefficient of Ga to GaAs is unity, whereas that for As is zero, unless there is a previously deposited Ga layer. For a silicon MBE system, an electron gun is used to evaporate silicon. One or more effusion ovens are used for the dopants. Effusion ovens behave like small-area sources and exhibit a $\cos\theta$ emission, where θ is the angle between the direction of the source and the normal to the substrate surface.

MBE uses an evaporation method in a vacuum systems. An important parameter for vacuum technology is the molecular impingement rate, that is, how many molecules impinge on a unit area of the substrate per unit time. The impingement rate ϕ is a function of the molecular weight, temperature, and pressure. The rate is derived in Appendix K and can be expressed as[18]

$$\phi = P(2\pi mkT)^{-1/2} \tag{26}$$

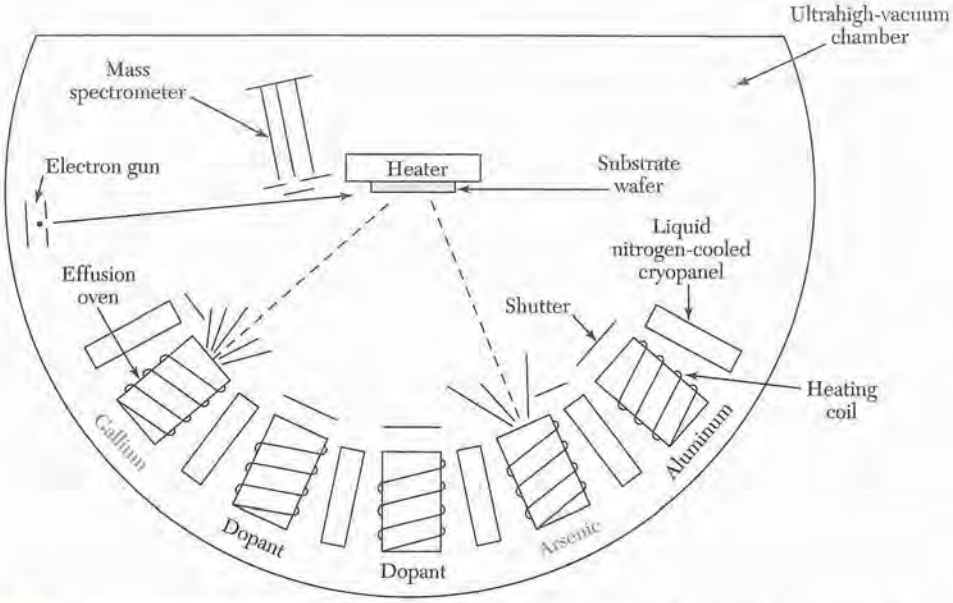


Fig. 24 Arrangement of the sources and substrate in a conventional molecular-beam epitaxy (MBE) system. (Courtesy of M. B. Panish, Bell Laboratories, Lucent Technologies.)

or

$$\boxed{\phi = 2.64 \times 10^{20}\left(\frac{P}{\sqrt{MT}}\right) \text{ molecules / cm}^2\text{-s,}} \tag{26a}$$

where P is the pressure in Pa, m is the mass of a molecule in kg , k is Boltzmams constant in J/K, T is the temperature in Kelvin, and M is the molecular weight. Therefore, at 300 K and 10^{-4} Pa pressure, the impingement rate is 2.7×10^{14} molecules/cm^2-s for oxygen (M = 32).

EXAMPLE 3

At 300 K, the molecular diameter of oxygen is 3.64 Å, and the number of molecules per unit area N_s is 7.54×10^{14} cm^{-2}. Find the time required to form a monolayer of oxygen at pressures of 1, 10^{-4}, and 10^{-8} Pa.

SOLUTION The time required to form a monolayer (assuming 100% sticking) is obtained from the impingement rate:

$$t = \frac{N_s}{\phi} = \frac{N_s\sqrt{MT}}{2.64 \times 10^{20} P}.$$

therefore,

$$\begin{aligned} t &= 2.8\times10^{-4} \approx 0.28 \text{ ms} && \text{at } 1 \text{ Pa,} \\ &= 2.8 \text{ s} && \text{at } 10^{-4}\text{Pa,} \\ &= 7.7 \text{ hr} && \text{at } 10^{-8}\text{Pa.} \end{aligned}$$

To avoid contamination of the epitaxial layer, it is of paramount importance to maintain ultrahigh-vacuum conditions (~10^{-8} Pa) for the MBE process.

During molecular motion, they will collide with other molecules. The average distance traversed by all the molecules between successive collisions with each other is defined as the mean free path. It can be derived from a simple collision theory. A molecule having a diameter d and a velocity v will move a distance $v\delta t$ in the time δt. The molecule suffers a collision with another molecule if its center is anywhere within the distance d of the center of another molecule. Therefore, it sweeps out (without collision) a cylinder of diameter $2d$. The volume of the cylinder is

$$\delta V = \frac{\pi}{4}(2d)^2 v\delta t = \pi d^2 v\delta t, \tag{27}$$

Since there are n molecules/cm^3, the volume associated with one molecule is on the average $1/n$ cm^3. When the volume δV is equal to $1/n$, it must contain on the average one other molecule; thus, a collision would have occurred. Setting $\tau = \delta t$ as the average time between collision,

$$\frac{1}{n} = \pi d^2 v\tau, \tag{28}$$

and the mean free path λ is then

$$\lambda = v\tau = \frac{1}{\pi n d^2} = \frac{kT}{\pi P d^2}. \tag{29}$$

A more rigorous derivation gives

$$\lambda = \frac{kT}{\sqrt{2}\pi P d^2} \tag{30}$$

and

$$\boxed{\lambda = \frac{0.66}{P(\text{in Pa})} \text{ cm}} \tag{31}$$

for air molecules (equivalent molecular diameter of 3.7Å) at room temperature. Therefore, at a system pressure of 10^{-8} Pa, λ would be 660 km.

EXAMPLE 4

Assume an effusion oven geometry of area $A = 5$ cm^2 and a distance L between the top of the oven and the gallium arsenide substrate of 10 cm. Calculate the MBE growth rate for the effusion oven filled with gallium arsenide at 900°C. The surface density of gallium atom is 6×10^{14}cm^{-2}, and the average thickness of a monolayer is 2.8 Å.

SOLUTION On heating gallium arsenide, the volatile arsenic vaporizes first, leaving a gallium-rich solution. Therefore, only the pressures marked Ga-rich in Fig.11 are of interest. The pressure at 900°C is 5.5×10^{-2} Pa for gallium and 1.1 Pa for arsenic (As_2). The arrival rate can be obtained from the impingement rate (Eq. 26a) by multiplying it by $A/\pi L^2$:

$$\text{Arrival rate} = 2.64 \times 10^{20}\left(\frac{P}{\sqrt{MT}}\right)\left(\frac{A}{\pi L^2}\right) \text{ molecules / cm}^2\text{-s}.$$

The molecular weight M is 69.72 for Ga and 74.92×2 for As_2. Substituting values of P, M, and T (1173 K) into the above equation gives

$$\text{Arrival rate} = 8.2 \times 10^{14}/\text{cm}^2\text{-s} \quad \text{for Ga,}$$
$$= 1.1 \times 10^{16}/\text{cm}^2\text{-s} \quad \text{for } As_2.$$

The growth rate of gallium arsenide is found to be governed by the arrival rate of gallium. The growth rate is

$$\frac{8.2 \times 10^{14} \times 2.8}{6 \times 10^{14}} \approx 3.8\ \text{Å}/s = 23\ \text{nm}/\text{min}.$$

Note that the growth rate is relatively low compared with that of VPE.

There are two ways to clean a surface in situ for MBE. High-temperature baking can decompose native oxide and remove other adsorbed species by evaporation or diffusion into the wafer. Another approach is to use a low-energy ion beam of an inert gas to sputter-clean the surface, followed by a low-temperature annealing to reorder the surface lattice structure.

MBE can use a wide variety of dopants (compared with CVD and MOCVD), and the doping profile can be exactly controlled. However, the doping process is similar to the vapor-phase growth process: a flux of evaporated dopant atoms arrives at a favorable lattice site and is incorporated along the growing interface. Fine control of the doping profile is achieved by adjusting the dopant flux relative to the flux of silicon atoms (for silicon epitaxial films) or gallium atoms (for gallium arsenide epitaxial films). It is also possible to dope the epitaxial film using a low-current, low-energy ion beam to implant the dopant (see Chapter 13).

The substrate temperatures for MBE range from 400°–900°C; and the growth rates range from 0.001 to 0.3 μm/min. Because of the low-temperature process and low-growth rate, many unique doping profiles and alloy compositions not obtainable from conventional CVD can be produced in MBE. Many novel structures have been made using MBE. These include the *superlattice*, which is a periodic structure consisting of alternating ultrathin layers with its period less than the electron mean free path (e.g., $GaAs/Al_xGa_{1-x}As$, with each layer 10 nm or less in thickness), and the heterojunction field-effect transistors discussed in Chapter 7.

A further development in MBE has replaced the group III elemental sources by metalorganic compounds such as trimethygallium (TMG) or triethylegallium (TEG). This approach is called metalorganic molecular-beam epitaxy (MOMBE) and is also referred to as chemical-beam epitaxy (CBE). Although closely related to MOCVD, it is considered a special form of MBE. The metalorganics are sufficiently volatile that they can be admitted directly into the MBE growth chamber as a beam and are not decomposed before forming the beam. The dopants are generally elemental sources, typically Be for *p*-type and Si or Sn for *n*-type GaAs epitaxial layers.

10.6 STRUCTURES AND DEFECTS IN EPITAXIAL LAYERS

10.6.1 Lattice-Matched and Strained-Layer Epitaxy

For conventional homoepitaxial growth, a single-crystal semiconductor layer is grown on a single-crystal semiconductor substrate. The semiconductor layer and the substrate are the same material having the same lattice constant. Therefore, homoepitaxy is, by definition, a lattice-matched epitaxial process. The homoepitaxial process offers one important means of controlling the doping profiles so that device and circuit performance can be optimized. For example, an *n*-type silicon layer with a relatively low doping concentration

can be grown epitaxially on an n^+-silicon substrate. This structure substantically reduces the series resistance associated with the substrate.

For heteroepitaxy, the epitaxial layer and the substrate are two different semiconductors, and the epitaxial layer must be grown in such a way that an idealized interfacial structure is maintained. This implies that atomic bonding across the interface must be continuous without interruption. Therefore, the two semiconductor must either have the same lattice spacing or be able to deform to adopt a common spacing. These two cases are referred to as lattice-matched epitaxy and strained-layer epitaxy.

Figure 25*a* shows a lattice-matched epitaxy where the substrate and the film have the same lattice constant. An important example is the epitaxial growth of $Al_xGa_{1-x}As$ on a GaAs substrate where for any x between 0 and 1, the lattice constant of $Al_xGa_{1-x}As$ differs from that of GaAs by less than 0.13%.

For the lattice-mismatched case, if the epitaxial layer has a larger lattice constant and is flexible, it will be compressed in the plane of growth to conform to the substrate spacing. Elastic forces then compel it to dilate in a direction perpendicular to the interface. This type of structure is called strained-layer epitaxy and is illustrated[19] in Fig. 25*b*. On the other hand, if the epitaxial layer has a smaller lattice constant, it will be dilated in the plane of growth and compressed in a direction perpendicular to the interface. In the above strained-layer epitaxy, as the strained-layer thickness increases, the total number of atoms under strain or the distorted atomic bonds grows, and at some point misfit dislocations are nucleated to relieve the homogeneous strain energy. This thickness is referred to as the *critical layer thickness* for the system. Figure 25*c* shows the case in which there are edge dislocations at the interface.

The critical layer thicknesses for two material systems are shown[20] in Fig. 26. The upper curve is for the strained-layer epitaxy of a Ge_xSi_{1-x} layer on a silicon substrate, and

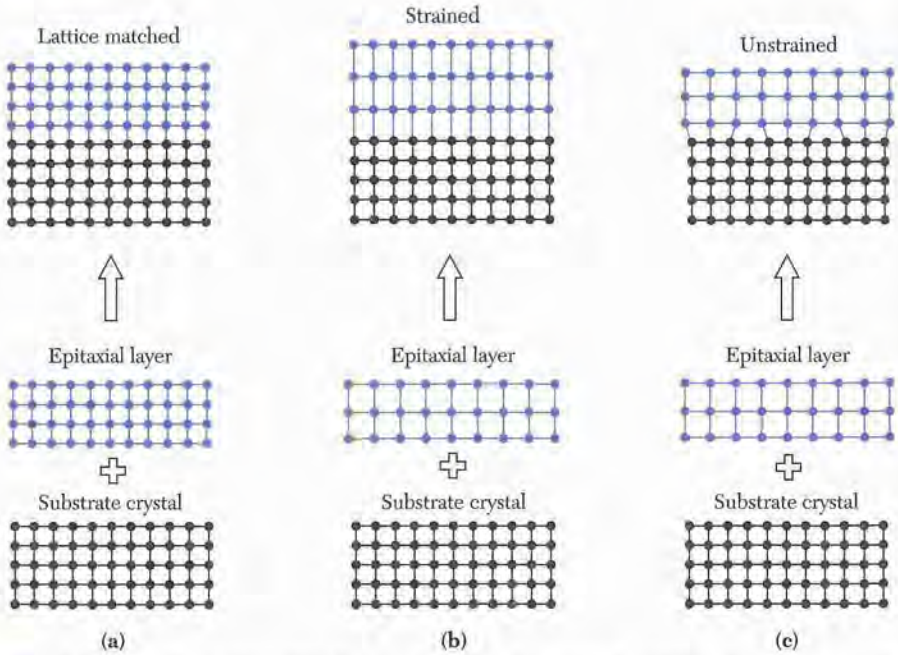


Fig. 25 Schematic illustration of (*a*) lattice-matched, (*b*) strained, and (*c*) relaxed heteroepitaxial structures.[19] Homoepitaxy is structurally identical to the lattice-matched heteroepitaxy.

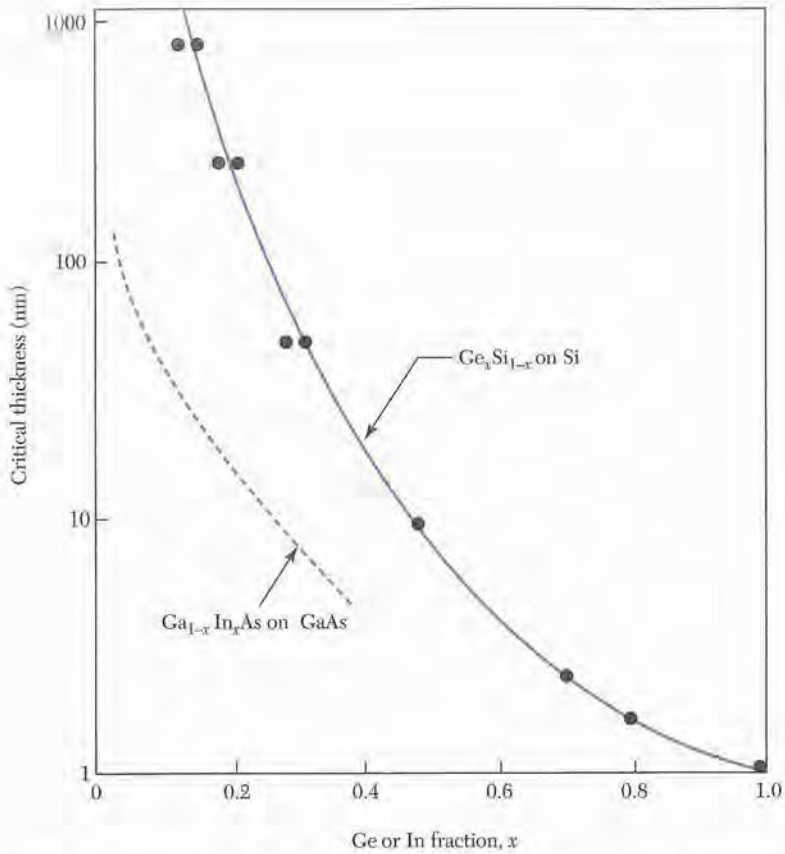


Fig. 26 Experimentally determined critical layer thickness for defect-free, strained-layer epitaxy[20] of Ge_xSi_{1-x} on Si, and $Ga_{1-x}In_xAs$ on GaAs.

the lower curve is for a $Ga_{1-x}In_xAs$ layer on a GaAs substrate. For example, for $Ge_{0.3}Si_{0.7}$ on silicon, the maximum epitaxial thickness is about 70 nm. For thicker films, edge dislocations will occur.

A related heteroepitaxial structure is the strained-layer superlattice (SLS). A superlattice is an artificial one-dimensional periodic structure constituted by different materials with a period of about 10 nm. Figure 27 shows[17] a SLS having two semiconductors with different equilibrium lattice constants $a_1 > a_2$, grown in a structure with a common inplane lattice constant b, where $a_1 > b > a_2$. For sufficiently thin layers, the lattice mismatch is accommodated by uniform strains in the layers. Under these condition, no misfit dislocations are generated at the interfaces, so high-quality crystalline materials can be obtained. These artificially structured materials can be grown by MBE. These materials provide a new area in semiconductor research and permit new solid-state devices, especially for high-speed and photonic applications.

10.6.2 Defects in Epitaxial Layers

Defects in epitaxial layers will degrade device properties. For example, defects can result in reduced mobility or increased leakage current. The defects in epitaxial layers can be categorized into five groups.

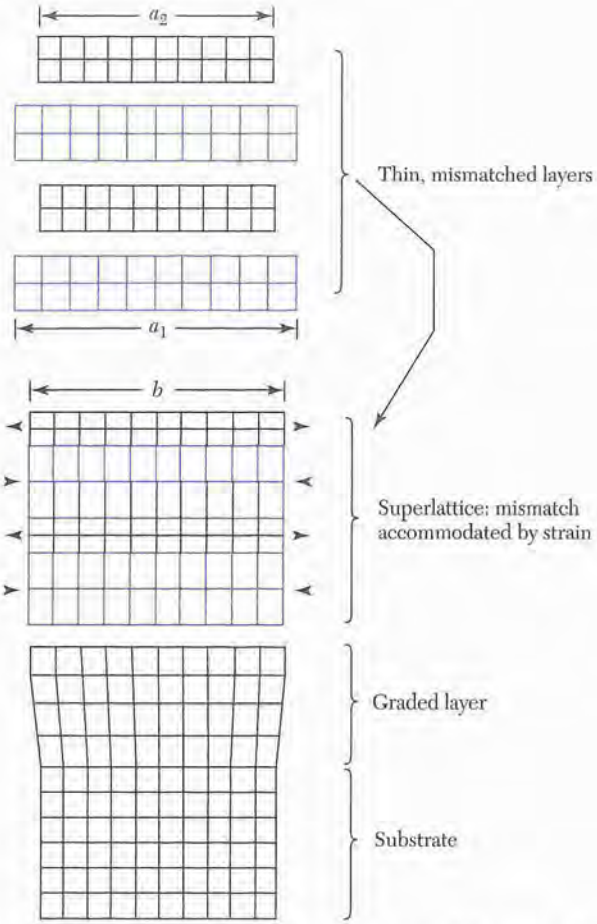


Fig. 27 Illustration of the elements and formation of an strained-layer superlattice.[17] Arrows show the direction of the strain.

1. Defects from the substrates. These defects may propagate from the substrate into the epitaxial layer. To avoid these defects, dislocation-free semiconductor substrates are required.
2. Defects from the interface. The oxide precipitates or any contamination at the interface of the epitaxial layer and substrate may cause the formation of misoriented clusters or nuclei containing stacking faults. These clusters and stacking faults may coalesce with normal nuclei and grow into the film in the shape of an inverted pyramid. To avoid these defects, the surface of the substrate must be thoroughly cleaned. In addition, an in-situ etch back may be used such as the reversable reaction of Eq. 22.
3. Precipitates or dislocation loops. Their formation is due to supersaturation of impurities or dopants. Epitaxial layers containing very high intentional or unintentional dopants or impurity concentrations are susceptible to such defects.
4. Low-angle grain boundaries and twins. Any misoriented areas of an epitaxial film during growth may meet and coalesce to form these defects.

5. Edge dislocations. These are formed in the heteroepitaxy of two lattice-mismatched semiconductors. If both lattices are rigid, they will retain their fundamental lattice spacings, and the interface will contain rows of misbonded atoms described as misfit or edge dislocations. The edge dislocations can also form in a strained layer when the layer thickness becomes larger than the critical layer thickness.

SUMMARY

Several techniques are available to grow single crystals of silicon and gallium arsenide. For silicon crystals, we use sand (SiO_2) to produce polycrystalline silicon, which then serves as the raw material in a Czochralski puller. A seed crystal with the desired orientation is used to grow a large ingot from the melt. Over 90% of silicon crystals are prepared by this technique. During crystal growth, the dopant in the crystal will redistribute. A key parameter is the segregation coefficient, i.e., the ratio of the dopant concentration in the solid to that in the melt. Since most of the coefficients are less than 1, the melt becomes progressively enriched with the dopant as the crystal grows.

Another growth technique for silicon is the float-zone process. It offers lower contamination than that normally obtained from the Czochralski technique. Float-zone crystals are used mainly for high-power, high-voltage devices where high-resistivity materials are required.

To make GaAs, we use chemically pure gallium and arsenic as the starting materials that are synthesized to form polycrystalline GaAs. Single crystals of GaAs can be grown by the Czochralski technique. However, a liquid encapsulant (e.g., B_2O_3) is required to prevent decomposition of GaAs at the growth temperature. Another technique is the Bridgman process, which uses a two-zone furnace for gradual solidification of the melt.

After a crystal is grown, it usually goes through wafer-shaping operations to give an end product of highly polished wafers with a specified diameter, thickness, and surface orientation. For example, 200 mm silicon wafers for a MOSFET (metal-oxide-semiconductor field-effect transistor) fabrication line should have a diameter of 200 ± 1 mm, a thickness of 0.725 ± 0.01 mm, and a surface orientation of $(100) \pm 1°$. Wafers with diameters larger than 200 mm are being manufactured for future integrated circuits. Their specifications are listed in Table 3.

A real crystal has defects that influence the electrical, mechanical, and optical properties of the semiconductor. These defects are point defects, line defects, area defects, and volume defects. We also discussed means to minimize such defects. For the more demanding ULSI applications, the dislocation density must be less than 1 per square centimeter. Other important requirements are listed in Table 4.

A technology closely related to crystal growth is the epitaxial process. In this process, the substrate wafer is the seed. High-quality, single-crystal films can be grown at a temperature 30%–50% lower than the melting point. The common techniques for epitaxial growth are chemical-vapor deposition (CVD), metalorganic CVD (MOCVD), and molecular-beam epitaxy (MBE). CVD and MOCVD are chemical deposition processes. Gases and dopants are transported in vapor form to the substrate, where a chemical reaction occurs that results in the deposition of the epitaxial layer. Inorganic compounds are used for CVD, whereas metalorganic compounds are used for MOCVD. MBE, on the other hand, is a physical deposition process. It is done by the evaporation of a species in an ultrahigh vacuum system. Because it is a low-temperature process that has a low growth rate, MBE can grow single-crystal, multilayer structures with dimensions on the order of atomic layers.

In addition to conventional homoepitaxy, such as n-type silicon on an n^+-silicon substrate, we have also considered heteroepitaxy that includes lattice-matched and strained-layer structures. For strained-layer epitaxy, there is a critical layer thickness above which edge dislocations will nucleate to relieve the strain energy.

Besides the edge dislocations in an epitaxial layer, we have defects from the substrate, defects from the interface, precipitates, and low-angle grain boundaries and twins. These defects degrade device performance. Various means have been presented to minimize or even to eliminate these defects so that a defect-free semiconductor layer can be grown either homoepitaxially or heteroepitaxially.

REFERENCES

1. C. W. Pearce, "Crystal Growth and Wafer Preparation" and "Epitaxy," in S. M. Sze, Ed., *VLSI Technology*, McGraw-Hill, New York,1983.
2. T. Abe, "Silicon Crystals for Giga-Bit Scale Integration," in T. S. Moss, Ed., *Handbook on Semiconductors*, Vol. 3, Elsevier Science B. V., Amsterdam/New York, 1994.
3. W. R. Runyan, *Silicon Semiconductor Technology*, McGraw-Hill, New York, 1965.
4. W. G. Pfann, *Zone Melting*, 2nd Ed., Wiley, New York, 1966.
5. E. W. Hass and M. S. Schnoller, "Phosphorus Doping of Silicon by Means of Neutron Irradiation," *IEEE Trans. Electron Devices*, **ED-23**, 803 (1976).
6. M. Hansen, *Constitution of Binary Alloys*, McGraw-Hill, New York, 1958.
7. S. K. Ghandhi, *VLSI Fabrication Principles*, Wiley, New York, 1983.
8. J. R. Arthur, "Vapor Pressures and Phase Equilibria in the GaAs System," *J. Phys. Chem. Solids*, 28, 2257 (1967).
9. B. El-Kareh, *Fundamentals of Semiconductor Processing Technology*, Kluwer Academic, Boston, 1995.
10. C. A. Wert and R. M. Thomson, *Physics of Solids*, McGraw-Hill, New York, 1964.
11. (a) F. A. Trumbore, "Solid Solubilities of Impurity Elements in Germanium and Silicon," *Bell Syst. Tech. J.*, **39**, 205 (1960); (b) R. Hull, *Properties of Crystalline Silicon*, INSPEC, London, 1999.
12. Y. Matsushita, "Trend of Silicon Substrate Technologies for 0.25 μm Devices," *Proc. VLSI Technol. Workshop*, Honolulu, (1996).
13. *The International Technology Roadmap for Semiconductors*, Semiconductor Industry Association, San Jose, CA, 1999.
14. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967.
15. R. Reif, T. I. Kamins, and K. C. Saraswat, "A Model for Dopant Incorporation into Growing Silicon Epitaxial Films," *J. Electrochem. Soc.*, **126**, 644, 653 (1979).
16. R. D Dupuis, *Science*, "Metalorganic Chemical Vapor Deposition of III–V Semiconductors", **226**, 623 (1984).
17. M. A. Herman and H. Sitter, *Molecular beam Epitaxy*, Springer-Verlag, Berlin, 1996.
18. A. Roth, *Vacuum Technology*, North-Holland, Amsterdam, 1976.
19. M. Ohring, *The Materials Science of Thin Films*, Academic, New York, 1992.
20. J. C. Bean, "The Growth of Novel Silicon Materials," *Physics Today*, **39**, 10, 36 (1986).

PROBLEMS (* DENOTES DIFFICULT PROBLEMS)

FOR SECTION 10.1 SILICON CRYSTAL GROWTH FROM THE MELT

1. Plot the doping distribution of arsenic at distances of 10, 20, 30, 40, and 45 cm from the seed in a silicon ingot 50 cm long that has been pulled from a melt with an initial doping concentration of 10^{17} cm^{-3}.

2. In silicon, the lattice constant is 5.43 Å. Assume a hard-sphere model. (a) Calculate the radius of a silicon atom. (b) Determine the density of silicon atoms in atoms/cm^3. (c) Use the Avogadro constant to find the density of silicon.

3. Assuming that a 10 kg pure silicon charge is used, what is the amount of boron that must be added to get the boron-doped silicon having a resistivity of 0.01 Ω-cm when one half of the ingot is grown?

4. A silicon wafer 1-mm thick having a diameter of 200 mm contains 5.41 mg of boron uniformly distributed in substitutional sites. Find (a) the boron concentration in atoms/cm^3 and, (b) the average distance between boron atoms.

5. The seed crystal used in Crochralski process is usually necked down to a small diameter (5.5 mm) as a means to initiate dislocation-free growth. If the critical yield strength of silicon is 2×10^6 g/cm^2, calculate the maximum length of a silicon ingot 200 mm diameter that can be supported by such a seed.

6. Plot the curve of C_s/C_0 value for k_0= 0.05 in Czochralski technique.

7. A Czochralski grown crystal is doped with boron. Why is the boron concentration larger at the tail end of the crystal than at the seed end?

8. Why is the impurity concentration larger in the center of the wafer than at its perimeter?

FOR SECTION 10.2 SILICON FLOAT-ZONE PROCESS

9. We use the float-zone process to purify a silicon ingot that contains a uniform gallium concentration of 5×10^{16} cm^{-3}. One pass is made with a molten zone 2 cm long. Over what distance is the resulting gallium concentration below 5×10^{15} cm^{-3}?

10. From Eq. 18 find the C_s/C_0 value at x/L = 1 and 2 with k_e = 0.3.

11. If p^+-n abrupt-junction diodes are fabricated using the silicon materials shown in Fig. 9, find the percentage change of breakdown voltages for the conventionally doped silicon and the neutron-irradiated silicon.

FOR SECTION 10.3 GaAs CRYSTAL-GROWTH TECHNIQUES

12. From Fig. 10, if C_m is 20%, what fraction of the liquid at T_b will remain?

13. From Fig. 11, explain why the GaAs liquid always becomes gallium rich?

FOR SECTION 10.4 MATERIAL CHARACTERIZATION

14. The equilibrium density of vacancy n_s is given by $N\exp(-E_s/kT)$, where N is the density of semiconductor atoms and E_s is the energy of formation. Calculate n_s in silicon at 27°C, 900°C, and 1200°C. Assume E_s = 2.3 eV.

15. Assume the energy of formation (E_f) of a Frenkel-type defect to be 1.1 eV and estimate the defect density at 27°C and 900°C. The equilibrium density of Frenkel type defects is given by $n_f = \sqrt{NN'}\,e^{-E_f/2kT}$, where N is the atomic density of silicon (cm^{-3}), and N' is the density of available interstitial sites (cm^{-3}) and it is represented by $N' = 1 \times 10^{27} e^{-3.8(\text{eV})/kT}\,\text{cm}^{-3}$.

16. How many chips of area 400 mm^2 can be placed on a wafer 300 mm in diameter? Explain your assumptions on the chip shape and unused wafer perimeter?

FOR SECTION 10.5 EPITAXIAL-GROWTH TECHNIQUES

*17. Find the average molecular velocity of air at 300 K (the molecular weight for air is 29).

18. The distance between source and wafer in a deposition chamber is 15 cm. Estimate the pressure at which this distance become 10% of the mean free path of source molecules.

*19. Find the number of atoms per unit area, N_s, needed to form a monolayer under close-packing condition (i.e., each atom is in contact with its six neighboring atoms), assuming the diameter d of the atom is 4.68 Å.

*20. Assume an effusion oven geometry of $A = 5$ cm^2 and $L = 12$ cm. (a) Calculate the arrival rate of gallium and the MBE growth rate for the effusion oven filled with gallium arsenide at 970°C. (b) For a tin effusion oven operated at 700°C under the same geometry, calculate the doping concentration (assuming tin atoms are fully incorporated in the gallium arsenide grown at the aforementioned rate). The molecular weight for tin is 118.69, and the pressure at 700°C for tin is 2.66×10^{-6} Pa.

FOR SECTION 10.6 STRUCTURES AND DEFECTS IN EPITAXIAL LAYERS

21. Find the maximum percentage of In, i.e., the x value for $Ga_xIn_{1-x}As$ film grown on GaAs substrate without formation of misfit disloation, if the final film thickness is 10 nm.

22. The lattice misfit, f, of a film is defined as $f \equiv [a_0(s) - a_0(f)]/a_0(f) = \Delta a_0/a_0$ where $a_0(s)$ and $a_0(f)$ are the unstrained lattice constants of the substrate and the film, respectively. Find the f values for InAs-GaAs and Ge-Si systems.

CHAPTER 11

Film Formation

- 11.1 THERMAL OXIDATION
- 11.2 DIELECTRIC DEOPOSITION
- 11.3 POLYSILICON DEPOSITION
- 11.4 METALLIZATION
- SUMMARY

To fabricate discrete devices and integrated circuits, we use many different kinds of thin films. We can classify thin films into four groups: thermal oxides, dielectric layers, polycrystalline silicon, and metal films. Figure 1 shows a schematic view of a conventional silicon n-channel MOSFET (metal-oxide-semiconductor field-effect transitor) that uses all four groups of films. The first important thin film from the thermal oxide group is the gate-oxide layer under which a conducting channel can be formed between the source and the drain. A related layer is the field oxide, which provides isolation from other devices. Both gate and field oxides generally are grown by a thermal oxidation process because only thermal oxidation can provide the highest-quality oxides having the lowest interface trap densities.

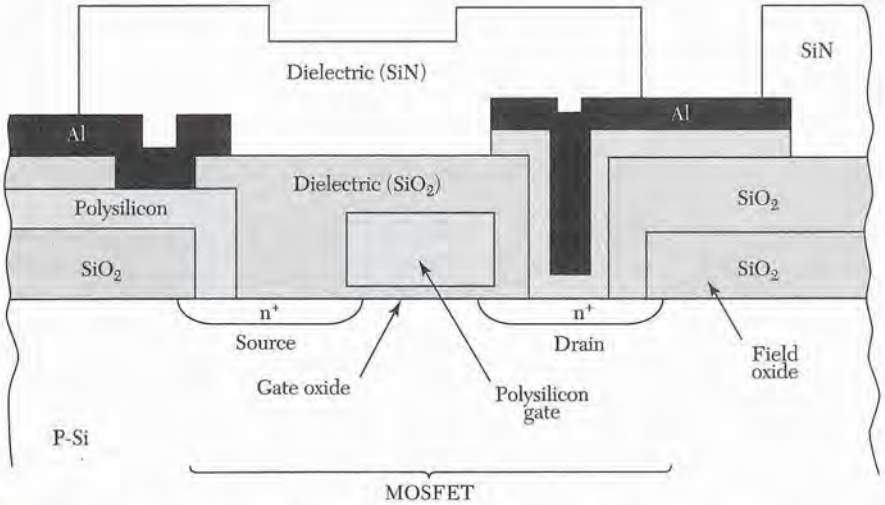


Fig. 1 Schematic cross section of a metal-oxide-semiconductor field-effect transitor (MOSFET).

Dielectric layers such as silicon dioxide and silicon nitride are used for insulation between conducting layers, for diffusion and ion implantation masks, for capping doped films to prevent the loss of dopants, and for passivation to protect devices from impurities, moisture, and scratches. Polycrystalline silicon, usually referred to as polysilicon, is used as a gate electrode material in MOS devices, a conductive material for multilevel metallization, and a contact material for devices with shallow junctions. Metal films such as aluminum and silicides are used to form low-resistance interconnections, ohmic contacts, and rectifying metal-semiconductor barriers.

Specifically, we cover the following topics:

- The thermal oxidation process to form silicon dioxide (SiO_2).
- Deposition techniques to form low–dielectric-constant and high–dielectric-constant films as well as polysilicon films.
- Deposition techniques to form aluminum and copper interconnections as well as the related global planarization process.
- Characteristics of these thin films and their compatibility with integrated-circuit processing.

11.1 THERMAL OXIDATION

Semiconductors can be oxidized by various methods. These include thermal oxidation, electrochemical anodization, and plasma reaction. Among these methods, thermal oxidation is by far the most important for silicon devices. It is the key process in modern silicon integrated-circuit technology. For gallium arsenide, however, thermal oxidation results in generally nonstoichiometric films. The oxides provide poor electrical insulation and semiconductor surface protection; hence, these oxides are rarely used in gallium arsenide technology. Consequently, in this section we concentrate on the thermal oxidation of silicon.

The basic thermal oxidation setup is shown[1] in Fig. 2. The reactor consists of a resistance-heated furnace, a cylindrical fused-quartz tube containing the silicon wafers held vertically in a slotted quartz boat, and a source of either pure dry oxygen or pure water vapor. The loading end of the furnace tube protudes into a vertical flow hood where a filtered flow of air is maintained. Flow is directed as shown by the arrow in Fig. 2. The hood reduces dust and particulate matters in the air surrounding the wafers and mini-

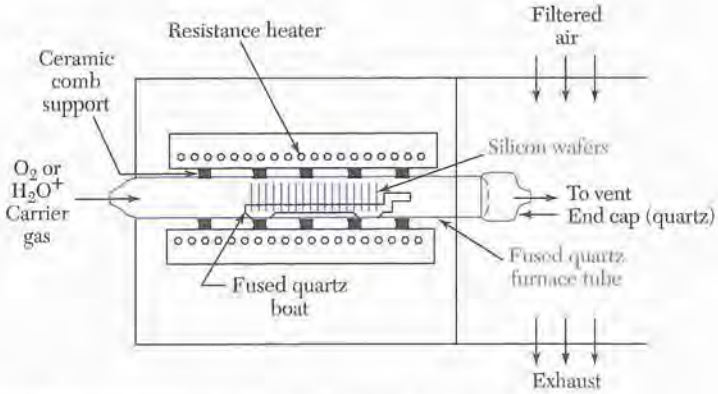


Fig. 2 Schematic cross section of a resistance-heated oxidation furnace.

mizes contamination during wafer loading. The oxidation temperature is generally in the range of 900°–1200°C and the typical gas flow rate is about 1 liter/min. The oxidation system uses microprocessors to regulate the gas flow sequence, to control the automatic insertion and removal of silicon wafers, to ramp the temperature up (i.e., to increase the furnace temperature linearly) from a low temperature to the oxidation temperature so that the wafers will not warp due to sudden temperature change, to maintain the oxidation temperature to within ±1°C, and to ramp the temperature down when oxidation is completed.

11.1.1 Kinetics of Growth

The following chemical reactions describe the thermal oxidation of silicon in oxygen or water vapor:

$$\text{Si (solid)} + O_2\,\text{(gas)} \rightarrow SiO_2\,\text{(solid)}, \tag{1}$$

$$\text{Si (solid)} + 2H_2O\,\text{(gas)} \rightarrow SiO_2\,\text{(solid)} + 2H_2\,\text{(gas)}. \tag{2}$$

The silicon-silicon dioxide interface moves into the silicon during the oxidation process. This creates a fresh interface region, with surface contamination on the original silicon ending up on the oxide surface. The densities and molecular weights of silicon and silicon dioxide are used in the following example to show that growing an oxide of thickness x consumes a layer of silicon $0.44x$ thick (Fig. 3).

EXAMPLE 1

If a silicon oxide layer of thickness x is grown by thermal oxidation, what is the thickness of silicon being consumed? The molecular weight of Si is 28.9 g/mol, and the density of Si is 2.33 g/cm^3. The corresponding values for SiO_2 are 60.08 g/mol and 2.21 g/cm^3.

SOLUTION The volume of 1 mol of silicon is

$$\frac{\text{Molecular weight of Si}}{\text{Density of Si}} = \frac{28.9\ \text{g/mole}}{2.33\ \text{g/cm}^3} = 12.06\ \text{cm}^3/\text{mol}.$$

The volume of 1 mol of silicon dioxide is

$$\frac{\text{Molecular weight of } SiO_2}{\text{Density of } SiO_2} = \frac{60.08\ \text{g/mol}}{2.21\ \text{g/cm}^3} = 27.18\ \text{cm}^3/\text{mol}.$$

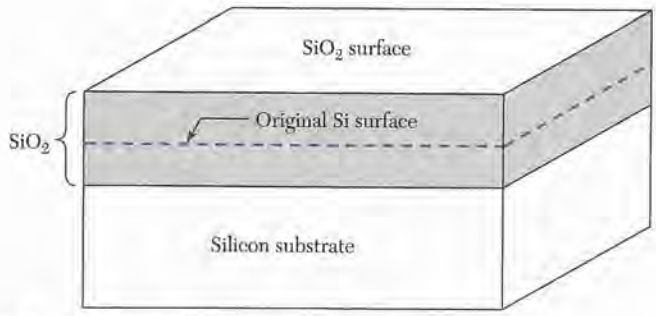


Fig. 3 Growth of silicon dioxide by thermal oxidation.

Since 1 mol of silicon is converted to 1 mol of silicon dioxide,

$$\frac{\text{Thickness of Si} \times \text{area}}{\text{Thickness of } SiO_2 \times \text{area}} = \frac{\text{volume of 1 mol of Si}}{\text{volume of 1 mol of } SiO_2},$$

$$\boxed{\frac{\text{Thickness of Si}}{\text{Thickness of } SiO_2} = \frac{12.06}{27.18} = 0.44,}$$

Thickness of silicon = 0.44 (thickness of SiO_2).

For example, to grow a silicon dioxide layer of 100 nm, a layer of 44 nm of silicon is consumed.

The basic structural unit of thermally grown silicon dioxide is a silicon atom surrounded tetrahedrally by four oxygen atoms, as illustrated[1] in Fig. 4*a*. The silicon-to-oxygen internuclear distance is 1.6 Å, and the oxygen-to-oxygen internuclear distance is 2.27 Å. These tetrahedra are joined together at their corners by oxygen bridges in a variety of ways to form the various phases or structures of silicon dioxide (also called silica). Silica has several crystalline structures (e.g., quartz) and an amorphous structure. When silicon is thermally oxidized, the silicon dioxide structure is amorphous. Typically amorphous silica has a density of 2.21 g/cm^3, compared with 2.65 g/cm^3 for quartz.

The basic difference between the crystalline and amorphous structures is that the former is a periodic structure, extending over many molecules, whereas the latter has no periodic structure at all. Figure 4*b* is a two-dimensional schematic diagram of a quartz crystalline structure made up of rings with six silicon atoms. Figure 4*c* is a two-dimensional schematic diagram of an amorphous structure for comparison. In the amorphous

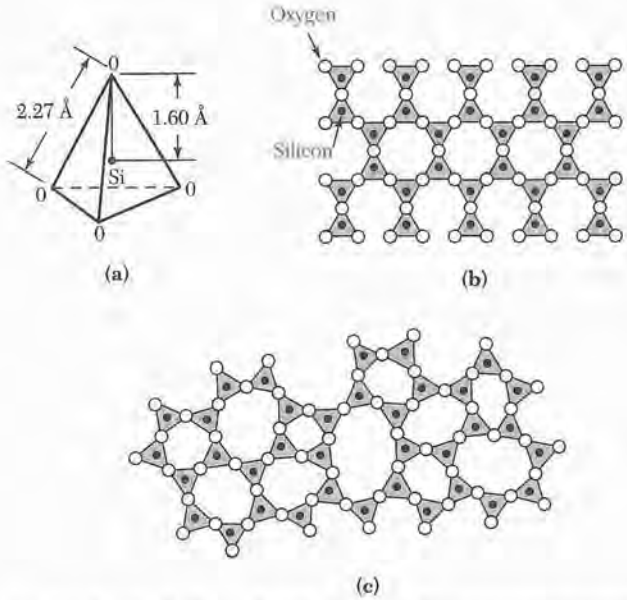


Fig. 4 (*a*) Basic structural unit of silicon dioxide. (*b*) Two dimensional representation of a quartz crystal lattice. (*c*) Two-dimensional representation of the amorphous structure of silicon dioxide.[1]

structure there is still a tendency to form characteristic rings with six silicon atoms. Note that the amorphous structure in Fig. 4c is quite open because only 43% of the space is occupied by silicon dioxide molecules. The relatively open structure accounts for the lower density and allows a variety of impurities (such as sodium) to enter and diffuse readily through the silicon dioxide layer.

The kinetics of thermal oxidation of silicon can be studied based on a simple model illustrated[2] in Fig. 5. A silicon slice contacts the oxidizing species (oxygen or water vapor), resulting in a surface concentration of C_0 molecules/cm^3 for these species. The magnitude of C_0 equals the equilibrium bulk concentration of the species at the oxidation temperature. The equilibrium concentration generally is proportional to the partial pressure of the oxidant adjacent to the oxide surface. At 1000°C and at a pressure of 1 atm, the concentration C_0 is 5.2×10^{16} molecules/cm^3 for dry oxygen and 3×10^{19} molecules/cm^3 for water vapor.

The oxidizing species diffuses through the silicon dioxide layer, resulting in a concentration C_s at the surface of silicon. The flux F_1 can be written as

$$F_1 = D\frac{dC}{dx} \cong \frac{D(C_0 - C_s)}{x}, \tag{3}$$

where D is the diffusion coefficient of the oxidizing species and x is the thickness of the oxide layer already present.

At the silicon surface, the oxidizing species reacts chemically with silicon. Assuming the rate of reaction is proportional to the concentration of the species at the silicon surface, the flux F_2 is given by

$$F_2 = \kappa C_s, \tag{4}$$

where κ is the surface reaction rate constant for oxidation. At the steady state, $F_1 = F_2 = F$. Combining Eqs. 3 and 4 gives

$$F = \frac{DC_0}{x + (D/\kappa)}. \tag{5}$$

The reaction of the oxidizing species with silicon forms silicon dioxide. Let C_1 be the number of molecules of the oxidizing species in a unit volume of the oxide. There are 2.2×10^{22} silicon dioxide molecules/cm^3 in the oxide, and we add one oxygen molecule

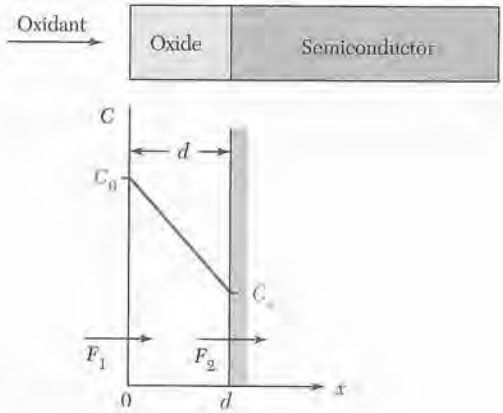


Fig. 5 Basic model for the thermal oxidation of silicon.[2]

(O_2) to each silicon dioxide molecule, whereas we add two water molecules (H_2O) to each silicon dioxide molecule. Therefore, C_1 for oxidation in dry oxygen is 2.2×10^{22} cm^{-3}, and for oxidation in water vapor it is twice this number (4.4×10^{22} cm^{-3}). Thus, the growth rate of the oxide layer thickness is given by

$$\frac{dx}{dt} = \frac{F}{C_1} = \frac{DC_0/C_1}{x + (D/\kappa)}. \tag{6}$$

We can solve this differential equation subject to the initial condition, $x(0) = d_0$, where d_0 is the initial oxide thickness; d_0 can also be regarded as the thickness of oxide layer grown in an earlier oxidation step. Solving Eq. 6 yields the general relationship for the oxidation of silicon,

$$x^2 + \frac{2D}{\kappa}x = \frac{2DC_0}{C_1}(t+\tau), \tag{7}$$

where $\tau \equiv (d_0^2 + 2Dd_0/\kappa)C_1/2DC_0$, which represents a time coordinate shift to account for the initial oxide layer d_0.

The oxide thickness after an oxidizing time t is given by

$$x = \frac{D}{k}\left[\sqrt{1 + \frac{2C_0\kappa^2(t+\tau)}{DC_1}} - 1\right]. \tag{8}$$

For small values of t, Eq. 8 reduces to

$$x \cong \frac{C_0\kappa}{C_1}(t+\tau), \tag{9}$$

and for larger values of t, it reduces to

$$x \cong \sqrt{\frac{2DC_0}{C_1}(t+\tau)}. \tag{10}$$

During the early stages of oxide growth, when surface reaction is the rate-limiting factor, the oxide thickness varies linearly with time. As the oxide layer becomes thicker, the oxidant must diffuse through the oxide layer to react at the silicon-silicon dioxide interface and the reaction becomes diffusion limited. The oxide growth then becomes proportional to the square root of the oxidizing time, which results in a parabolic growth rate.

Equation 7 is often written in a more compact form:

$$\boxed{x^2 + Ax = B(t + \tau),} \tag{11}$$

where $A = 2D/\kappa$, $B = 2DC_0/C_1$ and $B/A = \kappa C_0/C_1$. Using this form, Eqs. 9 and 10 can be written as

$$\boxed{x = \frac{B}{A}(t+\tau)} \tag{12}$$

for the linear region and as

$$x^2 = B(t + \tau). \tag{13}$$

for the parabolic region. For this reason, the term B/A is referred to as the linear rate constant and B is the parabolic rate constant. Experimentally measured results agree with the predictions of this model over a wide range of oxidation conditions. For wet oxidation, the initial oxide thickness d_0 is very small, or $\tau \cong 0$. However, for dry oxidation, the extrapolated value of d_0 at $t = 0$ is about 20 nm.

The temperature dependence of the linear rate constant B/A is shown in Fig. 6 for both dry and wet oxidation and for (111)- and (100)-oriented silicon wafers.[2] The linear rate constant varies as exp $(-E_a/kT)$, where the activation energy E_a is about 2 eV for both dry and wet oxidation. This closely agrees with the energy required to break silicon-silicon bonds, 1.83 eV/molecule. Under a given oxidation condition, the linear rate constant depends on crystal orientation. This is because the rate constant is related to the rate of incorporation of oxygen atoms into the silicon. The rate depends on the surface bond structure of silicon atoms, making it orientation dependent. Because the density of available bonds on the (111)-plane is higher than that on the (100)-plane, the linear rate constant for (111)-silicon is larger.

Figure 7 shows the temperature dependence of the parabolic rate constant B, which can also be described by exp $(-E_a/kT)$. The activation energy E_a is 1.24 eV for dry oxidation. The comparable activation energy for oxygen diffusion in fused silica is 1.18 eV. The corresponding value for wet oxidation, 0.71 eV, compares favorably with the value of 0.79 eV for the activation energy of diffusion of water in fused silica. The parabolic rate constant is independent of crystal orientation. This independence is expected because

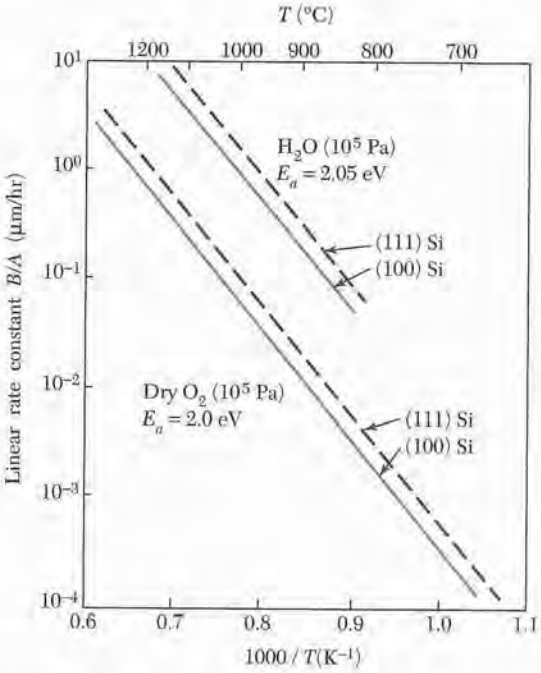


Fig. 6 Linear rate constant versus temperature.[2]

it is a measure of the diffusion process of the oxidizing species through a random network layer of amorphous silica.

Although oxides grown in dry oxygen have the best electrical properties, considerably more time is required to grow the same oxide thickness at a given temperature in dry oxygen than in water vapor. For relatively thin oxides such as the gate oxide in a MOSFET (typically ≤ 20 nm), dry oxidation is used. However, for thicker oxides such as field oxides (≥20 nm) in MOS integrated circuits, and for bipolar devices, oxidation in water vapor (or steam) is used to provide both adequate isolation and passivation.

Figure 8 shows the experimental results of silicon dioxide thickness as a function of reaction time and temperature for two substrate orientations.[3] Under a given oxidation condition, the oxide thickness grown on a (111)-substrate is larger than that grown on a (100)-substrate because of the larger linear rate constant of the (111)-orientation. Note that for a give temperature and time, the oxide film obtained using wet oxidation is about 5–10 times thicker than that using dry oxidation.

11.1.2 Thin Oxide Growth

Relative slow growth rates must be used to reproducibly grow thin oxide films of precise thickness. Various approaches to achieve such slower growth rates have been reported, including growth in dry O_2 at atmospheric pressure and lower temperatures (800°–900°C); growth at pressures lower than atmospheric pressure; growth in a reduced partial pressures of O_2 by using a diluent inert gas, such as N_2, Ar, or He, together with the gas containing the oxidizing species; and the use of composite oxide films with the gate-oxide films consisting of a layer of thermally grown SiO_2 and an overlayer of chemical-vapor

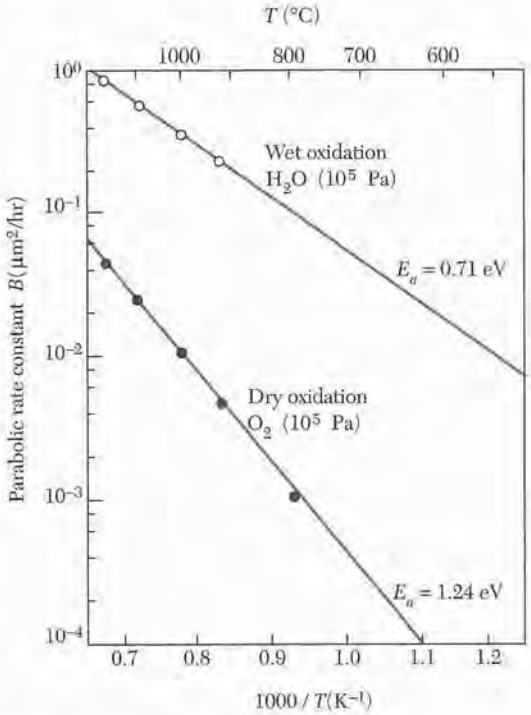


Fig. 7 Parabolic rate constant versus temperature.[2]

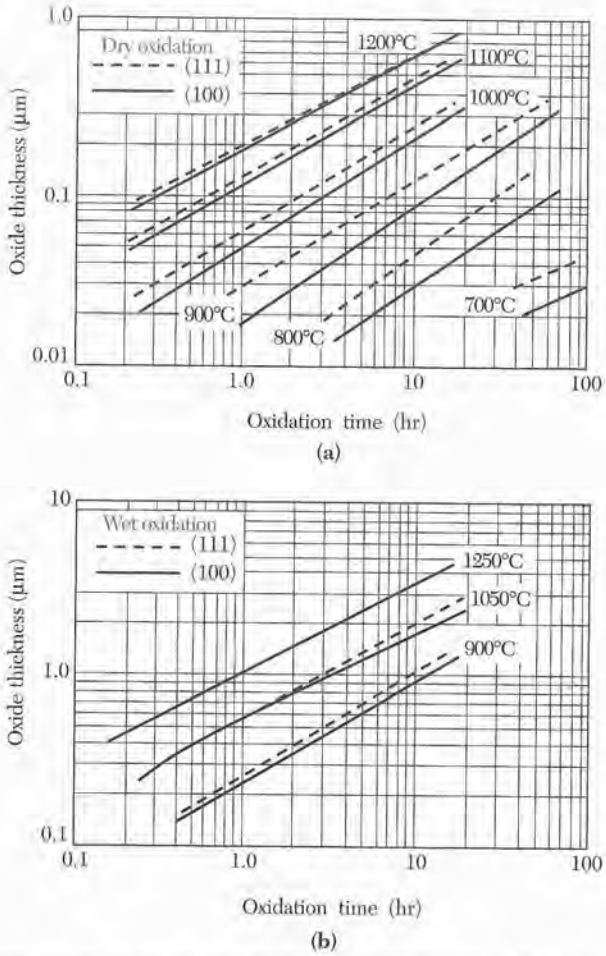


Fig. 8 Experimental results of silicon dioxide thickness as a function of reaction time and temperature for two substrate orientations. (*a*) Growth in dry oxygen. (*b*) Growth in steam.[3]

deposition (CVD) SiO_2. However, the mainstream approach for gate oxides 10–15 nm thick is to grow the oxide film at atmospheric pressure and lower temperatures (800°–900°C). With this approach, processing using modern *vertical* oxidation furnaces can grow reproducible, high-quality 10 nm oxides to within 0.1 nm across the wafer.

We noted earlier that for dry oxidation, there is an apparently rapid oxidation that gives rise to an initial oxide thickness d_0 of about 20 nm. Therefore, the simple model presented in Section 11.1.1 is not valid for dry oxidation with an oxide thickness ≤ 20 nm. For ultralarge-scale integration (ULSI), the ability to grow thin (5 ~ 20 nm), uniform, high-quality reproducible gate oxides has become increasingly important. We briefly consider the growth mechanisms of such thin oxides.

In the early stage of growth in dry oxidation, there is a large compressive stress in the oxide layer that reduces the oxygen diffusion coefficient in the oxide. As the oxide becomes thicker, the stress will be reduced due to the viscous flow of silica and the diffusion coefficient will approach its stress-free value. Therefore, for thin oxides, the value of D/κ may be sufficiently small that we can neglect the term Ax in Eq. 11 and obtain

$$x^2 - d_0^2 = Bt, \tag{14}$$

where d_0 is equal to $\sqrt{2DC_0\tau/C_1}$, which is the initial oxide thickness when time is extrapolated to zero, and B is the parabolic rate constant defined previously. We therefore expect the initial growth in dry oxidation to follow a parabolic form.

11.2 DIELECTRIC DEPOSITION

Deposited dielectric films are used mainly for insulation and passivation of discrete devices and integrated circuits. There are three commonly used deposition methods: atmospheric-pressure CVD, low-pressure CVD (LPCVD), and plasma-enhanced chemical vapor deposition (PECVD, or plasma deposition). PECVD is an energy-enhanced CVD method, in which plasma energy is added to the thermal energy of a conventional CVD system. Considerations in selecting a deposition process are the substrate temperature, the deposition rate and film uniformity, the morphology, the electrical and mechanical properties, and the chemical composition of the dielectric films.

The reactor for atmospheric-pressure CVD is similar to the one shown in Fig. 2, except that different gases are used at the gas inlet. In a hot-wall, reduced-pressure reactor as shown in Fig. 9*a*, the quartz tube is heated by a three-zone furnace, and gas is introduced at one end and pumped out at the opposite end. The semiconductor wafers are held vertically in a slotted quartz boat.[4] The quartz tube wall is hot because it is adjacent to the

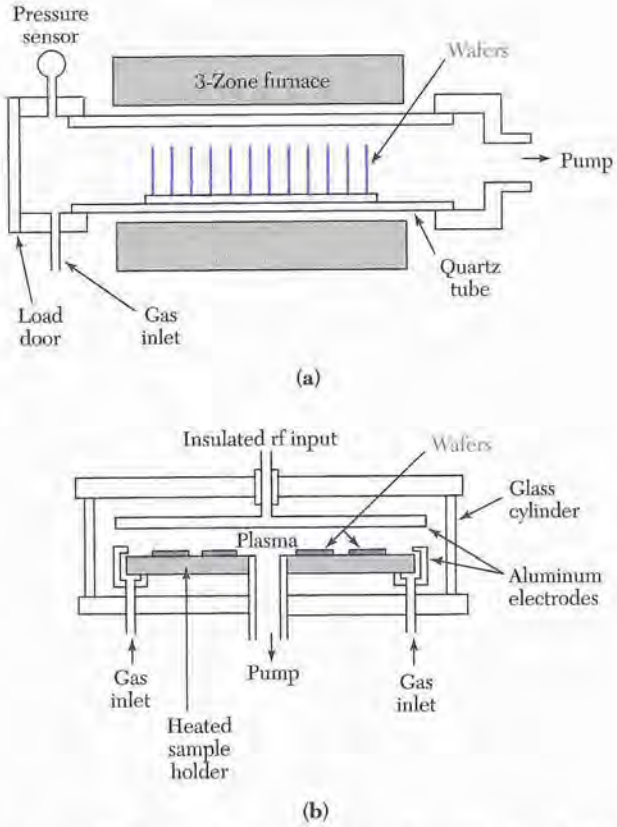


Fig. 9 Schematic diagrams of chemical-vapor deposition reactors. (*a*) Hot-wall, reduced-pressure reactor. (*b*) Parallel-plate plasma deposition reactor.[4] rf, radio frequency.

furnace, in contrast to a cold-wall reactor such as the horizontal epitaxial reactor that uses radio frequency (rf) heating.

The parallel-plate, radial-flow, PECVD reactor shown in Fig. 9*b* consists of a cylindrical glass or aluminum chamber sealed with aluminum endplates. Inside are two parallel aluminum electrodes. An rf voltage is applied to the upper electrode, whereas the lower electrode is grounded. The rf voltage causes a plasma discharge between the electrodes. Wafers are placed on the lower electrode, which is heated between 100° and 400°C by resistance heaters. The reaction gases flow through the discharge from outlets located along the circumference of the lower electrode. The main advantage of this reactor is its low deposition temperature. However, its capacity is limited, especially for large-diameter wafers, and the wafers may become contaminated if loosely adhering deposits fall on them.

11.2.1 Silicon Dioxide

CVD silicon dioxide does not replace thermally grown oxides because the best electrical properties are obtained with thermally grown films. CVD oxides are used instead to complement the thermal oxides. A layer of undoped silicon dioxide is used to insulate multilevel metallization, to mask ion implantation and diffusion, and to increase the thickness of thermally grown field oxides. Phosphorus-doped silicon dioxide is used both as an insulator between metal layers and as a final passivation layer over devices. Oxides doped with phosphorus, arsenic, or boron are used occasionally as diffusion sources.

Deposition Methods

Silicon dioxide films can be deposited by several methods. For low-temperature deposition (300°–500°C), the films are formed by reacting silane, dopant, and oxygen. The chemical reactions for phosphorus-doped oxides are

$$SiH_4 + O_2 \xrightarrow{450°C} SiO_2 + 2H_2, \tag{15}$$

$$4PH_3 + 5O_2 \xrightarrow{450°C} 2P_2O_5 + 6H_2. \tag{16}$$

The deposition process can be performed either at atmospheric pressure in a CVD reactor or at reduced pressure in an LPCVD reactor (Fig. 9*a*). The low deposition temperature of the silane-oxygen reaction makes it a suitable process when films must be deposited over a layer of aluminum.

For intermediate-temperature deposition (500°–800°C), silicon dioxide can be formed by decomposing tetraethylorthosilicate, $Si(OC_2H_5)_4$, in an LPCVD reactor. The compound, abbreviated TEOS, is vaporized from a liquid source. The TEOS compound decomposes as follows:

$$Si(OC_2H_5)_4 \xrightarrow{700°C} SiO_2 + \text{by-products}, \tag{17}$$

forming both SiO_2 and a mixture of organic and organosilicon by-products. Although the higher temperature required for the reaction prevents its use over aluminum, it is suitable for polysilicon gates requiring a uniform insulating layer with good step coverage. The good step coverage is a result of enhanced surface mobility at higher temperatures. The oxides can be doped by adding small amounts of the dopant hydrides (phosphines, arsine, or diborane), similar to the process in epitaxial growth.

The deposition rate as a function of temperature varies as $e^{-E_a/kT}$, where E_a is the activation energy. The E_a of the silane-oxygen reaction is quite low: about 0.6 eV for undoped oxides and almost zero for phosphorus doped oxide. In contrast, E_a for the TEOS reaction

is much higher: about 1.9 eV for undoped oxide and 1.4 eV when phosphorus doping compounds are present. The dependence of the deposition rate on TEOS partial pressure is proportional to $(1 - e^{-P/P_0})$, where P is the TEOS partial pressure and P_0 is about 30 Pa. At low TEOS partial pressures, the deposition rate is determined by the rate of the surface reaction. At high partial pressures, the surface becomes nearly saturated with adsorbed TEOS and the deposition rate becomes essentially independent of TEOS pressure.[4]

Recently, atmospheric-pressure and low-temperature CVD processes using TEOS and ozone (O_3) have been proposed,[5] as shown in Fig. 10. This CVD technology produces oxide films with high conformality and low viscosity under low deposition temperature. In addition, the shrinkage of oxide film during annealing is also a function of ozone concentration, as shown in Fig. 11. Because of their porosity, O_3-TEOS CVD oxides are often accompanied by plasma-assisted oxides to permit planarization in ULSI processing.

For high-temperature deposition (900°C), silicon dioxide is formed by reacting dichlorosilane, $SiCl_2H_2$, with nitrous oxide at reduced pressure:

$$SiCl_2H_2 + 2N_2O \xrightarrow{900°C} SiO_2 + 2N_2 + 2HCl. \tag{18}$$

This deposition gives excellent film uniformity and is sometimes used to deposit insulating layers over polysilicon.

Properties of Silicon Dioxide

Deposition methods and properties of silicon dioxide films are listed[4] in Table 1. In general, there is a direct correlation between deposition temperature and film quality. At higher temperatures, deposited oxide films are structurally similar to silicon dioxide that has been thermally grown.

The lower densities occur in films deposited below 500°C. Heating deposited silicon dioxide at temperatures between 600° and 1000°C causes densification, during which the oxide thickness decreases, whereas the density increases to 2.2 g/cm^3. The refractive index of silicon dioxide is 1.46 at a wavelength of 0.6328 μm. Oxides with lower indices are porous, such as the oxide from the silane-oxygen deposition, which has a refractive index of 1.44. The porous nature of the oxide also is responsible for the lower dielectric strength, which is the applied electric field that will cause a high current to flow in the oxide film. The etch rates of oxides in a hydrofluoric acid solution depend on deposition

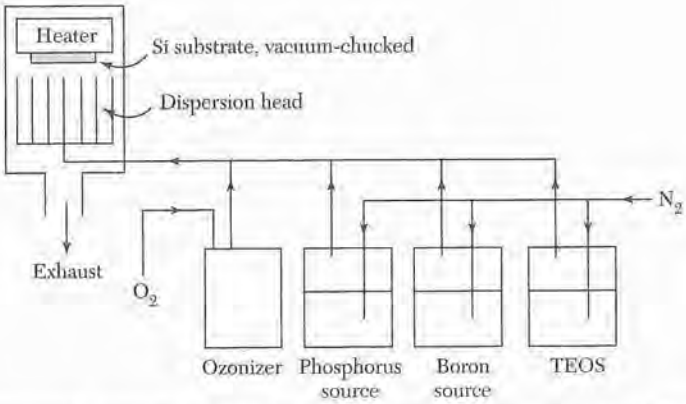


Fig. 10 Experimental apparatus for the O_3–TEOS chemical-vapor deposition (CVD) system.

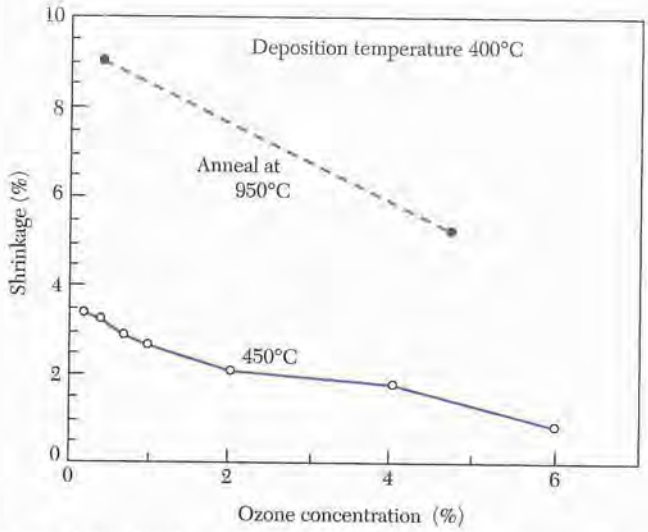


Fig. 11 Dependence of the shrinkage of the O_3–TEOS CVD film on ozone concentration using annealing. (Courtesy of SAMCO Company, Japan.)

temperature, annealing history, and dopant concentration. Usually higher-quality oxides are etched at lower rates.

Step Coverage

Step coverage relates the surface topography of a deposited film to the various steps on the semiconductor substrate. In the illustration of ideal, or conformal, step coverage shown in Fig. 12*a*, film thickness is uniform along all surfaces of the step. The uniformity of the film thickness, regardless of topography, is due to the rapid migration of reactants after adsorption on the step surfaces.[6]

Figure 12*b* shows an example of nonconformal step coverage, which results when the reactants adsorb and react without significant surface migration. In this instance, the deposition rate is proportional to the arrival angle of the gas molecules. Reactants arriving along the top horizontal surface come from many different angles and ϕ_1, the arrival

TABLE 1 Properties of SiO_2 Films

Property	Thermally grown at 1000°C	$SiH_4 + O_2$ at 450°C	TEOS at 700°C	$SiCl_2H_2 + N_2O$ at 900°C
Composition	SiO_2	SiO_2 (H)	SiO_2	SiO_2(Cl)
Density (g/cm^3)	2.2	2.1	2.2	2.2
Refractive index	1.46	1.44	1.46	1.46
Dielectric strength (10^6 V/cm)	>10	8	10	10
Etch rate (Å /min) (100:1 H_2O:HF)	30	60	30	30
Etch rate (Å /min) (buffered HF)	440	1200	450	450
Step coverage	—	Nonconformal	Conformal	Conformal

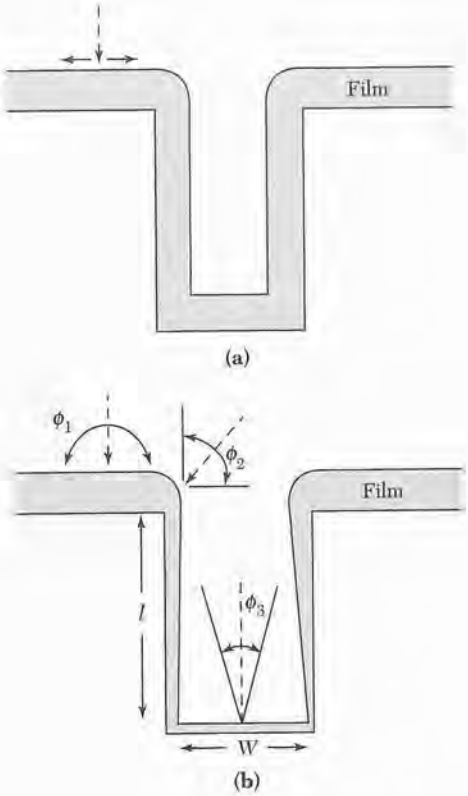


Fig. 12 Step coverage of deposited films. (*a*) Conformal step coverage. (*b*) Nonconformal step coverage.[4]

angle, varies in two dimensions, from 0 to 180°, whereas reactants arriving at the top of a vertical wall have an arrival angle ϕ_2 that varies from 0° to 90°. Thus, the film thickness on the top surface is double that of a wall surface. Further down the wall, ϕ_3 is related to the width of the opening, and the film thickness is proportional to

$$\phi_3 \cong \arctan \frac{W}{l}, \tag{19}$$

where l is the distance from the top surface and W is the width of the opening. This type of step coverage is thin along the vertical walls, with a possible crack at the bottom of step caused by self-shadowing.

Silicon dioxide formed by TEOS decomposition at reduced pressure gives a nearly conformal coverage due to rapid surface migration. Similarly, the high-temperature dichlorosilane-nitrous oxide reaction also results in conformal coverage. However, during silane-oxygen deposition, no surface migration takes place and the step coverage is determined by the arrival angle. Most evaporated or sputtered materials have a step coverage similar to that in Fig. 12*b*.

P-Glass Flow

A smooth topography is usually required for the deposited silicon dioxide used as an insulator between metal layers. If the oxide used to cover the lower metal layer is concave, circuit failure may result from an opening that may occur in the upper metal layer dur-

ing deposition. Because phosphorus-doped silicon dioxide (P-glass) deposited at low temperatures becomes soft and flows upon heating, it provides a smooth surface and is often used to insulate adjacent metal layers. This process is called P-glass flow.

Figure 13 shows four cross sections of scanning electron microscope photographs of P-glass covering a polysilicon step.[6] All samples are heated in steam at 1100°C for 20 min. Figure 13*a* shows a sample of glass that contains a negligibly small amount of phosphorus and does not flow. Note the concavity of the film and that the corresponding angle θ is about 120°. Figures 13*b*, 13*c*, and 13*d* show samples of P-glass with progressively higher phosphorus contents up to 7.2 wt% (weight percent). In these samples the decreasing step angles of the P-glass layer indicate how flow increases with phosphorus concentration. P-glass flow depends on annealing time, temperature, phosphorus concentration, and the annealing ambient.[6]

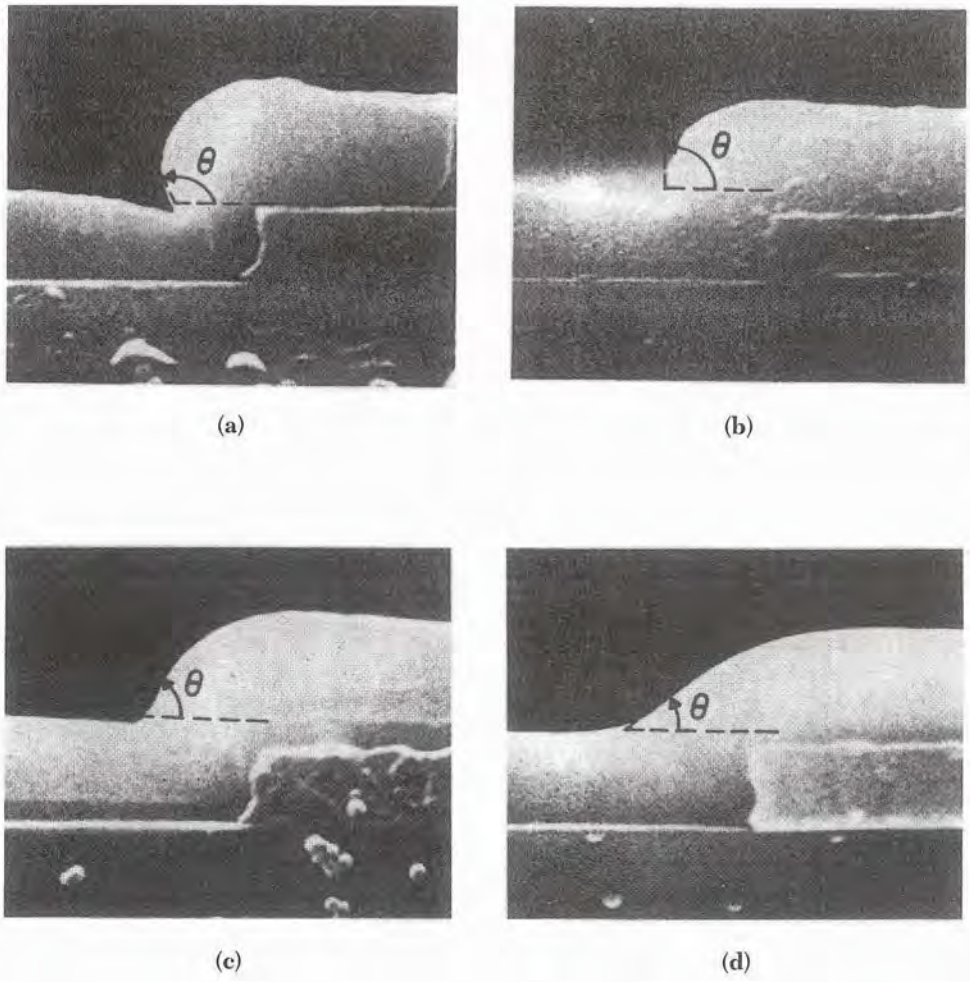


Fig. 13 Scanning-electron micrographs (10,000×) of samples annealed in steam at 1100°C for 20 minutes for the following weight percent of phosphorus.[6] (*a*) 0 wt%; (*b*) 2.2 wt%; (*c*) 4.6 wt%; and (*d*) 7.2 wt%.

The angle θ as a function of weight percent of phosphorus as shown in Fig. 13 can be approximated by

$$\theta \cong 120^\circ\left(\frac{10 - \text{wt\%}}{10}\right). \tag{20}$$

If we want an angle smaller than 45° we require a phosphorus concentration larger than 6 wt%. However, at concentrations above 8 wt% , the metal film (e.g., aluminum) may be corroded by the acid products formed during the reaction between the phosphorus in the oxide and atmospheric moisture. Therefore, the P-glass flow process uses phosphorus concentrations of 6–8 wt%.

11.2.2 Silicon Nitride

It is difficult to grow silicon nitride by thermal nitridation (e.g., with ammonia, NH_3) because of its low growth rate and high growth temperature. However, silicon nitride films can be deposited by an intermediate-temperature (750°C) LPCVD process or a low-temperature (300°C) plasma-assisted CVD process.[7,8] The LPCVD films are of stoichiometric composition (Si_3N_4) with high density (2.9–3.1 g/cm^3). These films can be used to passivate devices because they serve as good barriers to the diffusion of water and sodium. The films also can be used as masks for the selective oxidation of silicon because silicon nitride oxidizes very slowly and prevents the underlying silicon from oxidizing. The films deposited by plasma-assisted CVD are not stoichiometric and have a lower density (2.4–2.8 g/cm^3). Because of the low deposition temperature, silicon nitride films can be deposited over fabricated devices and serve as their final passivation. The plasma-deposited nitride provides excellent scratch protection, serves as a moisture barrier, and prevents sodium diffusion.

In the LPCVD process, dichlorosilane and ammonia react at reduced pressure to deposit silicon nitride at temperatures between 700° and 800°C. The reaction is

$$3SiCl_2H_2 + 4NH_3 \xrightarrow{\sim 750^\circ C} Si_3N_4 + 6HCl + 6H_2. \tag{21}$$

Good film uniformity and high wafer throughout (the number of wafers processed per hour) are advantages of the reduced-pressure process. As in the case of oxide deposition, silicon nitride deposition is controlled by temperature, pressure, and reactant concentration. The activation energy for deposition is about 1.8 eV. The deposition rate increases with increasing total pressure or dichlorosilane partial pressure and decreases with an increasing ammonia-to-dichlorosilane ratio.

Silicon nitride deposited by LPCVD is an amorphous dielectric containing up to 8 atomic percent (at %) hydrogen. The etch rate in buffered HF is less than 1 nm/min. The film has a very high tensile stress of approximately 10^{10} dynes/cm^2, which is nearly 10 times that of TEOS-deposited SiO_2. Films thicker than 200 nm may crack because of the very high stress. The resistivity of silicon nitride at room temperature is about 10^{16} Ω-cm. Its dielectric constant is 7 and its dielectric strength is 10^7 V/cm.

In the plasma-assisted CVD process, silicon nitride is formed either by reacting silane and ammonia in an argon plasma or by reacting silane in a nitrogen discharge. The reactions are as follows:

$$SiH_4 + NH_3 \xrightarrow{300^\circ C} SiNH + 3H_2, \tag{22a}$$

$$2SiH_4 + N_2 \xrightarrow{300^\circ C} 2SiNH + 3H_2. \tag{22b}$$

The products depend strongly on deposition conditions. The radial-flow, parallel-plate reactor (Fig. 9*b*) is used to deposit the films. The deposition rate generally increases with increasing temperature, power input, and reactant gas pressure.

Large concentrations of hydrogen are contained in plasma-deposited films. The plasma nitride (also referred to as SiN) used in semiconductor processing generally contains 20–25 at % hydrogen. Films with low tensile stress ($\sim 2 \times 10^9$ dynes/cm^2) can be prepared by plasma deposition. Film resistivities range from 10^5 to 10^{21} Ω-cm, depending on silicon-to-nitrogen ratio, whereas dielectric strengths are between 1×10^6 and 6×10^6 V/cm.

11.2.3 Low-Dielectric–Constant Materials

As devices continue to scale down to the deep submicron region, they require multilevel interconnection architecture to minimize the time delay due to parasitic resistance (R) and capacitance (C). The gain in device speed at the gate level will be offset by the propagation delay at the metal interconnects because of the increased RC time constant, as shown in Fig. 14. For example, in devices with gate length of 250 nm or less, up to 50% of the time delay is due to the RC delay of long interconnections.[9] Therefore, the device interconnection network becomes a limiting factor in determining chip performance such as device speed, cross talk, and power consumption of ULSI circuits.

In order to reduce the RC time constant of ULSI circuits, interconnection materials with low resistivity and interlayer films with low capacitance are required. On the low-capacitance topic ($C = \varepsilon_i A/d$, where ε_i is the dielecric permittivity, A the area, and d the

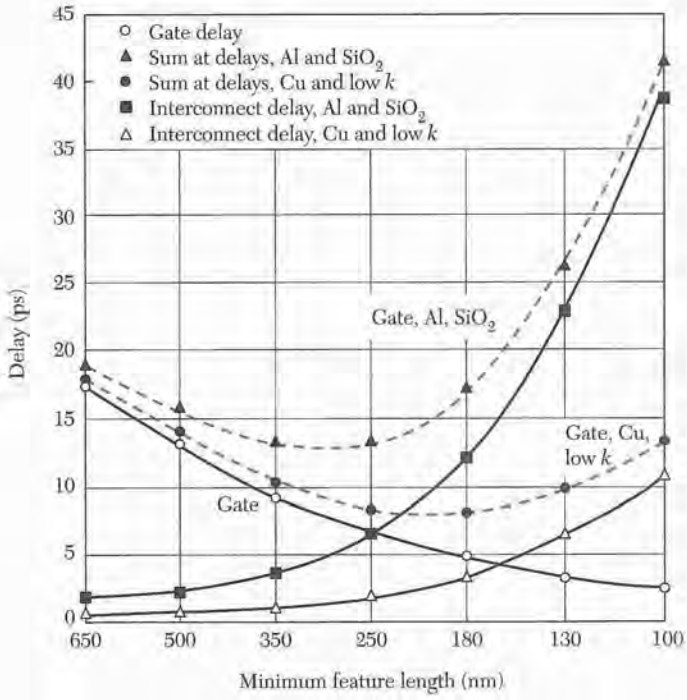


Fig. 14 Calculated gate and interconnect delay versus technology generation. The dielectric constant for the low-k material is 2.0. Both Al and Cu interconnects are 0.8 μm thick and 43 μm long.

thickness of the dielectric film), it is not easy to lower the parasitic capacitance by increasing thickness d of the interlayer dielectric (which makes gap filling more difficult), or decreasing wiring height and area A (which results in the increase of interconnect resistance). Therefore, materials with low dielectric constant (low-k) are required. The ε_i is equal to the multiplication of k and ε_0, where k and ε_0 are the dielectric constant and the vacuum permittivity, respectively.

Material Options

The properties of the interlayer dielectric film and how they are formed have to meet the following requirements: low dielectric constant, low residual stress, high planarization capability, high capability for gap filling, low deposition temperature, simplicity of process, and ease of integration.

A substantial number of low-k materials have been synthesized for the intermetal dielectric in ULSI circuits. Some of the promising low-k materials are shown in Table 2. These materials can be either inorganic or organic and can be deposited by either CVD or spin-on techniques.[9]

EXAMPLE 2

Estimate the intrinsic RC value of two parallel Al wires 0.5 μm × 0.5 μm in cross section, 1 mm in length, and separated by a polyimide (k~2.7) dielectric layer that is 0.5 μm thick. The resistivity of Al is 2.7 μΩ–cm.

SOLUTION

$$RC = \left(\rho \frac{\ell}{t_m^2}\right) \times \left(\varepsilon_i \frac{t_m \times \ell}{\text{spacing width}}\right) = \left(2.7 \times 10^{-6} \times \frac{1 \times 10^{-1}}{0.25 \times 10^{-8}}\right) \times$$

$$\left(8.85 \times 10^{-14} \times 2.7 \times \frac{0.5 \times 10^{-4} \times 10^{-1}}{0.5 \times 10^{-4}}\right) = 2.57 \text{ ps.}$$

TABLE 2 Low-k Materials

Determinant	Materials	Dielectric constant
Vapor-phase deposition polymers	Fluorosilicate glass (FSG)	3.5–4.0
	Parylene N	2.6
	Parylene F	2.4–2.5
	Black diamond (C-doped oxide)	2.7–3.0
	Fluorinated hydrocarbon	2.0–2.4
	Teflon-AF	1.93
Spin-on polymers	HSQ/MSQ	2.8–3.0
	Polyimide	2.7–2.9
	SiLK (aromatic hydrocarbon polymer)	2.7
	PAE [poly(arylene ethers)]	2.6
	Fluorinated amorphous carbon	2.1
	Xerogels (porous silica)	1.1–2.0

11.2.4 High-Dielectric–Constant Materials

High-k materials are required for ULSI circuits, especially for dynamic random access memory (DRAM). The storage capacitor in a DRAM has to maintain a certain value of capacitance for proper operation (e.g., 40 fF). For a given capacitance ($\varepsilon_i A/d$), usually a minimum d is selected to meet the conditions of the maximum allowed leakage current and the minimum required breakdown voltage. The area of the capacitor can be increased by using stacked or trench structures. These structures are considered in Chapter 14. However, for a planar structure, area A is reduced with increasing DRAM density. Therefore, the dielectric constant of the film must be increased.

Several high-k materials have been proposed, such as barium strontium titanate (BST) and lead zirconium titanate (PZT). They are shown in Table 3. In addition, there are titanates doped with one or more acceptors, such as alkaline earth metals, or doped with one or more donors, such as rare earth elements. The tantalum oxide (Ta_2O_5) has a dielectric constant in a range of 20–30. As a reference, the dielectric constant of Si_3N_4 is in a range of 6–7 and that for SiO_2 is 3.9. A Ta_2O_5 film can be deposited by a CVD process using gaseous $TaCl_5$ and O_5 as the starting materials.

EXAMPLE 3

A DRAM capacitor has the following parameters: capacitance C = 40 fF, cell size A = 1.28 μm², and dielectric constant k = 3.9 for silicon dioxide. If we replace SiO_2 with Ta_2O_5 (k = 25) without changing thickness, what is the equivalent cell area of the capacitor?

SOLUTION

$$C = \frac{\varepsilon_i A}{d},$$

$$\frac{3.9 \times 1.28}{d} = \frac{25 \times A}{d},$$

$$\therefore \text{Equivalent cell size } A = \frac{3.9}{25} \times 1.28 = 0.2\ \mu\text{m}^2.$$

TABLE 3 High-k Materials

	Materials	Dielectric constant
Binary	Ta_2O_5	25
	TiO_2	40
	Y_2O_3	17
	Si_3N_4	7
Paraelectric perovskite	$SrTiO_3$ (STO)	140
	$(Ba_{1-x}Sr_x)\ TiO_3$(BST)	300–500
	$Ba(Ti_{1-x}Zr_x)O_3$(BZT)	300
	$(Pb_{1-x}La_x)(Zr_{1-y}Ti_y)O_3$(PLZT)	800–1000
	$Pb(Mg_{1/3}Nb_{2/3})O_3$(PMN)	1000–2000
Ferroelectric perovskite	$Pb(Zr_{0.47}Ti_{0.53})O_3$(PZT)	>1000

11.3 POLYSILICON DEPOSITION

Using polysilicon as the gate electrode in MOS devices is a significant development in MOS technology. One important reason is that polysilicon surpasses aluminum for electrode reliability. Figure 15 shows the maximum time to breakdown for capacitors with both polysilicon and aluminum electrodes.[10] The polysilicon is clearly superior, especially for thinner gate oxides. The inferior time to breakdown of aluminum electrode is due to the migration of aluminum atoms into the thin oxide under an electrical field. Polysilicon is also used as a diffusion source to create shallow junctions and to ensure ohmic contact to crystalline silicon. Additional uses include the manufacture of conductors and high-value resistors.

A low-pressure reactor (Fig. 9*a*) operated between 600° and 650°C is used to deposit polysilicon by pyrolyzing silane according to the following reaction.

$$SiH_4 \xrightarrow{600°C} Si + 2H_2. \tag{23}$$

Of the two most common low-pressure processes one operates at a pressure of 25–130 Pa using 100% silane, whereas the other process involves a diluted mixture of 20%–30% silane in nitrogen at the same total pressure. Both processes can deposit polysilicon on hundreds of wafers per run with good uniformity (i.e., thickness within 5%)

Figure 16 shows the deposition rate at four deposition temperatures. At low silane partial pressure, the deposition rate is proportional to the silane pressure.[4] At higher silane

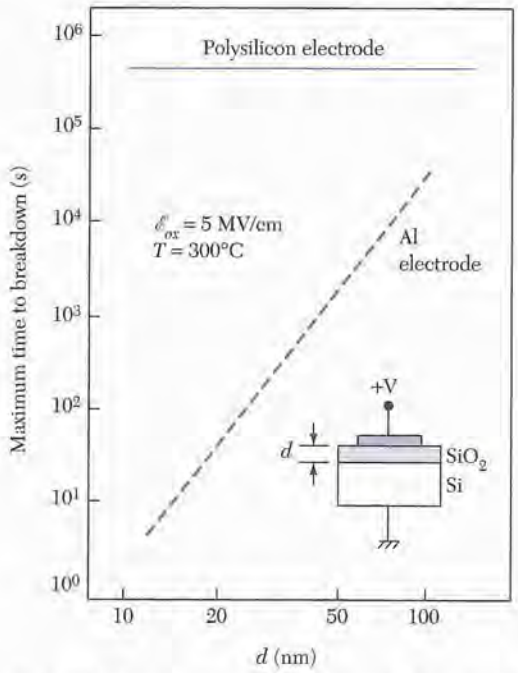


Fig. 15 Maximum time to breakdown versus oxide thickness for a polysilicon electrode and an aluminum electrode.[10]

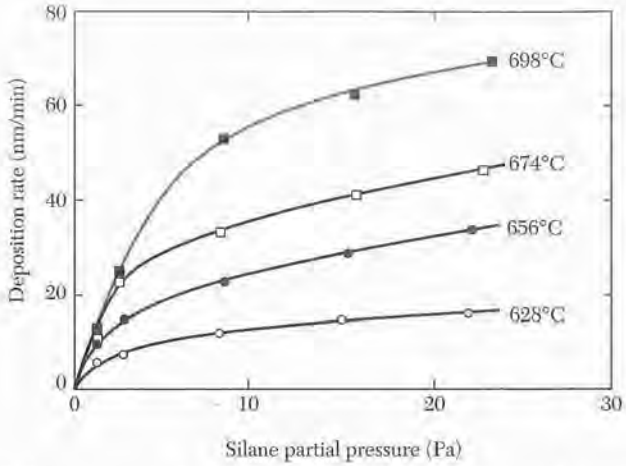


Fig. 16 Effect of silane concentration on the polysilicon deposition rate.[4]

concentrations, saturation of the deposition rate occurs. Deposition at reduced pressure is generally limited to temperatures between 600° and 650°C. In this temperature range, the deposition rate varies as exp $(-E_a/kT)$, where the activation energy E_a is 1.7 eV, which is essentially independent of the total pressure in the reactor. At higher temperatures, gas-phase reactions that result in a rough, loosely adhering deposit become significant and silane depletion will occur, causing poor uniformity. At temperatures much lower than 600°C, the deposition rate is too slow to be practical.

Process parameters that affect the polysilicon structure are deposition temperature, dopants, and the heat cycle applied following the deposition step. A columnar structure results when polysilicon is deposited at a temperature of 600°–650°C. This structure is comprised of polycrystalline grains ranging in size from 0.03 to 0.3 μm at a preferred orientation of (110). When phosphorus is diffused at 950°C, the structure changes to crystallite and grain size increases to a size between 0.5 and 1.0 μm. When temperature is increased to 1050°C during oxidation, the grains reach a final size of 1–3 μm. Although the initially deposited film appears amorphous when deposition occurs below 600°C, growth characteristics similar to the polycrystalline-grain columnar structure are observed after doping and heating.

Polysilicon can be doped by diffusion, ion implantation, or the addition of dopant gases during deposition, referred to as in situ doping. The implantation method is most commonly used because of its lower processing temperatures. Figure 17 shows the sheet resistance of single crystal silicon and of 500 nm polysilicon doped with phosphorus and antimony using ion implantation.[11] The ion implantation process is considered in Chapter 13. Implant dose, annealing temperature, and annealing time all influence the sheet resistance of implanted polysilicon. Carrier traps at the grain boundaries cause a very high resistance in the lightly implanted polysilicon. As Fig. 17 illustrates, resistance drops rapidly, approaching that of implanted single crystal silicon, as the carrier traps become saturated with dopants.

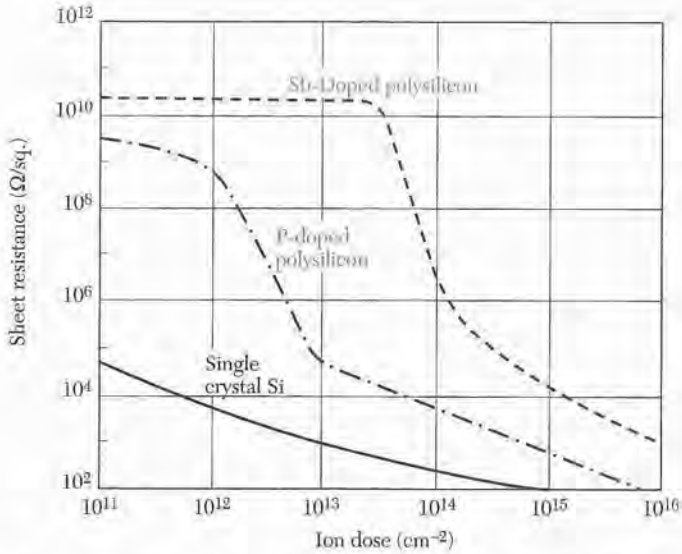


Fig. 17 Sheet resistance versus ion dose into 500 nm polysilicon at 30 keV.[11]

11.4 METALLIZATION

11.4.1 Physical-Vapor Deposition

The most common methods of physical-vapor deposition (PVD) of metals are evaporation, e-beam evaporation, plasma spray deposition, and sputtering. Metals and metal compounds such as Ti, Al, Cu, TiN, and TaN can be deposited by PVD. Evaporation occurs when a source material is heated above its melting point in an evacuated chamber. The evaporated atoms then travel at high velocity in straight-line trajectories. The source can be molten by resistance heating, by rf heating, or with a focused electron beam. Evaporation and e-beam evaporation were used extensively in earlier generations of integrated circuits, but they have been replaced by sputtering for ULSI circuits.

In ion-beam sputtering, a source of ions is accelerated toward the target and impinged on its surface. Figure 18*a* shows the standard sputtering system. The sputtered material deposits on a wafer that is placed facing the target. The ion current and energy can be independently adjusted. Since the target and wafer are placed in a chamber that has lower pressure, more target material and less contamination are transferred to the wafer.

One method to increase the ion density and, hence, the sputter-deposition rate is to use a third electrode that provides more electrons for ionization. Another method is to use a magnetic field, such as electron cyclotron resonance (ECR), to capture and spiral electrons, increasing their ionizing efficiency in the vicinity of the sputtering target. This technique, referred to as magnetron sputtering, has found widespread applications for the deposition of aluminum and its alloys at a rate that can approach 1 μm/min.

Long-throw sputtering is another technique used to control the angular distribution. Figure 18*b* shows the long throw-sputtering system. In standard sputtering configurations there are two primary reasons for a wide angular distribution of incident flux at the surface: the use of a small target to substrate separation, d_{ts}, and scattering of the flux by the working gas as the flux travels from the target to the substrate. These two factors are linked because a small d_{ts} is needed to achieve good throughput, uniformity, and film properties when there is substantial gas scattering. A solution to this problem is to sputter at very low pressures, a capability that has been developed using a variety of sys-

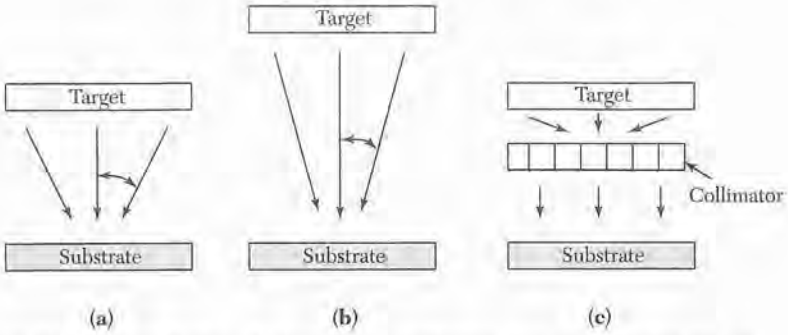


Fig. 18 (*a*) Standard sputtering, (*b*) long-throw sputtering, and (*c*) sputtering with a collimator.

tems, which can sustain the magnetron plasma under more rarefied conditions. These systems allow for sputtering at working pressures of less than 0.1 Pa. At these pressures, gas scattering is less important, and the target-to-substrate distance can be greatly increased. From a simple geometrical argument, this allows the angular distribution to be greatly narrowed, which permits more deposition at the bottom of high-aspect features such as contact holes.

Contact holes with large aspect ratio are difficult to fill with material, mainly because scattering events cause the top opening of the hole to seal before appreciable material has deposited on its floor. This problem can be overcome by collimating the sputtered atoms by placing an array of collimating tubes just above the wafer to restrict the depositing flux to normal ±5°. Sputtering with a collimator is shown in Fig. 18*c*. Atoms whose trajectory is more than 5° from normal are deposited on the inner surface of the collimators.

11.4.2 Chemical-Vapor Deposition

CVD is attractive for metallization because it offers coatings that are conformal, has good step coverage, and can coat a large number of wafers at a time. The basic CVD setup is the same as that used for deposition of dielectrics and polysilicon (see Fig. 9*a*). Low-pressure CVD (LPCVD) is capable of producing conformal step coverage over a wide range of topographical profiles, often with lower electrical resistivity than that from PVD.

One of the major new applications of CVD metal deposition for integrated circuit production is in the area of refractory-metal deposition. For example, tungsten's low electrical resistivity (5.3 μΩ-cm) and its refractory nature make it a desirable metal for use in integrated circuit fabrication.

CVD-W

Tungsten is used both as a contact plug and as a first-level metal. Tungsten can be deposited by using WF_6 as the W source gas, since it is a liquid that boils at room temperature. WF_6 can be reduced by silicon, hydrogen, or silane. The basic chemistry for CVD-W is as follows:

$$WF_6 + 3H_2 \rightarrow W + 6HF \text{ (hydrogen reduction)}, \tag{24}$$

$$2WF_6 + 3Si \rightarrow 2W + 3SiF_4 \text{ (silicon reduction)}, \tag{25}$$

$$2WF_6 + 3SiH_4 \rightarrow 2W + 3SiF_4 + 6H_2 \text{ (silane reduction)}. \tag{26}$$

On a Si contact, the selective process starts from a silicon reduction process. This process provides a nucleation layer of W grown on Si but not on SiO_2. The hydrogen reduction process can deposit W rapidly on the nucleation layer, forming the plug. The hydrogen reduction process provides excellent conformal coverage of the topography. This process, however, does not have perfect selectivity, and the HF gas by-product of the reaction is responsible for the encroachment of the oxide, as well as for the rough surface of deposited W films.

The silane reduction process gives a high deposition rate and much smaller W grain size than that obtained with the hydrogen reduction process. In addition, the problems of encroachment and rough W surface are eliminated because there is no HF by-product generation. Usually, a silane reduction process is used as the first step in blanket W deposition to serve as a nucleation layer and to reduce junction damage. After the silane reduction, hydrogen reduction is used to grow the blanket W layer.

CVD TiN

TiN is widely used as a diffusion barrier-metal layer in metallization and can be deposited by sputtering from a compound target or by CVD. The CVD TiN can provide better step coverage than PVD methods in deep submicron technology. CVD TiN can be deposited,[12–14] using $TiCl_4$ with NH_3, H_2/N_2, or NH_3/H_2:

$$6TiCl_4 + 8NH_3 \rightarrow 6\ TiN + 24HCl + N_2, \tag{27}$$

$$2TiCl_4 + N_2 + 4H_2 \rightarrow 2TiN + 8HCl, \tag{28}$$

$$2TiCl_4 + 2NH_3 + H_2 \rightarrow 2\ TiN + 8HCl. \tag{29}$$

The deposition temperature is about 400°–700°C for NH_3 reduction and is higher than 700°C for the N_2/H_2 reaction. The higher the deposition temperature, the better the TiN film and the less Cl incorporated in TiN (~5%).

11.4.3 Aluminum Metallization

Aluminum and its alloys are used extensively for metallization in integrated circuits. The Al film can be deposited by a PVD or CVD method. Because aluminum and its alloys have low resistivities, 2.7 μΩ-cm for Al and up to 3.5 μΩ-cm for its alloys, these metals satisfy the low-resistance requirements. Aluminum also adheres well to silicon dioxide. However, the use of aluminum in integrated circuits with shallow junctions often creates problems, such as spiking and eletromigration. We consider the problems of aluminum metallization and their solutions in this section.

Junction Spiking

Figure 19 shows the phase diagram of the Al-Si system at 1 atm.[15] The phase diagram relates these two components as a function of temperature. The Al-Si system exhibits eutectic characteristics; that is, the addition of either component lowers the system's melting point below that of either metal. Here, the minimum melting temperature, called eutectic temperature, is 577°C, corresponding to a 11.3% Si and 88.7% Al composition. The melting points of pure aluminum and pure silicon are 660°C and 1412°C, respectively. Because of the eutectic characteristics, during aluminum deposition the temperature on the silicon substrate must be limited to less than 577°C.

The inset of Fig. 19 also shows the solid solubility of silicon in aluminum. For example, the solubility of silicon in aluminum is 0.25 wt% at 400°C, 0.5 wt% at 450°C, and 0.8 wt% at 500°C. Therefore, wherever aluminum contacts silicon, the silicon will dis-

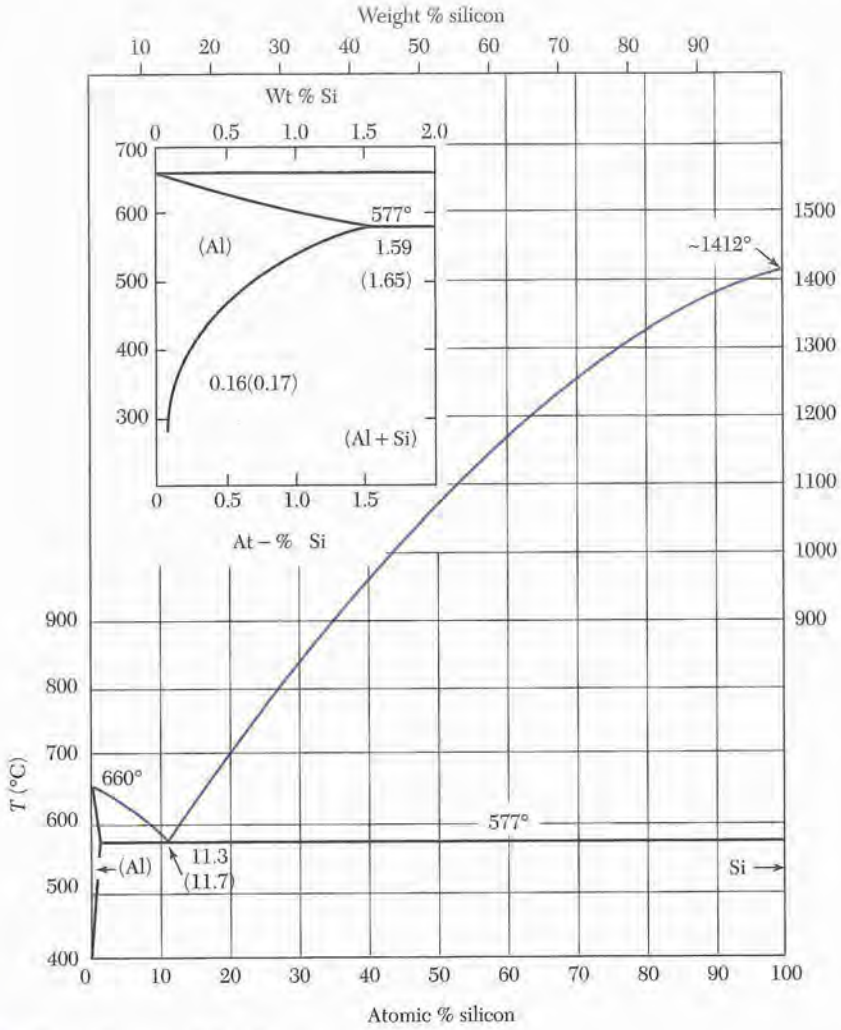


Fig. 19 Phase diagram of the aluminum–silicon system.[15]

solve into the aluminum during annealing. The amount of silicon dissolved will depend not only on the solubility at the annealing temperature but also on the volume of aluminum to be saturated with silicon. Consider a long aluminum metal line in contact with an area ZL of silicon as shown in Fig. 20. After an annealing time t, the silicon will diffuse a distance of approximately $\sqrt{Dt}$ along the aluminum line from the edge of the contact, where D is the diffusion coefficient given by $4 \times 10^{-2} \exp(-0.92/kT)$ for silicon diffusion in deposited aluminum films. Assuming that this length of aluminum is completely saturated with silicon, the volume of silicon consumed is then

$$\mathrm{Vol} \cong 2\sqrt{Dt}\,(HZ)S\left(\frac{\rho_{\mathrm{Al}}}{\rho_{\mathrm{Si}}}\right), \tag{30}$$

where ρ_{Al} and ρ_{Si} are the densities of aluminum and silicon, respectively, and S is the solubility of silicon in aluminum at the annealing temperature.[16] If the consumption takes

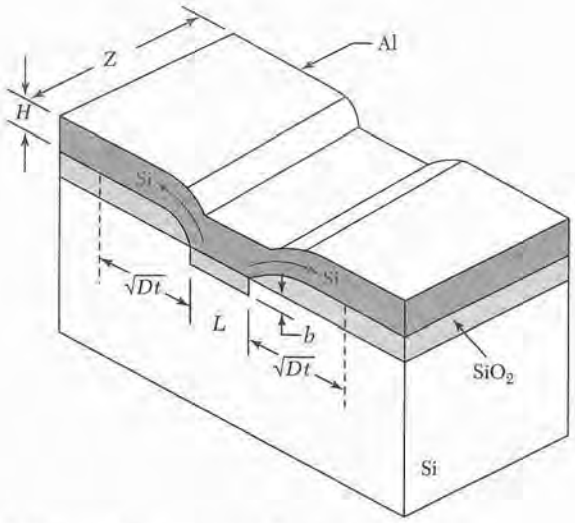


Fig. 20 Diffusion of silicon in aluminum metallization.[16]

place uniformly over the contact area A (where $A = ZL$ for uniform dissolution), the depth to which silicon would be consumed is

$$b \cong 2\sqrt{Dt}\left(\frac{HZ}{A}\right)S\left(\frac{\rho_{Al}}{\rho_{Si}}\right). \tag{31}$$

EXAMPLE 4

For $T = 500°C$, $t = 30$ min, $ZL = 16\ \mu m^2$, $Z = 5\ \mu m$, and $H = 1\ \mu m$. Find the depth b, assuming uniform dissolution.

SOLUTION The diffusion coefficient of silicon in aluminum at 500°C is about $2 \times 10^{-8}\ cm^2/s$; thus, $\sqrt{Dt}$ is 60 μm. The density ratio is 2.7/2.33 = 1.16.

At 500°C, S is 0.8 wt%. From Eq. 31 we have

$$b = 2 \times 60\left(\frac{1 \times 5}{16}\right)0.8\% \times 1.16 = 0.35\ \mu m.$$

Aluminum will fill a depth of $b = 0.35$ μm from which silicon is consumed. If at the contact point there is a shallow junction whose depth is less than b, the diffusion of silicon into aluminum can short-circuit the junction. ◀

In a practical situation, the dissolution of silicon does not take place uniformly but rather at only a few points. The effective area in Eq. 31 is less than the actual contact area; hence b is much larger. Figure 21 illustrates the actual situation in the p–n junction area of aluminum penetrating the silicon at only the few points where spikes are formed. One way to minimize aluminum spiking is to add silicon to the aluminum by coevaporation until the amount of silicon contained by the alloy satisfies the solubility requirement. Another method is to introduce a barrier metal layer between the aluminum and the silicon substrate (Fig. 22). This barrier metal layer must meet the following requirements: it forms low contact resistance with silicon, it will not react with aluminum, and

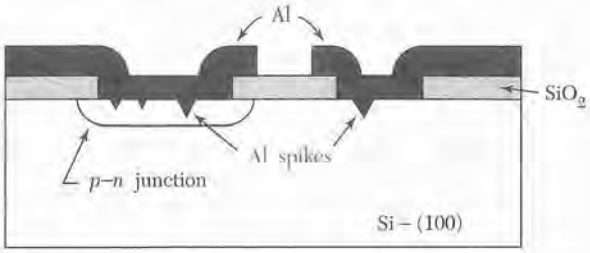


Fig. 21 Schematic view of aluminum films contacting silicon. Note the aluminum spiking in the silicon.

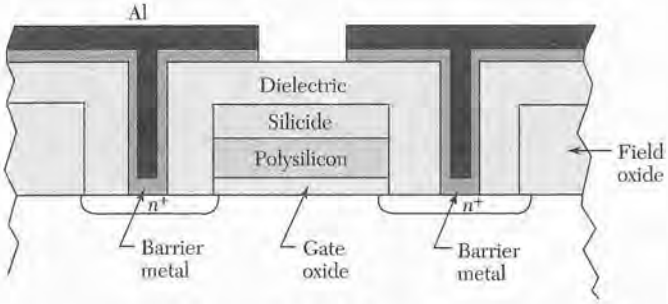


Fig. 22 Cross-sectional view of a MOSFET with a barrier metal between the aluminum and silicon and a composite gate electrode of silicide and polysilicon.

its deposition and formation are compatible with the overall process. Barrier metals such as titanium nitride (TiN) have been evaluated and found to be stable for contact annealing temperatures up to 550°C for 30 min.

Electromigration

In Chapter 6 we discussed scaled-down devices. As the device becomes smaller, the corresponding current density becomes larger. High current densities can cause device failure due to electromigration. The term electromigration refers to the transport of mass (i.e., atoms) in metals under the influence of current. It occurs by the transfer of momentum from the electrons to the positive metal ions. When a high current passes through thin metal conductors in integrated circuits, metal ions in some regions will pile up and voids will form in other regions. The pileup can short-circuit adjacent conductors, whereas the voids can result in an open circuit.

The mean time to failure (MTF) of a conductor due to electromigration can be related to the current density J and the activation energy E_a by

$$\mathrm{MTF} \sim \frac{1}{J^2} \exp\left(\frac{E_a}{kT}\right). \tag{32}$$

Experimentally, a value of $E_a \cong 0.5$ eV is obtained for deposited aluminum. This indicates that low-temperature, grain-boundary diffusion is the primary vehicle of material transport, since $E_a \cong 1.4$ eV would characterize the self-diffusion of single-crystal aluminum. The electromigration resistance of aluminum conductors can be increased by using several techniques. These techniques include alloying with copper (e.g., Al with 0.5% Cu), encapsulating the conductor in a dielectric, or incorporating oxygen during film deposition.

11.4.4 Copper Metallization

It is well known that both high conductivity wiring and low–dielectric-constant insulators are required to lower the *RC* time delay of the interconnect network. Copper is the obvious choice for a new interconnection metallization because it has higher conductivity and higher electromigration resistance than aluminum. Copper can be deposited by PVD, CVD, and electrochemical methods. However, the use of Cu as an alternative material to Al in ULSI circuits has drawbacks, such as its tendency to corrode under standard chip manufacture conditions, its lack of a feasible dry-etching method or a stable self-passivating oxide similar to Al_2O_3 on Al, and its poor adhesion to dielectric materials, such as SiO_2 and low-*k* polymers. In this section, we discuss the copper metallization techniques.

Several different techniques for fabrication of multilevel Cu interconnects have been reported.[17,18] The first method is a conventional method to pattern the metal lines followed by dielectric deposition. The second method is to pattern the dielectric layer first and fill copper metal into trenches. This step is followed by chemical mechanical polishing, discussed in Section 11.4.5, to remove the excess metal on the top surface of dielectric and leave Cu material in the holes and trenches. This method is also known as a damascene process.

Damascene technology

The approach for fabricating a copper/low-*k* dielectric interconnect structure is by the "damascene" or "dual damascene" process. Figure 23 shows the dual damascene sequence for an advanced Cu interconnection structure. For a typical damascene structure, trenches for metal lines are defined and etched in the interlayer dielectric (ILD) and then followed by metal deposition of TaN/Cu . The TaN layer serves as a diffusion barrier layer and prevents copper from penetrating the low-*k* dielectric. The excess copper metal on the surface is removed to obtain a planar structure with metal inlays in the dielectric.

For the dual damascene process, the vias and trenches in the dielectric are defined using two lithography and reactive ion etching (RIE) steps before depositing the Cu metal (Fig. 23*a–c*). Then a Cu chemical-mechanical polishing process is used to remove the metal on the top surface, leaving the planarized wiring and via imbedded in the insulator.[19] One special benefit of dual damascene is that the via plug is now of the same material as the metal line and the risk of via electromigration failure is reduced.

EXAMPLE 5

If we replace Al with Cu wire associated with some low-*k* dielectric ($k = 2.6$) instead of SiO_2 layer, what percentage of reduction of *RC* time constant will be achieved? (The resistivity of Al is 2.7 μΩ-cm, and the resistivity of Cu is 1.7 μΩ-cm.)

SOLUTION

$$\frac{1.7}{2.7} \times \frac{2.6}{3.9} \times 100\% = 42\%.$$

11.4.5 Chemical-Mechanical Polishing

In recent years, the development of chemical-mechanical polishing (CMP) has become increasingly important for multilevel interconnection because it is the only technology that allows global planarization (i.e., to make a flat surface across the whole wafer). It

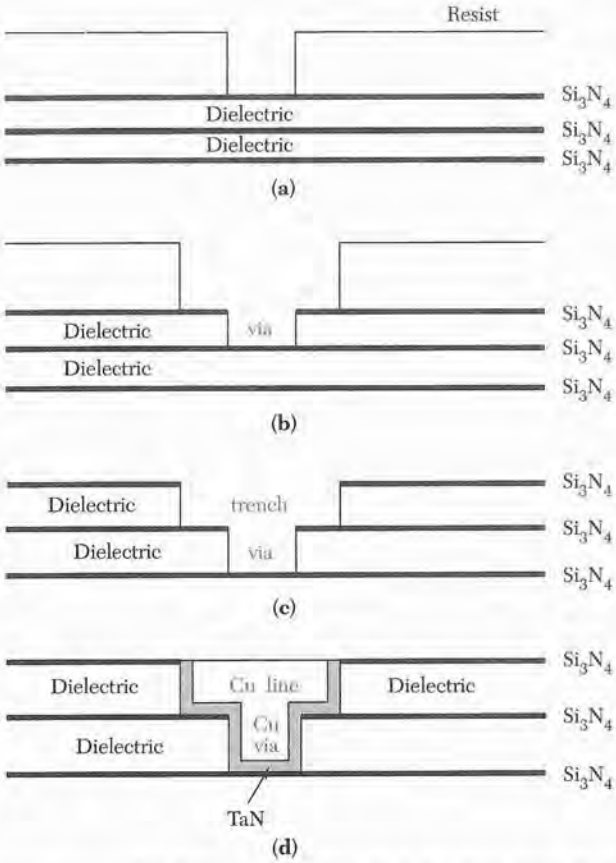


Fig. 23 Process sequence used to fabricate a Cu line-stud structure using dual damascene. (*a*) Resist stencil applied; (*b*) reactive ion etching dielectric and resist patterning; , (*c*) trench and via definition; and (*d*) Cu depositions followed by chemical-mechanical polishing (CMP).

that allows global planarization (i.e., to make a flat surface across the whole wafer). It offers many advantages over other types of technologies—better global planarization over large or small structures, reduced defect density, and the avoidance of plasma damage. Three CMP approaches are summarized in Table 4.

The CMP process consists of moving the sample surface against a pad that carries slurry between the sample surface and the pad. Abrasive particles in the slurry cause mechanical damage on the sample surface, loosening the material for enhanced chemical attack or fracturing off the pieces of surface into a slurry where they dissolve or are swept away. The process is tailored to provide an enhanced material removal rate from

TABLE 4 Three Methods of Chemical-Mechanical Polishing (CMP)

Method	Wafer facing	Platen movement	Slurry feeding
Rotary CMP	Down	Rotary against rotating wafer carrier	Dripping to pad surface
Orbital CMP	Down	Orbital against rotating wafer carrier	Through the pad surface
Linear CMP	Down	Linear against rotating wafer carrier	Dripping to pad surface

high points on surfaces, thus affecting the planarization because most chemical actions are isotropic. Mechanical grinding alone, theoretically, may achieve the desired planarization but is not desirable because of extensive associated damage to the material surfaces. There are three main parts of the process: the surface to be polished, the pad—the key media enabling the transfer of mechanical action to the surface being polished, and the slurry—which provides both chemical and mechanical effects. Figure 24 shows the CMP setup. [20]

EXAMPLE 6

The oxide removal rate and the removal rate of a layer underneath the oxide (called a stop layer) are $1r$ and $0.1r$, respectively. To remove 1 μm oxide and 0.01 μm stop layer, the total removal time is 5.5 min. Find the oxide removal rate.

SOLUTION

$$\frac{1}{1r} + \frac{0.01}{0.1r} = 5.5$$

$$\frac{1.1}{1r} = 5.5, \qquad r = 0.2\ \mu\text{m / min.}$$

11.4.6 Silicide

Silicon forms many stable metallic and semiconducting compounds with metals. Several metal silicides show low resisivity and high thermal stability, making them suitable for ULSI application. Silicides such as $TiSi_2$ and $CoSi_2$ have reasonably low resistivities and are generally compatible with integrated-circuit processing. Silicides become important metallization materials as devices become smaller. One important application of silicide is for the MOSFET gate electrode either alone or with doped polysilicon (polycide) above the gate oxide. Table 5 shows a comparison of titanium silicide and cobalt silicide.

Metal silicides have been used to reduce the contact resistance of the source and drain, the gate electrodes, and the interconnections. The self-aligned metal silicide technology (salicide) has been proven to be a highly attractive technique for improving the performance of submicron devices and circuits. The self-aligned process uses the silicide gate electrode as the mask to form the source and drain electrodes of a MOSFET (e.g., by ion implantation, considered in Chapter 13). This process can minimize the overlaps of these electrodes and thus reduce the parasitic capacitances.

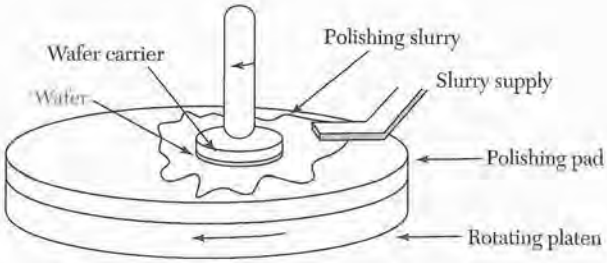


Fig. 24 A schematic of a CMP polisher.

TABLE 5 A Comparsion of $TiSi_2$ and $CoSi_2$ Films

Properties	$TiSi_2$	$CoSi_2$
Resistivity	13–16	22–28
Silicide/metal ratio	2.37	3.56
Silicide/Si ratio	1.04	0.97
Reactive to native oxide	Yes	No
Silicidation temperature (°C)	800–850	550–900
Film stress (dyne/cm^2)	1.5×10^{10}	1.2×10^{10}

Figure 25 shows the polycide and salicide process. A typical polycide formation sequence is shown in Fig. 25*a*. For sputter deposition, a high-temperature, high-purity compound target is used to ensure the quality of the silicide. The most commonly used silicides for the polycide process are WSi_2, $TaSi_2$, and $MoSi_2$. They are all refractory, thermally stable, and resistant to processing chemicals. A self-aligned silicide process is illustrated in Fig. 25*b*. In the process, the polysilicon gate is patterned without any silicide, and a sidewall spacer (silicon oxide or silicon nitride) is formed to prevent shorting the gate to the source and drain during the silicidation process. A metal layer, either Ti or Co, is blanket-sputtered on the entire structure, followed by silicide sintering. Silicide is formed, in principle, only where the metal is in contact with Si. A wet chemical wash then rinses off the unreacted metal, leaving only the silicide. This technique eliminates

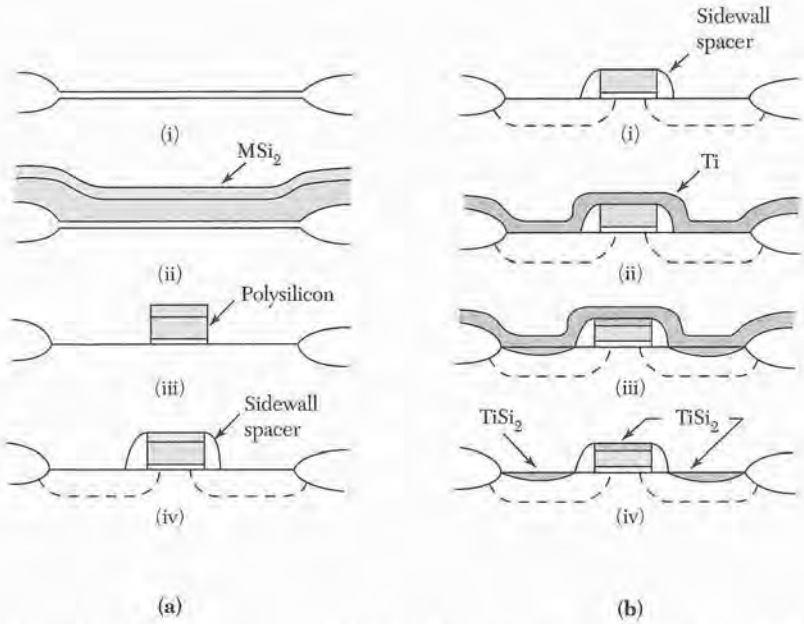


Fig. 25 Polycide and salicide processes. (*a*) Polycide structure: (i) gate oxide; (ii) polysilicon and silicide deposition, (iii) pattern polycide; and (iv) lightly doped drain (LDD) implant, sidewall formation, and *S/D* implant. (*b*) Salicide structure: (i) gate patterning (polysilicon only), LDD, sidewall, and *S/D* implant; (ii) metal (Ti,Co) deposition; (iii) anneal to form salicide; and (iv) selective (wet) etch to remove unreacted metal.

the need to pattern the composite polycide gate structure and adds silicide to the source/drain area to reduce the contact resistance.

The silicides are promising materials for ULSI circuits because of their low resistivity and excellent thermal stability. Cobalt silicide has been widely investigated recently because of its lowest resistivity and high-temperature thermal stability. However, cobalt is sensitive to native oxide as well as an oxygen-contained environment, and a large amount of silicon is consumed during silicidation.

EXAMPLE 7

Calculate the thickness of the cobalt silicide. For a desired sheet resistance of 0.6 Ω/sq, the resistivity is 18 μΩ-cm.

SOLUTION The resistivity is equal to the product of the sheet resistance and the film thickness:

$$\rho = R_s \times t.$$

Then

$$t = \frac{\rho}{R_s} = \frac{18 \times 10^{-6}}{0.6} = 3 \times 10^{-5} \text{ cm} = 300 \text{ nm}.$$

SUMMARY

Modern semiconductor device fabrication requires the use of thin films. Currently, there are four important types of films—thermal oxides, dielectric layers, polycrystalline silicon, and metal films. The major issues related to film formation are low-temperature processing, step coverage, selective deposition, uniformity, film quality, planarization, throughput, and large-wafer capacity.

Thermal oxidation offers the best quality for the Si-SiO_2 interface and has the lowest interface trap density. Therefore, it is used to form the gate oxide and the field oxide. LPCVD of dielectrics and polysilicon offer conformal step coverage. In contrast, PVD and atmospheric-pressure CVD generally result in noncomformal step coverage. CMP offers global planarization and reduces defect density. Conformal step coverage and planarization are also required for precise pattern transfer at the deep-submicron lithography level. Pattern transfer technology is discussed in the next chapter.

To minimize the *RC* time delay due to parasitic resistance and capacitance, the silicide process for ohmic contacts, copper metallization for interconnects, and low–dielectric-constant materials for interlayer films are extensively used to meet the requirements of the multilevel interconnect structures of ULSI circuits. In addition, we have investigated high-dielectric-constant materials to improve the gate insulator performance and to increase the capacitance per unit area for DRAM.

REFERENCES

1. E. H. Nicollian and J. R. Brews, *MOS Physics and Technology*, Wiley, New York, 1982.
2. B. E. Deal and A. S. Grove, "General Relationship for the Thermal Oxidation of Silicon," *J. Appl. Phys.*, **36**, 3770 (1965).
3. J. D. Meindl, et al., "Silicon Epitaxy and Oxidation," in F. Van de wiele, W. L. Engl, and P. O. Jespers, Eds., *Process and Device Modeling for Integrated Circuit Design*, Noorhoff, Leyden, 1977.

4. For a discussion on film deposition, see, for example, A.C. Adams, "Dielectric and Polysilicon Film Deposition," in S. M. Sze, Ed., *VLSI Technology*, McGraw-Hill, New York, 1983.

5. K. Eujino, et al., "Doped Silicon Oxide Deposition by Atmospheric Pressure and Low Temperature Chemical Vapor Deposition Using Tetraethoxysilane and Ozone," *J. Electrochem. Soc.*, **138**, 3019 (1991).

6. A. C. Adams and C. D. Capio, "Planarization of Phosphorus-Doped Silicon Dioxide," *J. Electrochem. Soc.*, **127**, 2222 (1980).

7. T. Yamamoto et al., "An Advanced 2.5nm Oxidized Nitride Gate Dielectric for Highly Reliable 0.25 μm MOSFETs," *Symp. on VLSI Technol. Dig. of Tech. Pap*, 1997, p. 45.

8. K. Kumar, et al., "Optimization of Some 3 nm Gate Dielectrics Grown by Rapid Thermal Oxidation in a Nitric Oxide Ambient," *Appl. Phys. Lett.*, **70**, 384 (1997).

9. T. Homma, "Low Dielectric Constant Materials and Methods for Interlayer Dielectric Films in Ultralarge-Scale Integrated Circuit Multilevel Interconnects," *Mater. Sci. Eng.*, **23**, 243 (1998).

10. H. N. Yu, et al., "1 μm MOSFET VLSI Technology. Part I—An Overview," *IEEE Trans. Electron Devices*, **ED-26**, 318 (1979).

11. J. M. Andrews, "Electrical Conduction in Implanted Polycrystalline Sillicon," *J. Electron. Mater.*, **8**, 3, 227 (1979).

12. M. J. Buiting, A. F. Otterloo, and A. H. Montree, "Kinetical Aspects of the LPCVD of Titanium Nitride from Titanium Tetrachloride and Ammonia," *J. Electrochem. Soc.*, **138**, 500 (1991).

13. R. Tobe, et al., "Plasma-Enhanced CVD of TiN and Ti Using Low-Pressure and High-Density Helicon Plasma," *Thin Solid Film*, **281–282**, 155 (1996).

14. J. Hu, et al., "Electrical Properties of Ti/TiN Films Prepared by Chemical Vapor Deposition and Their Applications in Submicron Structures as Contact and Barrier Materials, " *Thin Solid Film*, **308**, 589 (1997).

15. M. Hansen and A. Anderko, *Constitution of Binary Alloys*, McGraw-Hill, New York, 1958.

16. D. Pramanik and A. N. Saxena, "VLSI Metallization Using Aluminum and Its Alloys," *Solid State Tech.*, **26**, No. 1, 127 (1983), **26**. No. 3, 131 (1983).

17. C. L. Hu, and J. M. E. Harper, "Copper Interconnections and Reliability," *Matter. Chem. Phys.*, **52**, 5 (1998)

18. P. C. Andricacos, et al., "Damascene Copper Electroplating for Chip Interconnects," *193rd Meet. Electrochem. Soc.*, 1998, p. 3

19. J. M. Steigerwald, et al., "Chemical Mechanical Planarization of Microelectronic Materials," Wiley, New York, 1997.

20. L. M. Cook, et al.,*Theoretical and Practical Aspects of Dielectric and Metal CMP*, Semicond. Int., p. 141 (1995).

PROBLEMS (* DENOTES DIFFICULT PROBLEMS)

FOR SECTION 11.1 THERMAL OXIDATION

1. A p-type <100>-oriented, silicon wafer with a resistivity of 10 Ω-cm is placed in a wet oxidation system to grow a field oxide of 0.45 μm at 1050°C. Determine the time required to grow the oxide.

*2. After the first oxidation as given in Prob. 1, a window is opened in the oxide to grow a gate oxide at 1000°C for 20 minutes in dry oxidation. Find the thicknesses of the gate oxide and the total field oxide.

3. Show that Eq. 11 reduces to $x^2 = Bt$ for long times and to $x = B/A(t + \tau)$ for short times.

4. Determine the diffusion coefficient D for dry oxidation of <100>-oriented silicon samples at 980°C and 1 atm.

FOR SECTION 11.2 DIELECTRIC DEPOSITION

5. (a) In a plasma-deposited silicon nitride that contains 20 at% hydrogen and has a silicon-to-nitrogen ratio (Si/N) of 1.2, find x and y in the empirical formula of SiN_xH_y. (b) If the variation of film resistivity with Si/N ratio is given by $5 \times 10^{28} \exp(-33.3\gamma)$ for $2 > \gamma > 0.8$, where γ is the ratio, find the resistivity of the film in (a).

6. The dielectric constants of SiO_2, Si_3N_4, and Ta_2O_5 are about 3.9, 7.6, and 25, respectively. What is the capacitance ratio for the capacitors with the Ta_2O_5 and oxide/nitride/oxide dielectrics for the same dielectric thickness, provided the oxide/nitride/oxide has thickness ratio 1:1:1 for the oxide to the nitride?

7. In Prob. 6, if BST with a dielectric constant of 500 is chosen to replace Ta_2O_5, calculate the area reduction ratio to maintain the same capacitance if the two films have the same thickness.

8. In Prob. 6, calculate the equivalent thickness of the Ta_2O_5 in terms of SiO_2 thickness if both have the same capacitance. Assume the actual thickness of Ta_2O_5 is $3t$.

9. In a silane-oxygen reaction to deposit undoped SiO_2 film, the deposition rate is 15 nm/min at 425°C. What temperature is required to double the deposition rate?

10. The P-glass flow process requires temperatures above 1000°C. As device dimensions become smaller in ULSI, we must use lower temperatures. Suggest methods to obtain a smooth topography at < 900°C for deposited silicon dioxide that can be used as an insulator between metal layers.

FOR SECTION 11.3 POLYSILICON DEPOSITION

11. Why is silane more often used for polysilicon deposition than silicon chloride?

12. Explain why the deposition temperature for polysilicon films is moderately low, usually between 600°–650°C.

FOR SECTION 11.4 METALLIZATION

13. An e-beam evaporation system is used to deposit aluminum to form MOS capacitors. If the flatband voltage of the capacitance is shifted by 0.5 V because of e-beam radiation, find the number of fixed oxide charges (the silicon dioxide thickness is 50 nm). How can these charges be removed?

14. A metal line (L = 20 μm, W = 0.25 μm) has a sheet resistance 5 Ω/sq. Calculate the resistance of the metal line.

15. Calculate the thickness of the $TiSi_2$ and $CoSi_2$, where the initial Ti and Co film thickness are 30 nm.

16. Compare the advantages and disadvantages of $TiSi_2$ and $CoSi_2$ for salicide applications.

17. A dielectric material is placed between the two parallel metal lines. The length L = 1 cm, width W = 0.28 μm, thickness T = 0.3 μm, and spacing S = 0.36 μm. (a) Calculate the RC time delay. The metal is Al with a resistivity of 2.67 μΩ-cm and the dielectric is oxide with dielectric constant 3.9. (b) Calculate the RC time delay. The metal is Cu with a resistivity of 1.7 μΩ-cm and the dielectric is organic polymer with dielectric constant 2.8. (c) Compare the results in (a) and (b). How much can we decrease the RC time delay?

18. Repeat Prob. 17 (a) and (b) if the fringing factor for the capacitors is 3. The fringing factor is due to the spreading of the electric-field lines beyond the length and the width of the metal lines.

*19. To avoid electromigration problems, the maximum allowed current density in an aluminum runner is about 5×10^5 A/cm^2. If the runner is 2 mm long, 1 μm wide, and nominally 1 μm thick, and if 20% of the runner length passes over steps and is only 0.5 μm

thick there, find the total resistance of the runner if the resistivity is $3 \times 10^{-6}\,\Omega$-cm. Find the maximum voltage that can be applied across the runner.

*20. To use Cu for wiring one must overcome several obstacles: the diffusion of Cu through SiO_2, adhesion of Cu to SiO_2, and corrosion of Cu. One way to overcome these obstacles is to use a cladding/adhesion layer (e.g. Ta or TiN) to protect the Cu wires. Consider a cladded Cu wire with a square cross section of $0.5\ \mu m \times 0.5\ \mu m$ and compare it with a layered TiN/Al/TiN wire of the same size, with the top and bottom TiN layers 40 nm and 60 nm thick, respectively. What is the maximum thickness of the cladding layer if the resistance of the cladded Cu wire and the TiN/Al/TiN wire is the same?

CHAPTER 12

Lithography and Etching

12.1 OPTICAL LITHOGRAPHY
12.2 NEXT-GENERATION LITHOGRAPHIC METHODS
12.3 WET CHEMICAL ETCHING
12.4 DRY ETCHING
12. 5 MICROELECTROMECHANICAL SYSTEMS
SUMMARY

Lithography is the process of transferring patterns of geometric shapes on a mask to a thin layer of radiation-sensitive material (called resist) covering the surface of a semiconductor wafer.[1] These patterns define the various regions in an integrated circuit such as the implantation regions, the contact windows, and the bonding-pad areas. The resist patterns defined by the lithographic process are not permanent elements of the final device but only replicas of circuit features. To produce circuit features, these resist patterns must be transferred once more into the underlying layers comprising the device. The pattern transfer is accomplished by an etching process that selectively removes unmasked portions of a layer.[2] A brief description of the pattern transfer has been given in Section 4.1.

Specifically, we cover the following topics:

- The importance of a clean room for lithography.
- The most widely used lithographic method—the optical lithography and its resolution-enhancement techniques.
- Advantages and limitations of other lithographic methods.
- Mechanisms for wet chemical etching of semiconductors, insulator, and metal films.
- Plasma-assisted etching (also called dry etching) for high-fidelity pattern transfer.
- Microelectromechanical systems (MEMS) formed by orientation-dependent etching, sacrificial etching, or LIGA (lithography, electroplating, and molding) process.

12.1 OPTICAL LITHOGRAPHY

The vast majority of lithographic equipment for integrated-circuit (IC) fabrication is optical equipment using ultraviolet light ($\lambda \cong 0.2$–$0.4\ \mu m$). In this section we consider the exposure tools, the masks, the resists , and resolution-enhancement techniques used for optical lithography. We also consider the pattern transfer process, which serves as a basis

for other lithographic systems. We first briefly consider the *clean room*, because all lithographic processes must be performed in an ultraclean environment.

12.1.1 The Clean Room

An IC fabrication facility requires a clean processing room, especially in the area used for lithography. The need for such a clean room arises because dust particles in the air can settle on semiconductor wafers and lithographic masks and can cause defects in the devices, which result in circuit failure. For example, a dust particle on a semiconductor surface can disrupt the single-crystal growth of an epitaxial film, causing the formation of dislocations. A dust particle incorporated into the gate oxide can result in enhanced conductivity and cause device failure due to low breakdown voltage. The situation is even more critical in the lithographic area. When dust particles adhere to the surface of a photomask, they behave as opaque patterns on the mask, and these patterns will be transferred to the underlying layer along with the circuit patterns on the mask. Figure 1 shows three dust particles on a photomask.[3] Particle 1 may result in the formation of a pinhole in the underlying layer. Particle 2 is located near a pattern edge and may cause a constriction of current flow in a metal runner. Particle 3 can lead to a short circuit between the two conducting regions and render the circuit useless.

In a clean room, the total number of dust particles per unit volume must be tightly controlled along with the temperature and humidity. Figure 2 shows the particle-size distribution curves for various *classes* of clean rooms. We have two systems to define the classes of clean room.[4] For the English system, the numerical designation of the class is taken from the maximum allowable number of particles 0.5 μm and larger, per *cubic foot*. For the metric system, the class is taken from the logarithm (base 10) of the maximum allowable number of particles 0.5 μm and larger, per *cubic meter*. For example, a class 100 clean room (English system) has a dust count of 100 particles/ft^3 with particle diameters of 0.5 μm and larger, whereas a class M 3.5 clean room (metric system) has a dust count of $10^{3.5}$ or about 3500 particles/m^3 with particle diameters of 0.5 μm or large. Since

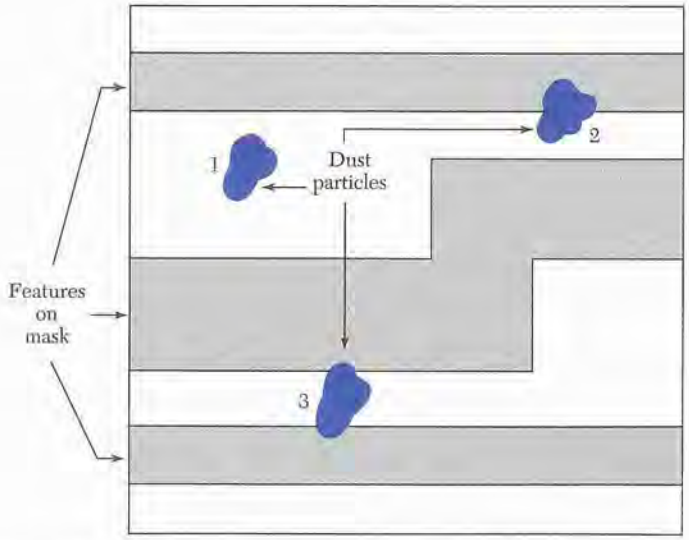


Fig. 1 Various ways in which dust particles can interfere with photomask patterns.[3]

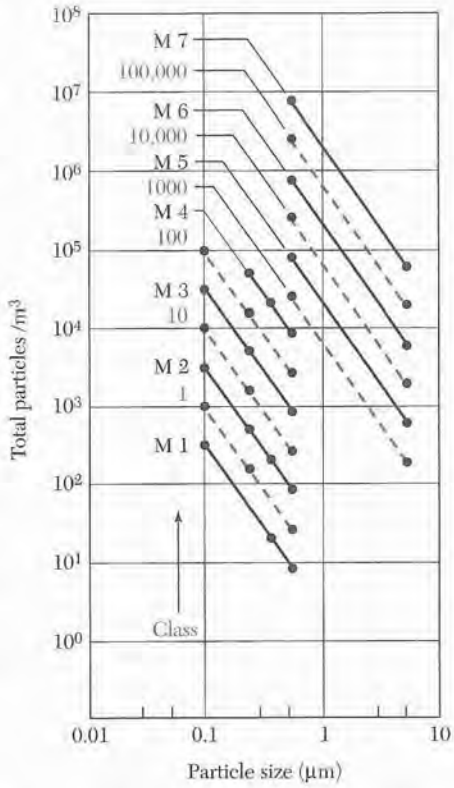


Fig. 2 Particle-size distribution curve for English (- - -) and metric (—) classes of clean rooms.[4]

100 particles/ft^3 = 3500 particles/m^3, a class 100 in English system corresponds to a class M 3.5 in the metric system.

Since the number of dust particles increases as particle size decreases, a more stringent control of the clean room environment is required when the minimum feature lengths of ICs are reduced to the deep-submicron range. For most IC fabrication areas, a class 100 clean room is required, that is, the dust count must be about four orders of magnitude lower than that of ordinary room air. However, for the lithography area, a class 10 clean room or one with a lower dust count is required.

EXAMPLE 1

If we expose a 200-mm wafer for 1 minute to an air stream under a laminar-flow condition at 30 m/min, how many dust particles will land on the wafer in a class 10 clean room?

SOLUTION For a class 10 clean room, there are 350 particles (0.5 μm and larger) per cubic meter. The air volume that goes over the wafer in 1 minute is

$$(30\ \text{m/min}) \times \pi \left(\frac{0.2\ \text{m}}{2}\right)^2 \times 1\ \text{minute} = 0.942\ \text{m}^3.$$

The number of dust particles (0.5 μm and larger) contained in the air volume is

$350 \times 0.942 = 330$ particles.

Therefore, if there are 400 IC chips on the wafer, the particle count amounts to one particle on each of 82% of the chips. Fortunately, only a fraction of the particles that land adhere to the wafer surface, and of those only a fraction are at a circuit location critical enough to cause a failure. However, the calculation indicates the importance of the clean room.

12.1.2 Exposure Tools

The pattern transfer process is accomplished by using a lithographic exposure tool. The performance of an exposure tool is determined by three parameters: resolution, registration, and throughput. *Resolution* is the minimum feature dimension that can be transferred with high fidelity to a resist film on a semiconductor wafer. *Registration* is a measure of how accurately patterns on successive masks can be aligned (or overlaid) with respect to previously defined patterns on the wafer. *Throughput* is the number of wafers that can be exposed per hour for a given mask level.

There are basically two optical exposure methods: shadow printing and projection printing.[5,6] Shadow printing may have the mask and wafer in direct contact with one another as in *contact printing*, or in close proximity as in *proximity printing*. Figure 3*a* shows a basic setup for contact printing where a resist-coated wafer is brought into physical contact with a mask, and the resist is exposed by a nearly collimated beam of ultraviolet light through the back of the mask for a fixed time. The intimate contact between resists and mask provides a resolution of ~1 μm. However, contact printing suffers a major drawback caused by dust particles. A dust particle or a speck of silicon dust on the wafer can be imbedded into the mask when the mask makes contact with the wafer. The imbedded particle causes permanent damage to the mask and results in defects in the wafer with each succeeding exposure.

To minimize mask damage, the proximity exposure method is used. Figure 3*b* shows the basic setup. It is similar to the contact printing method except that there is a small gap, 10–50 μm, between the wafer and the mask during exposure. The small gap, however, results in optical diffraction at feature edges on the photomask, that is, when light

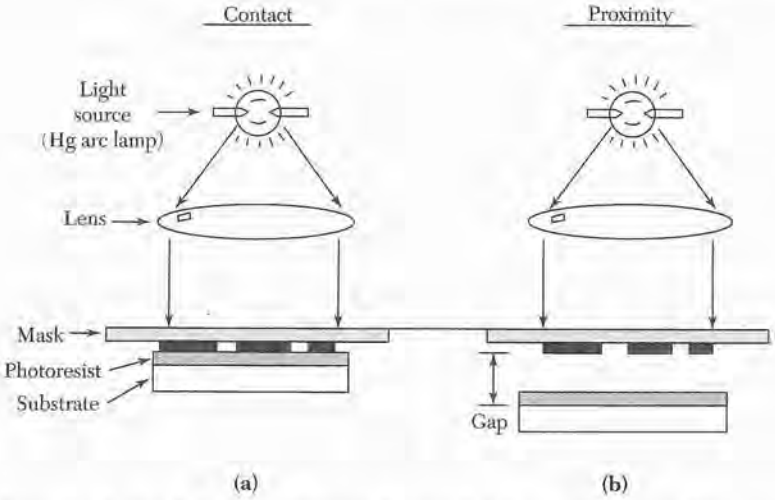


Fig. 3 Schematic of optical shadow printing techniques[1]: (*a*) contact printing, (*b*) proximity printing.

passes by the edges of an opaque mask feature, fringes are formed and some light penetrates into the shadow region. As a result, the resolution is degraded to the 2–5 μm range.

In shadow printing, the minimum linewidth [or critical dimension (CD)] that can be printed is roughly

$$\boxed{\mathrm{CD} \cong \sqrt{\lambda g}\,;} \tag{1}$$

when λ is the wavelength of the exposure radiation and g is the gap between the mask and the wafer and includes the thickness of the resist. For $\lambda = 0.4$ μm and $g = 50$ μm, the CD is 4.5 μm. If we reduce λ to 0.25 μm (wavelength range of 0.2–0.3 μm is in the deep-UV spectral region) and g to 15 μm, CD becomes 2 μm. Thus, there is an advantage in reducing both λ and g. However, for a given distance g, any dust particle with a diameter larger than g potentially can cause mask damage.

To avoid the mask damage problem associated with shadow printing, projection-printing exposure tools have been developed to project an image of the mask patterns onto a resist-coated wafer many centimeters away from the mask. To increase resolution, only a small portion of the mask is exposed at a time. The small image area is scanned or stepped over the wafer to cover the entire wafer surface. Figure 4*a* shows a 1:1 wafer scan projection system.[6,7] A narrow, arc-shaped image field ~1 mm in width serially transfers the slit image of the mask onto the wafer. The image size on the wafer is the same as that on the mask.

The small image field can also be stepped over the surface of the wafer by two-dimensional translations of the wafer only, whereas the mark remains stationary. After the exposure of one chip site, the wafer is moved to the next chip site and the process is repeated.

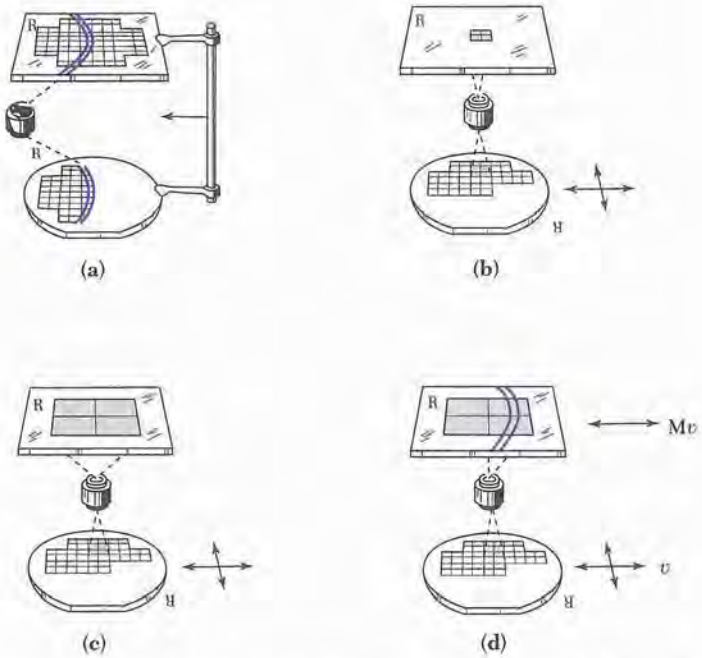


Fig. 4 Image partitioning techniques for projection printing: (*a*) annual-field wafer scan, (*b*) 1:1 step-and-repeat, (*c*) M:1 reduction step-and repeat, and (*d*) M:1 reduction step-and-scan.[6,7]

Figure 4*b* and 4*c* show the partitioning of the wafer image by *step-and-repeat projection* with a ratio of 1:1 or at a demagnification ratio M:1 (e.g., 10:1 for a 10 times reduction on the wafer), respectively. The demagnification ratio is an important factor in our ability to produce both the lens and the mask from which we wish to print. The 1:1 optical systems are easier to design and fabricate than a 10:1 or a 5:1 reduction systems, but it is much more difficult to produce defect-free masks at 1:1 than it is at a 10:1 or a 5:1 demagnification ratio.

Reduction projection lithography can also print larger wafers without redesigning the stepper lens as long as the field size (i.e., the exposure area onto the wafer per se) of the lens is large enough to contain one or more IC chips. When the chip size exceeds the field size of the lens, further partitioning of the image on the reticle is necessary. In Fig. 4*d* the image field on the reticle can be a narrow, arc-shaped for M:1 step-and-scan projection lithography. For the step-and-scan system, we have two-dimensional translations of the wafer with speed v, and one-dimensional translation of the mask with M times that of the wafer speed.

The resolution of a projection system is given by

$$l_m = k_1 \frac{\lambda}{\mathrm{NA}}, \tag{2}$$

where λ is again the exposure wavelength, k_1 is a process dependent factor, and NA is the numerical aperture, which is given by

$$\mathrm{NA} = \overline{n}\sin\theta \tag{3}$$

with $\overline{n}$ the index of refraction in the image medium (usually air, where $\overline{n} = 1$), and θ the half-angle of the cone of light converging to a point image at the wafer, as shown[5] in Fig. 5. Also shown in the figure is the depth of focus (DOF), which can be expressed as

$$\boxed{\mathrm{DOF} = \frac{\pm l_m/2}{\tan\theta} \approx \frac{\pm l_m/2}{\sin\theta} = k_2 \frac{\lambda}{(\mathrm{NA})^2},} \tag{4}$$

where k_2 is another process-dependent factor.

Equation 2 indicates that resolution can be improved (i.e., smaller l_m) by either reducing the wavelength or increasing NA or both. However, Eq. 4 indicates that the DOF

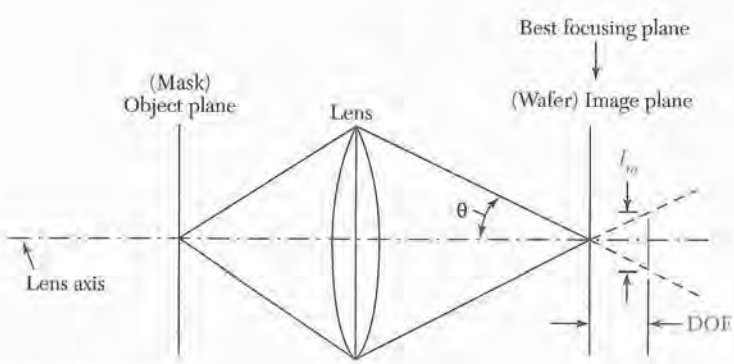


Fig. 5 Simple image system.[5]

degrades much more rapidly by increasing NA than by decreasing λ. This explains the trend toward shorter-wavelength sources in optical lithography.

The high-pressure mercury-arc lamp is widely used in exposure tools because of its high intensity and reliability. The mercury-arc spectrum is composed of several peaks, as shown in Fig. 6. The terms G-line, H-line, and I-line refer to the peaks at 436 nm, 405 nm, and 365 nm, respectively. I-line lithography with 5:1 step-and-repeat projection can offer a resolution of 0.3 μm with resolution enhancement techniques (see Section 12.1.6). Advanced exposure tools such as the 248 nm lithographic system using a KrF excimer laser, the 193 nm lithographic system using an ArF excimer laser, and the 157 nm lithographic system using a F_2 excimer laser have been developed for mass production with a resolution of 0.18 μm (180 nm), 0.10 μm (100 nm), and 0.07 μm (70 nm), respectively.

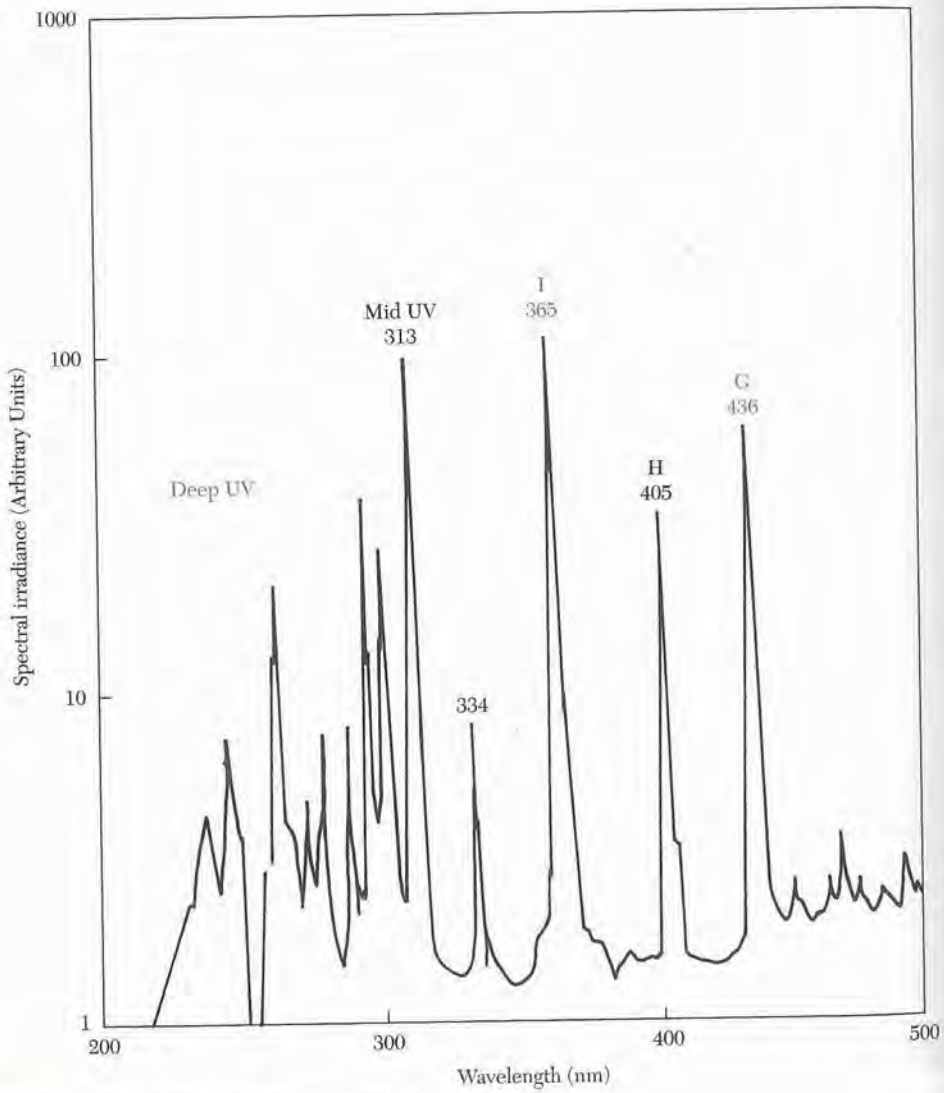


Fig. 6 Typical high-pressure mercury-arc lamp spectrum.

12.1.3 Masks

Masks used for IC manufacturing are usually reduction reticles. The first step in mask making is to use a computer-aided design (CAD) system in which designers can completely describe the circuit patterns electrically. The digital data produced by the CAD system then drives a pattern generator, which is an electron-beam lithographic system (see Section 12.2.1) that transfers the patterns directly to electron-sensitized mask. The mask consists of a fused silica substrate covered with a chrominum layer. The circuit pattern is first transferred to the electron-sensitized layer (electron resist), which is transferred once more into the underlying chrominum layer for the finished mask. The details of pattern transfer are considered in Section 12.1.5.

The patterns on a mask represent one level of an IC design. The composite layout is broken into mask levels that correspond to the IC process sequence such as the isolation region on one level, the gate region on another, and so on. Typically, 15–20 different mask levels are required for a complete IC process cycle.

The standard-size mask substrate is a fused silica plate 15 × 15 cm square, 0,6 cm thick. The size is needed to accommodate the lens field sizes for 4:1 or 5:1 optical exposure tools, whereas the thickness is required to minimize pattern placement errors due to substrate distortion. The fused silica plate is needed for its low coefficient of thermal expansion, its high transmission at shorter wavelengths, and its mechanical strength. Figure 7 shows a mask on which patterns of geometric shapes have been formed. A few secondary-chip sites, used for process evaluation, are also included in the mask.

One of the major concerns about masks is the defect density. Mask defects can be introduced during the manufacture of the mask or during subsequent lithographic processes. Even a small mask-defect density has a profound effect on the final IC yield. The *yield* is defined as the ratio of good chips per wafer to the total number of chips per wafer. As a first-order approximation, the yield Y for a given masking level can be expressed as

$$Y \cong e^{-DA}, \tag{5}$$

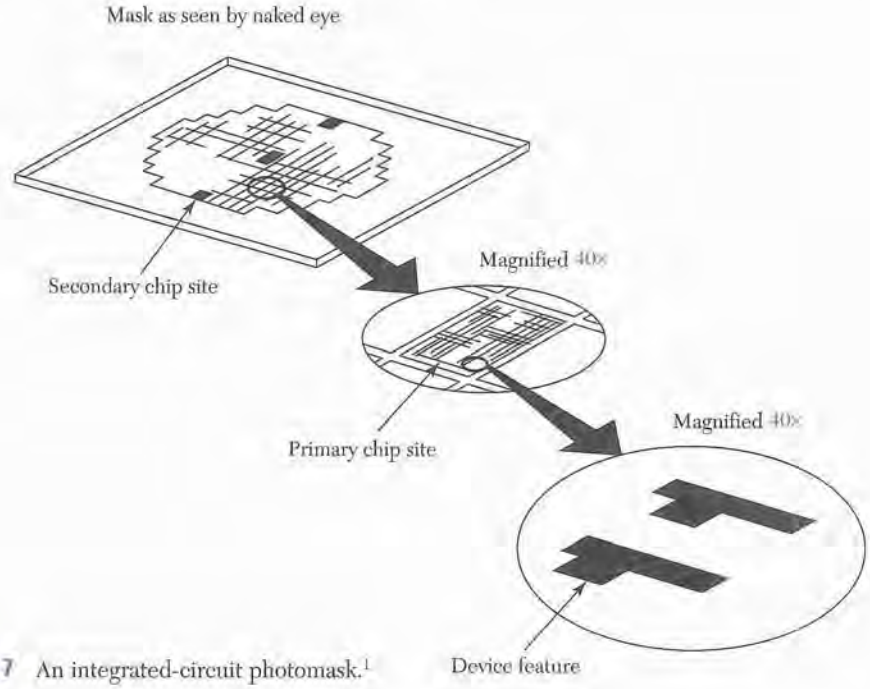


Fig. 7 An integrated-circuit photomask.[1]

where D is the average number of "fatal" defects per unit area and A is the area of an IC chip. If D remains the same for all mask levels (e.g., $N = 10$ levels), then the final yield becomes

$$Y \cong e^{-NDA}. \tag{6}$$

Figure 8 shows the mask-limit yield for a 10-level lithographic process as a function of chip size for various values of defect densities. For example, for $D = 0.25$ defect/cm^2, the yield is 10% for a chip size of 90 mm^2, and it drops to about 1% for a chip size of 180 mm^2. Therefore, inspection and cleaning of masks are important to achieve high yields on large chips. Of course, an ultraclean processing area is mandatory for lithographic processing.

12.1.4 Photoresist

The photoresist is a radiation-sensitive compound. Photoresists can be classified as positive and negative, depending on how they respond to radiation. For positive resists, the exposed regions become more soluble and thus more easily removed in the development process. The net result is that the patterns formed (also called images) in the positive resist are the same as those on the mask. For negative resists, the exposed regions become less soluble, and the patterns formed in the negative resist are the reverse of the mask patterns.

Positive photoresists consist of three components: a photosensitive compound, a base resin, and an organic solvent. Prior to exposure, the photosensitive compound is insoluble in the developer solution. After exposure, the photosensitive compound absorbs radiation in the exposed pattern areas, changes its chemical structure, and becomes soluble in the developer solution. After development, the exposed areas are removed.

Negative photoresists are polymers combined with a photosensitive compound. After exposure, the photosensitive compound absorbs the optical energy and converts it into

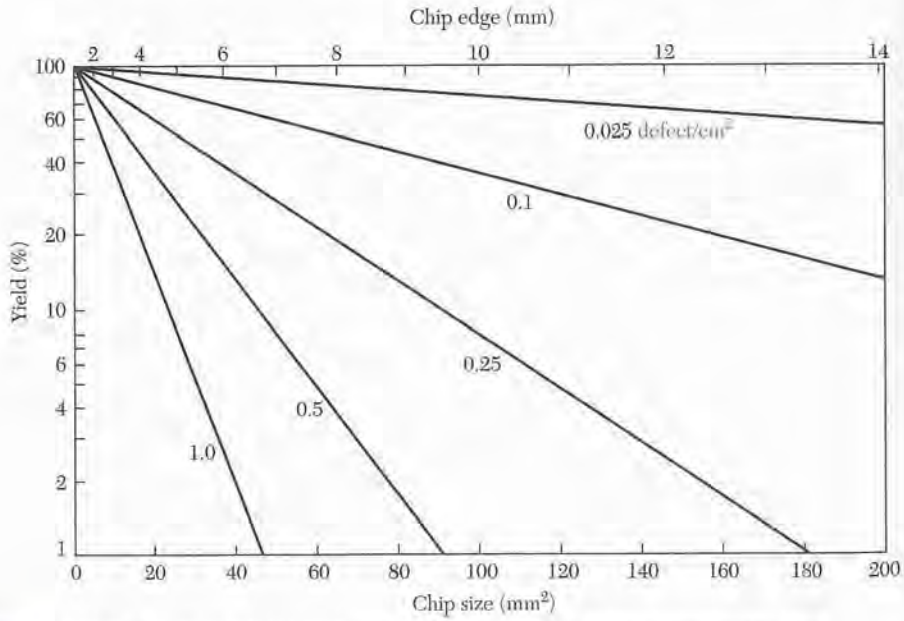


Fig. 8 Yield for a 10-mask lithographic process with various defect densities per level.

chemical energy to initiate a polymer linking reaction. This reaction causes cross linking of the polymer molecules. The cross-linked polymer has a higher molecular weight and becomes insoluble in the developer solution. After development, the unexposed areas are removed. One major drawback of a negative photoresist is that in the development process, the whole resist mass swells by absorbing developer solvent. This swelling action limits the resolution of negative photoresists.

Figure 9*a* shows a typical exposure response curve and image cross section for a positive resist.[1] The response curve describes the percentage of resist remaining after exposure and development versus the exposure energy. Note that the resist has a finite solubility in its developer, even without exposure to radiation. As the exposure energy increases, the solubility gradually increases until at a threshold energy E_T, the resist becomes completely soluble. The sensitivity of a positive resist is defined as the energy required to produce complete solubility in the exposed region. Thus, E_T corresponds to the sensitivity. In addition to E_T, a parameter γ, the contrast ratio, is defined to characterize the resist:

$$\boxed{\gamma \equiv \left[\ln\left(\frac{E_T}{E_1}\right)\right]^{-1},} \tag{7}$$

where E_1 is the energy obtained by drawing the tangent at E_T to reach 100% resist thickness as shown in Fig. 9*a*. A larger γ implies a higher solubility of the resist with an incremental increase of exposure energy and results in sharper images.

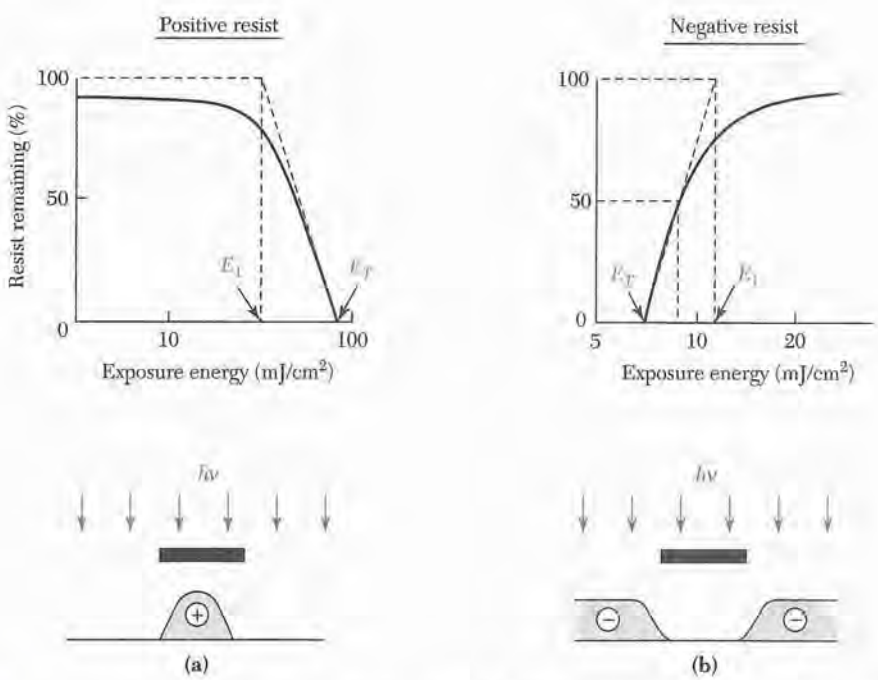


Fig. 9 Exposure-response curve and cross section of the resist image after development.[1] (*a*) Positive photoresist; (*b*) negative photoresist.

The image cross section in Fig. 9*a* illustrates the relationship between the edges of a photomask image and the corresponding edges of the resist images after development. The edges of the resist image are generally not at the vertically projected positions of the mask edges because of *diffraction*. The edge of the resist image corresponds to the position where the total absorbed optical energy equals the threshold energy E_T.

Figure 9*b* shows the exposure-response curve and image cross section for a negative resist. The negative resist remains completely soluble in the developer solution for exposure energies lower than the threshold energy E_T. Above E_T, more of the resist film remains after development. At exposure energies twice the threshold energy, the resist film becomes essentially insoluble in the developer. The sensitivity of a negative resist is defined as the energy required to retain 50% of the original resist film thickness in the exposed region. The parameter γ is defined similarly to γ in Eq. 7 except that E_1 and E_T are interchanged. The image cross section for the negative resist (Fig. 9*b*) is also influenced by the diffraction effect.

EXAMPLE 2

Find the parameter γ for the photoresists shown in Fig. 9

SOLUTION For the positive resist, $E_T = 90$ mJ/cm^2 and $E_1 = 45$ mJ/cm^2:

$$\gamma = \left[\ln\left(\frac{E_1}{E_T}\right)\right]^{-1} = \left[\ln\left(\frac{90}{45}\right)\right]^{-1} = 1.4.$$

For the negative resist, $E_T = 7$ mJ/cm^2 and $E_1 = 12$ mJ/cm^2:

$$\gamma = \left[\ln\left(\frac{E_T}{E_1}\right)\right]^{-1} = \left[\ln\left(\frac{12}{7}\right)\right]^{-1} = 1.9.$$

For deep UV lithography (e.g., 248 and 193 nm), we cannot use conventional photoresists because these resists require a high-dose exposure in deep UV, which will cause lens damage and lower throughput. The chemical-amplified resist (CAR) has been developed for the deep UV process. CAR consists of a photo-acid generator, a resin polymer, and a solvent. CAR is very sensitive to deep UV radiation and the exposed and unexposed regions differ greatly in their solubility in the developer solution.

12.1.5 Pattern Transfer

Figure 10 illustrates the steps to transfer IC patterns from a mask to a silicon wafer that has an insulating SiO_2 layer formed on its surface.[8] The wafer is placed in a clean room, which, typically, is illuminated with yellow light, since photoresists are not sensitive to wavelengths greater than 0.5 μm. To ensure satisfactory adhesion of the resist the surface must be changed from hydrophilic to hydrophobic. This change can be made by the application of an adhesion promoter, which can provide a chemically compatible surface for the resist. The most common adhesion promoter for silicon ICs is hexa-methylene-di-siloxane (HMDS). After the application of this adhesion layer, the wafer is held on a vacuum spindle, and 2–3 cc of liquidous resist is applied to the center of wafer. The wafer is then rapidly accelerated up to a constant rotational speed, which is maintained for about

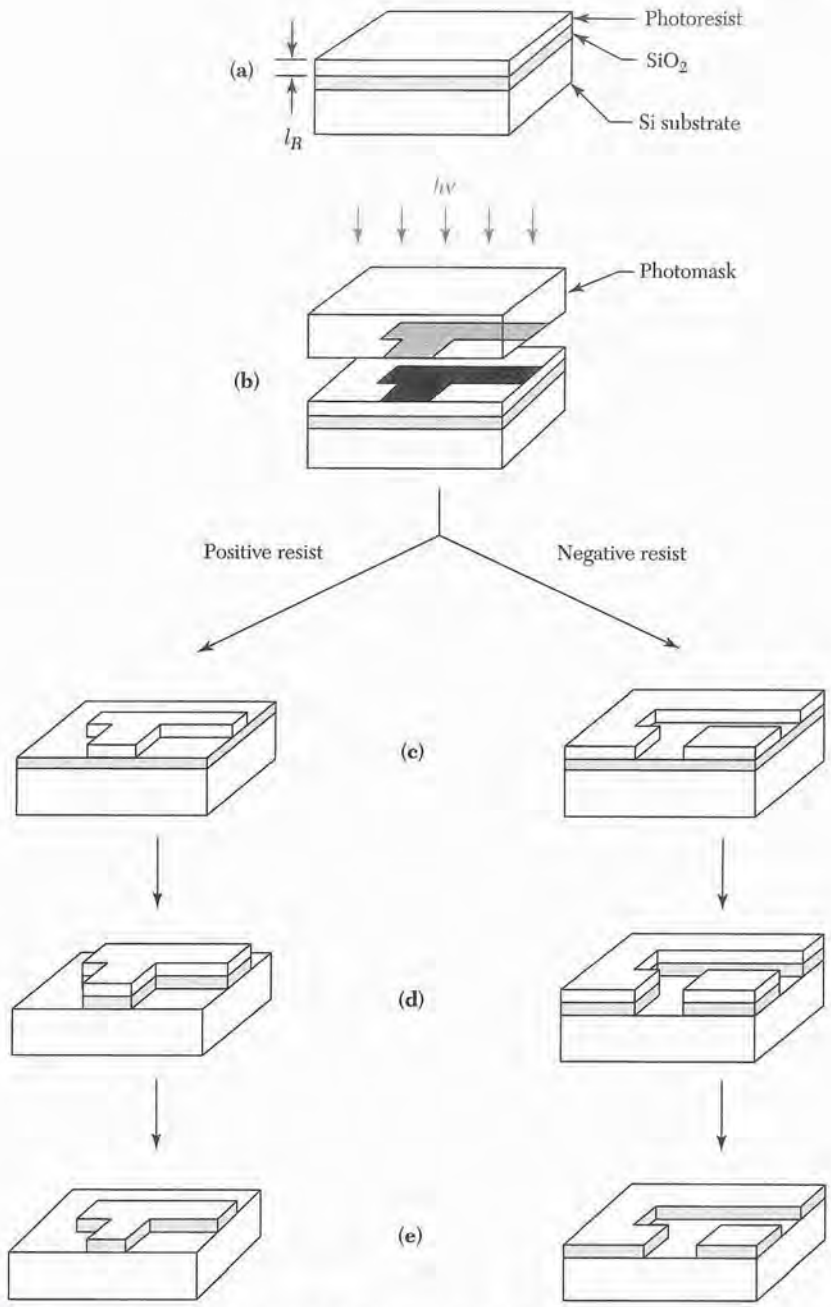


Fig. 10 Details of the optical lithographic pattern transfer process.[8]

30 seconds. Spin speed is generally in the range of 1000–10,000 rpm to coat a uniform film about 0.5 to 1 μm thick, as shown in Fig. 10*a*. The thickness of photoresist is correlated with its viscosity.

After the spinning step, the wafer is given a soft bake (typically at 90°–120°C for 60–120 seconds) to remove the solvent from the photoresist film and to increase resist adhesion

to the wafer. The wafer is aligned with respect to the mask in an optical lithographic system, and the resist is exposed to UV light, as shown in Fig. 10*b*. If a positive photoresist is used, the exposed resist is dissolved in the developer, as shown in the left side of Fig. 10*c*. The photoresist development is usually done by flooding the wafer with the developer solution. The wafer is then rinsed and dried. After development, a postbaking at ~100°–180°C may be required to increase the adhesion of the resist to the substrate. The wafer is then put in an ambient that etches the exposed insulation layer but does not attack the resist, as shown in Fig. 10*d*. Finally, the resist is stripped (e.g., using solvent or plasma oxidation), leaving behind an insulator image (or pattern) that is the same as the opaque image on the mask (left side of Fig. 10*e*).

For the negative photoresist, the procedures described are also applicable, except that the unexposed areas are removed. The final insulator image (right side of Fig. 10*e*) is the reverse of the opaque image on the mask.

The insulator image can be used as a mask for subsequent processing. For example, an ion implantation can be done to dope the exposed semiconductor region, but not the area covered by the insulator. The dopant pattern is a duplicate of the design pattern on the photomask for a negative photoresist or is its complementary pattern for a positive photoresist. The complete circuit is fabricated by aligning the next mask in the sequence to the previous pattern and repeating the lithographic transfer process.

A related pattern transfer process is the liftoff technique, shown in Fig. 11. A positive resist is used to form the resist pattern on the substrate (Fig. 11*a* and 11*b*). The film (e.g., aluminum) is deposited over the resist and the substrate (Fig. 11*c*); the film thickness must be smaller than that of the resist. Those portions of the film on the resist are removed by selectively dissolving the resist layer in an appropriate liquid etchant so that the overlying film is lifted off and removed (Fig. 11*d*). The liftoff technique is capable of high resolution and is used extensively for discrete devices such as high-power

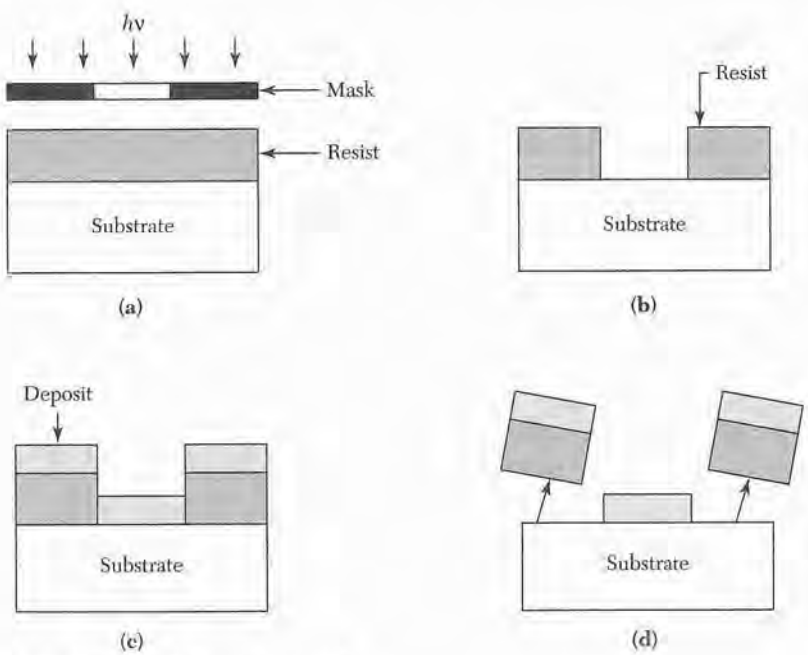


Fig. 11 Liftoff process for patter transfer.

MESFETs. However, it is not as widely applicable for ultralarge-scale integration, in which dry etching is the preferred technique.

12.1.6 Resolution Enhancement Techniques

Optical lithography has been continuously challenged to provide better resolution, greater depth of focus (DOF), and wider exposure latitude in IC processing. These challenges have been met by reducing the wavelength of the exposure tools and developing new resists. In addition, many resolution–enhancement techniques have been developed to extend the capability of optical lithography to even smaller feature lengths.

An important–resolution enhancement technique is the phase-shifting mask (PSM). The basic concept is shown[9] in Fig. 12. For a conventional mask, the electric field has the same phase at every aperture (clear area) in Fig. 12*a*. Diffraction and the limited resolution of the optical system spread the electric field at the wafer, as shown by the dotted lines. Interference between waves diffracted by the adjacent apertures enhances the field between them. Because the intensity I is proportional to the square of the electric field, it becomes difficult to separate the two images that are projected close to one another. The phase-shift layer that covers adjacent apertures reverses the sign of the electric field, as shown in Fig. 12*b*. Because the intensity at the mask is unchanged, the electric field of the images at the wafer can be cancelled. Therefore, images that are projected close to one another can be separated. A 180° phase change can be obtained by using a trans parent layer of thickness $d = \lambda / 2(\bar{n} - 1)$, where $\bar{n}$ is the refractive index and λ is the wavelength, which covers one aperture, as shown in Fig. 12*b*.

Another resolution-enhancement technique is the optical proximity correction (OPC), which uses modified shapes of adjacent subresolution geometry to improve imaging capability. For example, a square contact hole with dimensions near the resolution limit

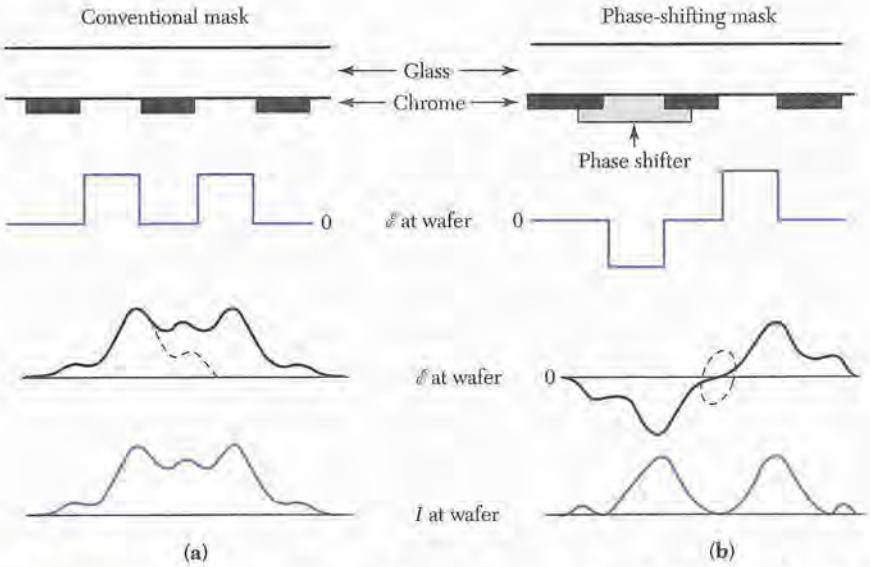


Fig. 12 The principle of phase-shift technology. (*a*) Conventional technology; (*b*) phase-shift technology.[9]

will print nearly as a circle. Modifying the contact-hole pattern with additional geometrics at the corners will help to print a more accurate square hole.

12.2 NEXT-GENERATION LITHOGRAPHIC METHODS

Why is optical lithography so widely used and what makes it such a promising method? The reasons are that it has high throughput, good resolution, low cost, and ease in operation.

However, due to deep-submicron IC process requirements, optical lithography has some limitations not yet solved. Although we can use PSM or OPC to extend its useful span, the complexity of mask production and mask inspection can not be easily resolved. In addition, the cost of the masks is very high. Therefore, we need to find postoptical lithography to process deep-submicron or even nanometer ICs.

Various types of next-generation lithographic methods for IC fabrication are discussed in this section. We consider electron-beam lithography, extreme UV lithography, X-ray lithography, and ion-beam lithography. We also consider the differences among these lithographic methods.

12.2.1 Electron-Beam Lithography

Electron-beam lithography is primarily used to produce photomasks. Relatively few tools are dedicated to direct exposure of the resist by a focused electron beam without a mask. Figure 13 shows a schematic of an electron-beam lithography system.[10] The electron gun is a device that can generate a beam of electrons with a suitable current density. A tung-

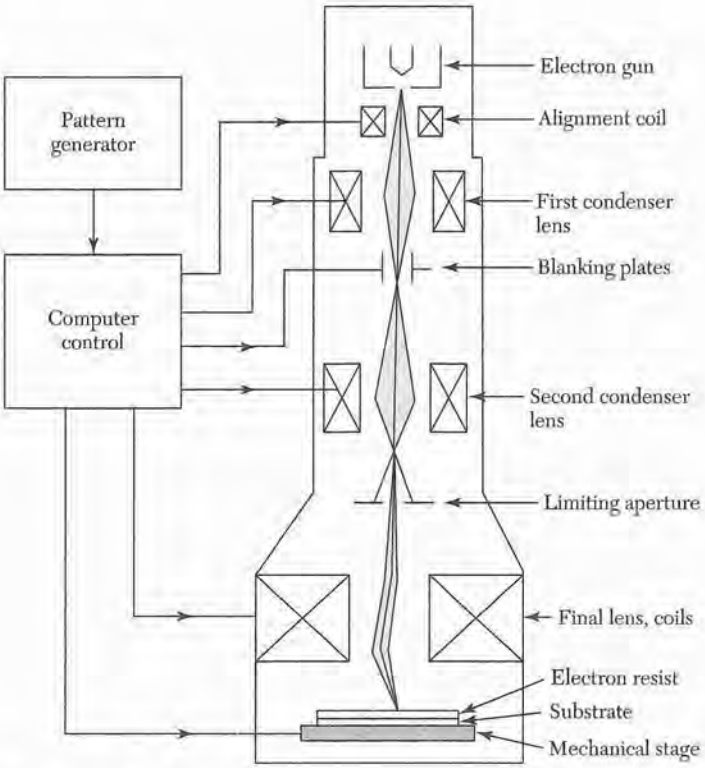


Fig. 13 Schematic of an electron-beam machine.[10]

sten thermionic-emission cathode or single-crystal lanthanum hexa-boride (LaB_6) is used for the electron gun. Condenser lenses are used to focus the electron beam to a spot size 10–25 nm in diameter. Beam-blanking plates that turn the electron beam on and off, and beam deflection coils are computer controlled and operated at MHz or higher rates to direct the focused electron beam to any location in the scan field on the substrate. Because the scan field (typically 1 cm) is much smaller than the substrate diameter, a precision mechanical stage is used to position the substrate to be patterned.

The advantages of electron-beam lithography include the generation of submicron resist geometries, highly automated and precisely controlled operation, greater depth of focus than that available from optical lithography, and direct patterning on a semiconductor wafer without using a mask. The disadvantage is that electron beam lithographic machines have low throughput—approximately 10 wafers per hour at less than 0.25 μm resolution. This throughput is adequate for the production of photomasks, for situations that require small numbers of custom circuits, and for design verification. However, for maskless direct writing, the machine must have the highest possible throughput and therefore the largest beam diameter possible consistent with the minimum device dimensions.

There are basically two ways to scan the focused electron beam: raster scan and vector scan.[11] In a raster scan system, resist patterns are written by a beam that moves through a regular mode,vertically oriented, as shown in Fig. 14*a*. The beam scans sequentially over every possible location on the mask and is blanked (turned off) where no exposure is required. All patterns on the area to be written must be subdivided into individual addresses, and a given pattern must have a minimum incremental interval that is evenly divisible by the beam address size.

In the vector scan system, as shown in Fig. 14*b*, the beam is directed only to the requested pattern features and jumps from feature to feature, rather than scanning the whole chip, as in raster scan. For many chips, the average exposed region is only 20% of the chip area, so we can save time by using a vector scan system.

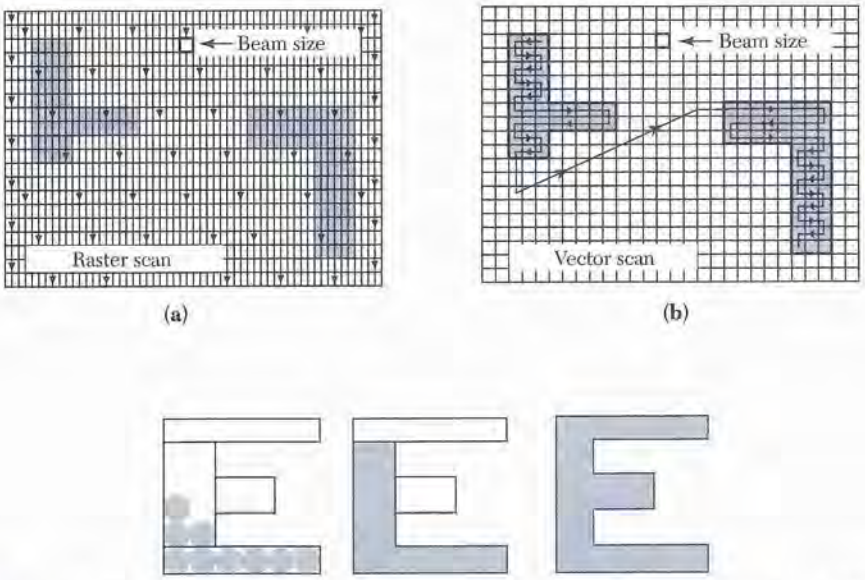


Fig. 14 (*a*) Raster scan writing scheme; (*b*) vector scan writing schemes; and (*c*) shapes of electron beam: round, variable, cell projection.[12]

Figure 14*c* shows several types of electron beam employed for e-beam lithography: Gaussian spot beam (round beam), variable-shaped beam, and cell projection. In variable-shaped beam system, the patterning beam has a rectangular cross section of variable size and aspect ratio. It offers the advantage of exposing several address units simultaneously. Therefore, the vector scan method using variable-shaped beam has higher throughout than the conventional Gaussian sport beam. It is also possible to pattern a complex geometric shape in one exposure with an electron beam system; this is called cell projection as shown in the far right of Fig. 14*c*. The cell projection technique[12] is particularly suitable for highly repetitive designs, as in an MOS memory cells, since several memory cell patterns can be exposed at once. Cell projection has not yet achieved the throughput of optical exposure tools.

SCALPEL System

A new approach for electron-beam technology is the electron-beam *projection* lithography. An example is the SCALPEL (Scattering with *a*ngular *l*imitation *p*rojection *e*lectron-beam *l*ithography) system.[13] SCALPEL combines high resolution with the wide process latitude inherent in electron-beam lithography (depth of focus about 20–30 μm compared with <1 μm for KrF lithography) and the high throughput of a projection system. SCALPEL's masks are expected to be considerably less expensive than optical masks, which require optical proximity correction, and phase-shift masks for the 100 nm generation, because SCALPEL adopts a 4× projection reduction.

The system has an initial design for a 3 mm field, which would be capable of delivering a throughput of 200 mm wafers at a rate of 30 per hour. As shown in Fig. 15, SCALPEL uses a scattering mask to conduct the reduction projection lithography through a scanner. The mask is uniformly illuminated by a parallel beam of 100 keV electrons

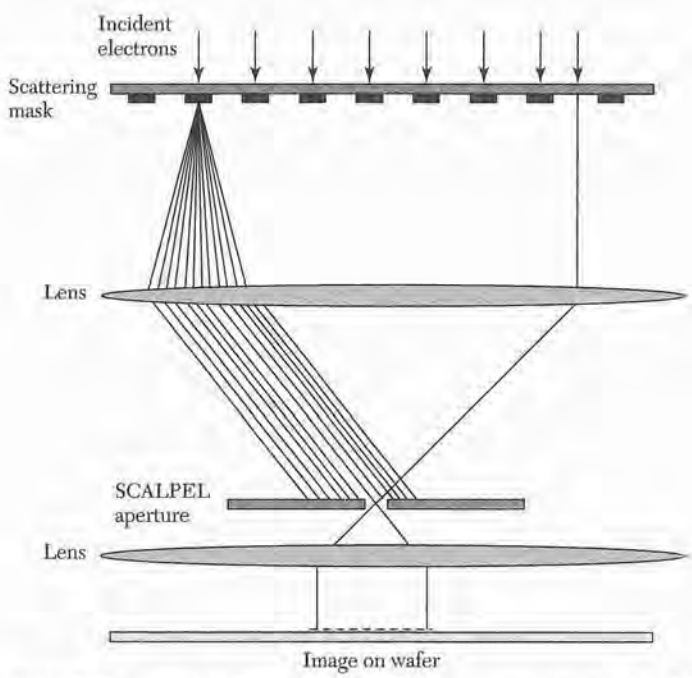


Fig. 15 Basic SCALPEL principle of operation showing contrast generation by differentiating more- or less-scattered electrons.[13]

over an area 1 mm × 1 mm. The key aspect of SCALPEL that differentiates it from previous attempts at projection electron-beam lithography is the mask design. The mask is comprised of a thin (100–150 nm), low-atomic-number membrane (SiN_x), and a thin (30–60 nm) high-atomic-number pattern layer (Cr/W).

Although the mask is almost completely electron transparent at the energies used (100 keV), contrast is generated by utilizing the difference in electron-scattering characteristics between the membrane and patterned materials. The membrane of low-atomic-number material scatters electrons weakly with small angles, whereas the pattern layer of high-atomic-number material scatters most electrons strongly with high angles. A SCALPEL aperture in the back-focal plane of the projection optics blocks the strongly scattered electrons, forming a high-contrast aerial image on the wafer.

Electron Resist

Electron resists are polymers. The behavior of an electron-beam resist is similar to that of a photoresist, that is, a chemical or physical change is induced in the resist by irradiation. This change allows the resist to be patterned. For a positive electron resist, the polymer-electron interaction causes chemical bonds to be broken (chain scission) to form shorter molecular fragments, as shown[14] in Fig. 16*a*. As a result, the molecular weight is reduced in the irradiated area, which can be dissolved subsequently in a developer solution that attacks the low-molecular-weight material. Common positive electron resists include poly-methyl methacrylate (PMMA) and poly-butene-1 sulfone (PBS). Positive electron resists can achieve resolution of 0.1 μm or better.

For a negative electron resist, the irradiation causes radiation-induced polymer linking, as shown in Fig. 16*b*. The cross linking creates a complex three-dimensional structure with a molecular weight higher than that of the nonirradiated polymer. The nonirradiated resist can be dissolved in a developer solution that does not attack the high-molecular-weight material. Poly-glycidyl methacrylate-co-ethyl acrylate (COP) is a common

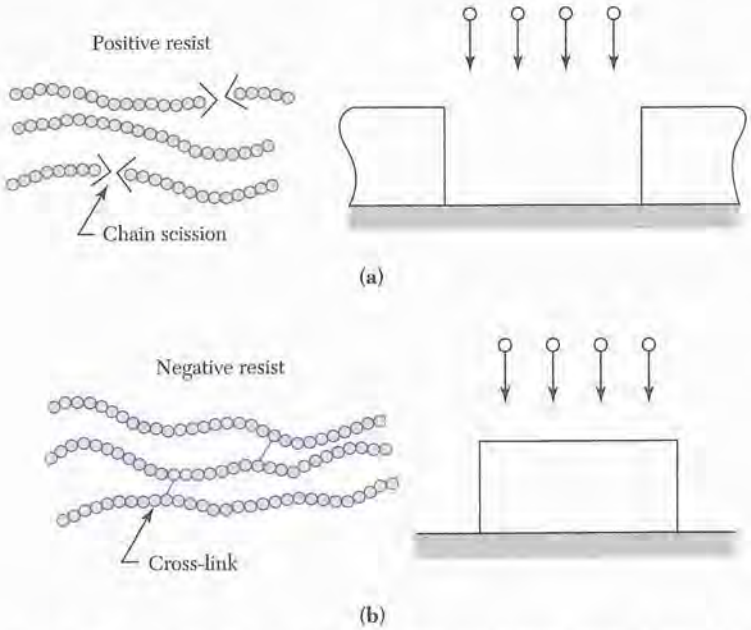


Fig. 16 Schematic of positive and negative resists used in electron-beam lithography.[14]

negative electron resist. COP, like most negative photoresists, also swells during development so the resolution is limited to about 1 μm.

The Proximity Effect

In optical lithography, the resolution is limited by diffraction of light. In electron-beam lithography, the resolution is not limited by diffraction (because the wavelengths associated with electrons of a few keV and higher energies are less than 0.1 nm) but by electron scattering. When electrons penetrate the resist film and underlying substrate, they undergo collisions. These collisions lead to energy losses and path changes. Thus, the incident electrons spread out as they travel through the material until either all of their energy is lost or they leave the material because of backscattering.

Figure 17*a* shows computed electron trajectories of 100 electrons with initial energy of 20 keV incident at the origin of a 0.4 μm PMMA film on a thick silicon substrate.[15] The electron beam is incident along the z-axis, and all trajectories have been projected onto the xz plane. This figure shows qualitatively that the electrons are distributed in an oblong pear-shaped volume with a diameter on the same order of magnitude as the electron penetration depth (~3.5 μm). Also, there are many electrons that undergo backscat-

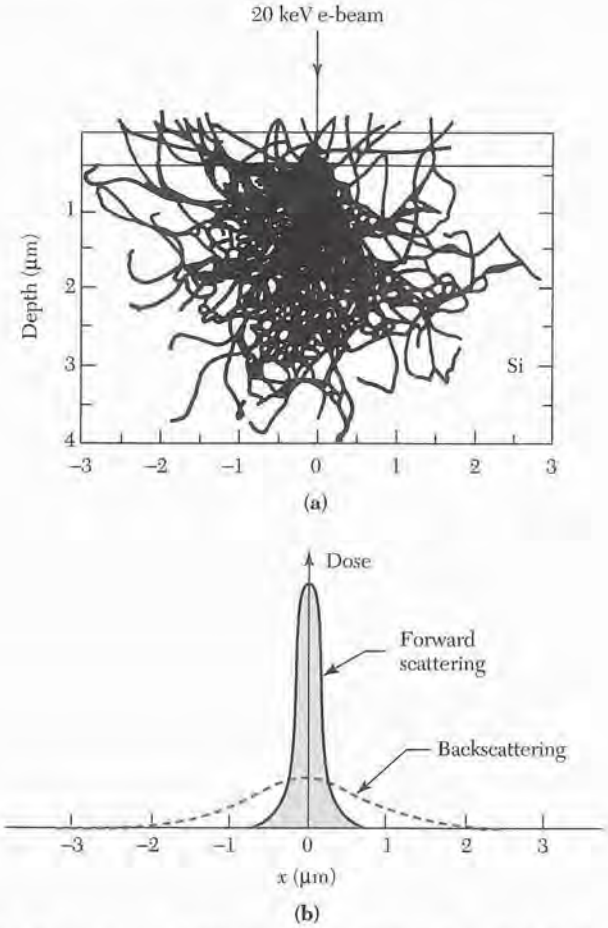


Fig. 17 (*a*) Simulated trajectories of 100 electrons in PMMA for a 20-keV electron beam.[15] (*b*) Dose distribution for forward scattering and backscattering at the resist-substrate interface.

tering collisions and travel backward from the silicon substrate into the PMMA resist film and leave the material.

Figure 17*b* shows the normalized distributions of the forward scattering and backscattering electrons at the resist-substrate interface. Because of the backscattering, electrons effectively can irradiate several micrometers away from the center of the exposure beam. Since the dose of a resist is given by the sum of the irradiations from all surrounding areas, the electron-beam irradiation at one location will affect the irradiation in neighboring locations. This phenomenon is called the *proximity effect*. The proximity effect places a limit on the minimum spacings between pattern features. To correct for the proximity effect, patterns are divided into smaller segments. The incident electron dose in each segment is adjusted so that the integrated dose from all its neighboring segments is the correct exposure dose. This approach further decreases the throughput of the electron-beam system,because of the additional computer time required to expose the subdivided resist patterns.

12.2.2 Extreme-Ultraviolet Lithography

Extreme-ultraviolet (EUV) lithography is a promising next-generation lithographic technology to extend the minimum linewidths to 30 nm without throughput loss.[16] Figure 18 shows a schematic diagram of an EUV lithographic system. A laser-produced plasma or a synchrotron radiation can serve as the EUV source of λ = 10–14 nm EUV light. The EUV radiation is reflected by a mask that is produced by patterning an absorber material deposited on a multilayer-coated flat silicon or glass-plate mask blank. EUV radiation is reflected from the nonpatterned regions (i.e., nonabsorbing regions) of the mask through a 4× reduction camera and imaged into a thin layer of resist on the wafer.

Since the EUV radiation beam is narrow, the mask must be scanned by the beam to illuminate the entire pattern field that describes the circuit mask layer. Also, for a 4× four-mirror (i.e., the one-paraboloid, two-ellipsoid, and one-plane mirrors) reduction camera, the wafer must be scanned at one-fourth the mask speed in a direction opposite to the mask movement to reproduce the image field on all chip sites on the wafer surface. A

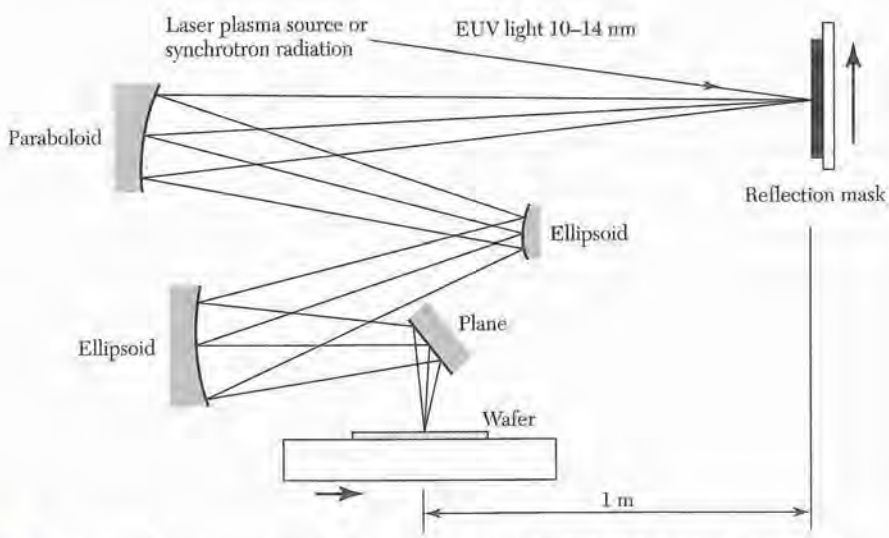


Fig. 18 Schematic representation of an extreme-ultraviolet (EUV) lithography system.[16]

precision system is required to perform the chip-site alignment and to control the wafer and mask stage movements and the exposure dose during the scanning process.

EUV lithography is capable of printing 50-nm features with PMMA resist using 13-nm radiation. However, the production of EUV exposure tools has a number of challenges. Since EUV is strongly absorbed in all materials, the lithography process must be performed in vacuum. The camera must use reflective-lens elements, and the mirrors must be coated with multilayer coatings that produce distributed quarter-wave Bragg reflectors. In addition, the mask blank must also be multilayer coated to maximize its reflectivity at $\lambda = 10$–14 nm.

12.2.3 X-Ray Lithography

X-ray lithography[17] (XRL) is a potential candidate to succeed optical lithography for the fabrication of integrated circuits at 100 nm. The synchrotron storage ring is the choice of X-ray source for high-volume manufacturing. It can provide a large amount of collimated flux and can easily accommodate 10–20 exposure tools.

XRL uses a shadow printing method similar to optical proximity printing. Figure 19 shows a schematic XRL system. The X-ray wavelength is about 1 nm, and the printing is through a 1× mask in close proximity (10–40 μm) to the wafer. Since X-ray absorption depends on the atomic number of the material and most materials have low transparency at $\lambda \cong 1$ nm, the mask substrate must be a thin membrane (1–2 μm thick) made of low-atomic-number material, such as silicon carbide or silicon. The pattern itself is defined in a thin (~ 0.5 μm) relatively high-atomic-number material, such as tantalum, tungsten, gold, or one of their alloys, which is supported by the thin membrane.

Masks are the most difficult and critical element of an XRL system, and the construction of an X-ray mask is much more complicated than that of a photomask. To avoid absorption of the X-rays between the source and mask, the exposure generally takes place in a helium environment. The X rays are produced in vacuum, which is separated from the helium by a thin vacuum window (usually of beryllium). The mask substrate will absorb 25%–35% of the incident flux and must therefore be cooled. An X-ray resist 1 μm thick

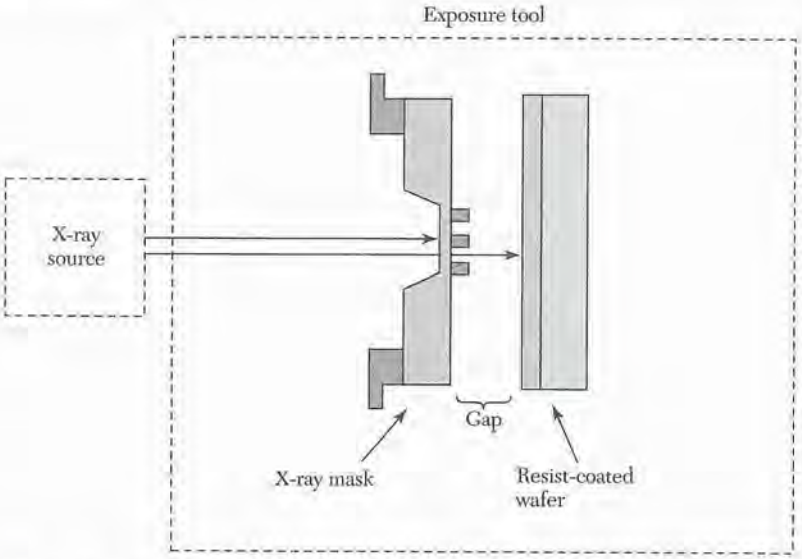


Fig. 19 Schematic representation of a proximity X-ray lithography system.[17]

will absorb about 10% of the incident flux, and there are no reflections from the substrate to create standing waves, so antireflection coatings are unnecessary.

We can use electron beam resists as X-ray resists because when an X-ray is absorbed by an atom, the atom goes to an excited state with the emission of an electron. The excited atom returns to its ground state by emitting an X-ray having a different wavelength than the incident X-ray. This X-ray is absorbed by another atom, and the process repeats. Since all the processes result in the emission of electrons, a resist film under X-ray irradiation is equivalent to one being irradiated by a large number of secondary electrons from any of the other processes. Once the resist film is irradiated, chain cross linking or chain scission will occur, depending on the type of resist.

12.2.4 Ion-Beam Lithography

Ion-beam lithography can achieve higher resolution than optical, X-ray, or electron-beam lithographic techniques because ions have a higher mass and therefore scatter less than electrons. The most important application is the repair of masks for optical lithography, a task for which commercial systems are available.

Figure 20 shows the computer-simulated trajectories of 50 H^+ ions implanted at 60 keV into PMMA and various substrates.[18] Note that the spread of the ion beam at a depth of 0.4 μm is only 0.1 μm in all cases (compare with Fig. 17*a* for electrons). The backscattering is completely absent for the silicon substrate, and there is only a small amount of backscattering for the gold substrate. However, ion-beam lithography may suffer from random (or stochastic) space-charge effect, causing broadening of the ion beam.

There are two types of ion-beam lithography systems: a scanning focused-beam system and a mask-beam system. The former system is similar to the electron-beam machine (Fig. 13), in which the ion source can be Ga^+ or H^+. The latter system is similar to an optical 5× reduction projection step-and-repeat system, which projects 100 keV light ions such as H^+_2 through a stensil mask.

12.2.5 Comparison of Various Lithographic Methods

The lithographic methods discussed above all have 100 nm or better resolution. A comparison of various lithographic technologies is shown in Table 1. However, each method

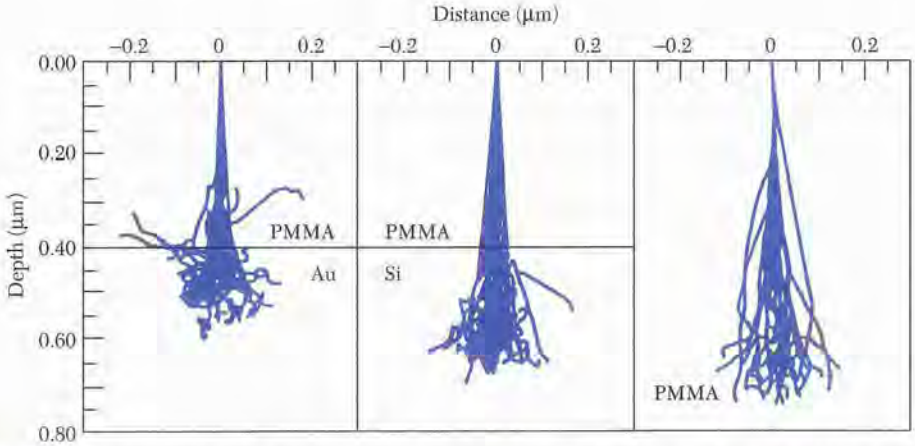


Fig. 20 Trajectories of 60 keV H^+ ions traveling through PMMA into Au, Si, and PMMA.[18]

TABLE 1 Comparison of Various Lithographic Technologies

		Optical 248/193 nm	SCALPEL[a]	EUV[a]	X-ray	Ion beam
Exposure Tool	Source	Laser	Filament	Laser plasma	Synchrotron	Multicusp
	Diffraction limited	Yes	No	Yes	Yes	No
	Optics	Refractive	Refractive	Refractive	No optics	Full-field refractive
	Step and scan	Yes	Yes	Yes	Yes	Stepper
	Throughput of 200 mm wafers/hr	40	30–35	20–30	30	30
Mask	Demagnification	4X	4X	4X	1X	4X
	Optical proximity correction	Yes	No	Yes	Yes	No
	Radiation path	Transmission	Transmission	Reflection	Transmission	Stencil
Resist	Single or Multilayer	Single	Single	Surface Imaging	Single	Single
	Chemical amplified resist	Yes	Yes	No	Yes	No

[a]SCALPEL, Scattering with *a*ngular *l*imitation *p*rojection *e*lectron-beam *l*ithoraph; EUV, extreme ultraviolet.

has its own limitations: the diffraction effect in optical lithography, the proximity effect in electron-beam lithography, mask fabrication complexities in X-ray lithography, difficulty in mask blank production for EUV lithography, and stochastic space charge in ion projection lithography.

For IC fabrication, many mask levels are involved. However, it is not necessary to use the same lithographic method for all levels. A mix-and-match approach can take advantage of the unique features of each lithographic process to improve resolution and to maximize throughput. For example, a 4:1 SCALPEL or EUV method can be used for the most critical mask levels, whereas 4:1 or 5:1 optical system can be used for the rest.

According to the Roadmap of Semiconductor Industry Association, IC manufacturing technology will reach the 50 nm generation around 2010. With each new technology generation, lithography has become an even more important key driver for the semiconductor industry because of the requirements of smaller feature size and tighter overlay tolerance. In addition, lithography tool costs have become higher relative to the total equipment costs for IC manufacturing facility. Currently, the technology development of next-generation lithography is conducted by multinational research projects or by industrial partners.

12.3 WET CHEMICAL ETCHING

Wet chemical etching is used extensively in semiconductor processing. Starting from the sawed semiconductor wafers, chemical etchants are used for lapping and polishing to give

an optically flat, damage-free surface. Prior to thermal oxidation or epitaxial growth, the semiconductor wafers are chemically cleaned to remove contamination that results from handling and storing. Wet chemical etchings are especially suitable for blanket etches (i.e., over the whole wafer surface) of polysilicon, oxide, nitride, metals, and III-V compounds.

The mechanisms for wet chemical etching involve three essential steps, as illustrated in Fig. 21: the reactants are transported by diffusion to the reacting surface, chemical reactions occur at the surface, and the products from the surface are removed by diffusion. Both agitation and the temperature of the etchant solution will influence the etch rate, which is the amount of film removed by etching per unit time. In IC processing, most wet chemical etchings proceed by immersing the wafers in a chemical solution or by spraying the wafers with the etchant solution. For immersion etching, the wafer is immersed in the etch solution, and mechanical agitation is usually required to ensure etch uniformity and a consistent etch rate. Spray etching has gradually replaced immersion etching because it greatly increases the etch rate and uniformity by constantly supplying fresh etchant to the wafer surface.

For semiconductor production lines, highly uniform etch rates are important. Etch rates must be uniform across a wafer, from wafer to wafer, from run to run, and for any variations in feature sizes and pattern densities. Etch rate uniformity is given by:

$$\text{Etch rate uniformity } (\%) = \frac{(\text{maximum etch rate} - \text{minimum etch rate})}{\text{maximum etch rate} + \text{minimum etch rate}} \times 100\%. \qquad (8)$$

EXAMPLE 3

Calculate the Al average etch rate and etch rate uniformity on a 200 mm diameter silicon wafer, assuming the etch rates at the center, left, right, top, and bottom of the wafer are 750, 812, 765, 743, and 798 nm/min, respectively.

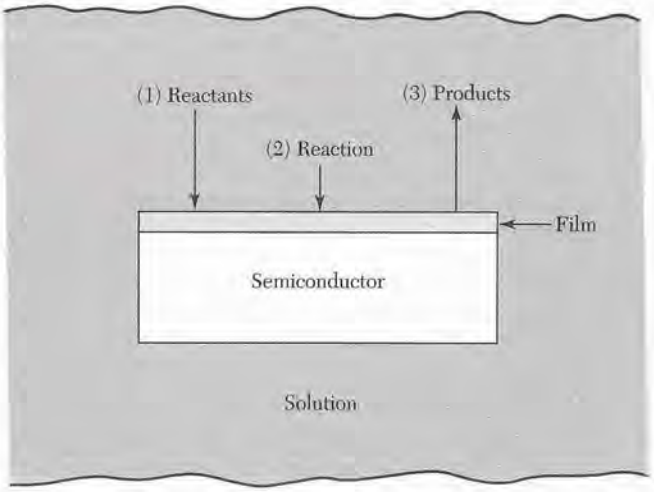


Fig. 21 Basic mechanisms in wet chemical etching.

SOLUTION

Al average etch rate = (750 + 812 + 765 + 743 + 798) ÷ 5 = 773.6 nm/min.

Etch rate uniformity = (812 − 743) ÷ (812 + 743) × 100% = 4.4 %.

12.3.1 Silicon Etching

For semiconductor materials, wet chemical etching usually proceeds by oxidation, followed by the dissolution of the oxide by a chemical reaction. For silicon, the most commonly used etchants are mixtures of nitric acid (HNO_3) and hydrofluoric acid (HF) in water or acetic acid (CH_3COOH). Nitric acid oxidizes silicon to form a SiO_2 layer.[19] The oxidation reaction is

$$Si + 4HNO_3 \rightarrow SiO_2 + 2H_2O + 4NO_2. \tag{9}$$

Hydrofluoric acid is used to dissolve the SiO_2 layer. The reaction is:

$$SiO_2 + 6HF \rightarrow H_2SiF_6 + 2H_2O. \tag{10}$$

Water can be used as a diluent for this etchant. However, acetic acid is preferred because it reduces the dissolution of the nitric acid.

Some etchants dissolve a given crystal plane of single-crystal silicon much faster than another plane; this results in orientation-dependent etching.[20] For a silicon lattice, the (111)-plane has more available bonds per unit area than the (110)- and (100)-planes; therefore, the etch rate is expected to be slower for the (111)-plane. A commonly used orientation-dependent etch for silicon consists of a mixture of KOH in water and isopropyl alcohol. For example, a solution with 19 wt % KOH in deionized (DI) water at about 80°C removes the (100)-plane at a much higher rate than the (110)- and (111)-planes. The ratio of the etch rates for the (100)-, (110)-, and (111)-planes is 100:16:1.

Orientation-dependent etching of <100>-oriented silicon through a patterned silicon dioxide mask creates precise V-shaped grooves,[10] the edges being (111)-planes at an angle of 54.7° from the (111)-surface, as shown at the left of Fig. 22*a*. If the window in the mask is sufficiently large or if the etching time is short, a U-shaped groove will be formed, as shown at the right of Fig. 22*a*. The width of the bottom surface is given by

$$W_b = W_0 - 2l\cot 54.7°$$

or

$$\boxed{W_b = W_0 - \sqrt{2}\ l,} \tag{11}$$

where W_0 is the width of the window on the wafer surface and l is the etched depth. If <$\bar{1}$10>-oriented silicon is used, essentially straight-walled grooves with sides of (111)-planes can be formed, as shown in Fig. 22*b*. We can use the large orientation dependence in the etch rates to fabricate device structures with submicron feature lengths.

12.3.2 Silicon Dioxide Etching

The wet etching of silicon dioxide is commonly achieved in a dilute solution of HF with or without the addition of ammonium fluoride (NH_4F). Adding NH_4F is referred to as a buffered HF solution (BHF), also called buffered-oxide-etch (BOE). The addition of NH_4F to HF controls the pH value and replenishes the depletion of the fluoride ions,

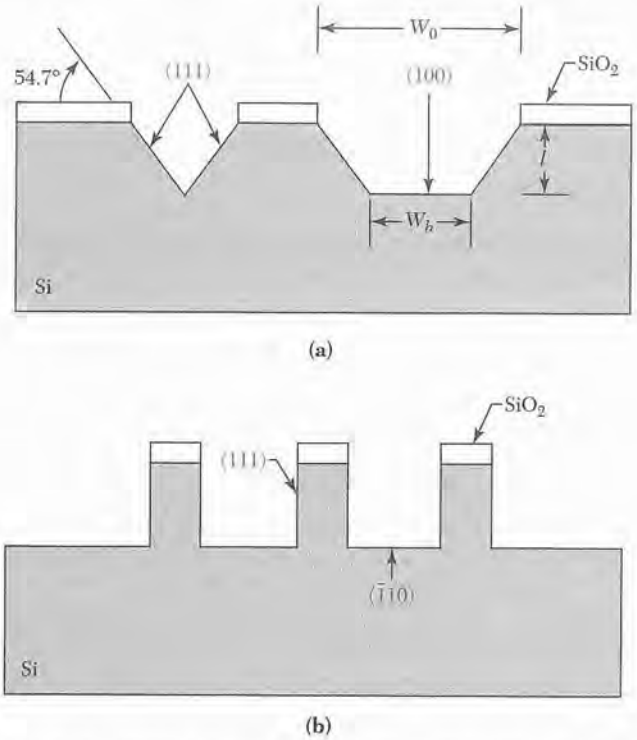


Fig. 22 Orientation-dependent etching. (*a*) Through window patterns on <100>-oriented silicon; (*b*) through window patterns on <$\bar{1}$10>-oriented silicon.[20]

thus maintaining stable etching performance. The overall reaction for SiO_2 etching is the same as that in Eq. 10. The etch rate of SiO_2 etching depends on etchant solution, etchant concentration, agitation, and temperature. In additional, density, porosity, microstructure, and the presence of impurities in the oxide also influence the etch rate. For example, a high concentration of phosphorus in the oxide results in a rapid increase in the etch rate, and a loosely structured chemical-vapor deposition (CVD) or sputtered oxide exhibits a faster etch rate than thermally grown oxide.

Silicon dioxide can also be etched in vapor-phase HF. Vapor-phase-HF oxide-etch technology has a potential for submicron feature etching because the process can be well controlled.

12.3.3 Silicon Nitride and Polysilicon Etching

Silicon nitride films are etchable at room temperature in concentrated HF or buffered HF and in a boiling H_3PO_4 solution. Selective etching of nitride to oxide is done with 85% H_3PO_4 at 180°C because this solution attacks silicon dioxide very slowly. The etch rate is typically 10 nm/min for silicon nitride, but less than 1 nm/min for silicon dioxide. However photoresist adhesion problems are encountered when etching nitride with boiling H_3PO_4 solution. Better patterning can be achieved by depositing a thin oxide layer on top of the nitride film before resist coating. The resist pattern is transferred to the oxide layer, which then acts as a mask for subsequent nitride etching.

Etching polysilicon is similar to etching single-crystal silicon. However, the etch rate is considerably faster because of grain boundaries. The etch solution is usually modified to ensure that it does not attack the underlying gate oxide. Dopant concentrations and temperature may affect the etch rate of polysilicon.

12.3.4 Aluminum Etching

Aluminum and aluminum alloy films are generally etched in heated solutions of phosphoric acid, nitric acid, acetic acid, and DI water. The typical etchant is a solution of 73% H_3PO_4, 4% HNO_3, 3.5% CH_3COOH, and 19.5% DI water at 30°–80°C. The wet etching of aluminum proceeds as follows: HNO_3 oxidizes aluminum, and H_3PO_4 then dissolves the oxidized aluminum. The etch rate depends on etchant concentration,

TABLE 2 Etchants for Insulators and Conductors

Material	Etchant composition	Etch rate (nm/min)
SiO_2	28 ml of HF 170 ml of HF } Buffered HF 113 g of NH_4F	100
	15 ml of HF 10 ml of HNO_3 } P – etch 300 ml of H_2O	12
Si_3N_4	Buffered HF	0.5
Al	H_3PO_4	10
	4 ml of HNO_3 3.5 ml of CH_3COOH 73 ml of H_3PO_4 19.5 ml of H_2O	30
Au	4 g KI	1000
Mo	1 g of I_2 40 ml of H_2O 5 ml of H_3PO_4 2 ml of HNO_3 4 ml of CH_3COOH 150 ml of H_2O	500
Pt	1 ml of HNO_3 7 ml of HCl 8 ml of H_2O	50
W	34 g of KH_2PO_4 13.4 g of KOH 33 g of $K_3Fe(CN)_6$ H_2O to make 1 liter	160

temperature, agitation of the wafers, and impurities or alloys in the aluminum film. For example, the etch rate is reduced when copper is added to the aluminum.

Wet etching of insulating and metal films is usually done with the similar chemicals that dissolve these materials in bulk form and involve their conversion into soluble salts or complexes. Generally, film materials will be etched more rapidly than their bulk counterparts. Also, the etch rates are higher for films that have a poor microstructure, built-in stress, or departure from stoichiometry, or that have been irradiated. Some useful etchants for insulating and metal films are listed in Table 2.

12.3.5 Gallium Arsenide Etching

A wide variety of etches has been investigated for gallium arsenide; however, few of them are truly isotropic.[21] This is because the surface activities of the (111)-Ga and (111)-As faces are very different. Most etches give a polished surface on the arsenic face, but the gallium face tends to show crystallographic defects and etches more slowly. The most commonly used etchants are the H_2SO_4-H_2O_2-H_2O and H_3PO_4-H_2O_2-H_2O systems. For an etchant with an 8:1:1 volume ratio of H_2SO_4:H_2O_2:H_2O, the etch rate is 0.8 μm/min for the ⟨111⟩-Ga face and 1.5 μm/min for all other faces. For an etchant with 3:1:50 volume ratio of H_3PO_4:H_2O_2:H_2O, the etch rate is 0.4 μm/min for ⟨111⟩-Ga face and 0.8 μm for all other faces.

12.4 DRY ETCHING

In pattern-transfer operations, a resist pattern is defined by a lithographic process to serve as a mask for etching of its underlying layer (Fig. 23*a*).[22] Most of the layer materials (e.g., SiO_2, Si_3N_4, and deposited metals) are amorphous or polycrystalline thin films. If they are etched in a wet chemical etchant, the etch rate is generally isotropic (i.e., the lateral and vertical etch rates are the same), as illustrated in Fig. 23*b*. If h_f is the thickness of the layer material and l the lateral distance etched underneath the resist mask, we can define the degree of anisotropy A_f by

$$\boxed{A_f \equiv 1 - \frac{1}{h_f} = 1 - \frac{R_l t}{R_v t} = 1 - \frac{R_l}{R_v},} \tag{12}$$

where t is the time and R_l and R_v are the lateral and vertical etch rates, respectively. For isotropic etching, $R_l = R_v$ and $A_f = 0$.

The major disadvantage of wet chemical etching for pattern transfer is the undercutting of the layer underneath the mask, resulting in a loss of resolution in the etched pattern. In practice, for isotropic etching, the film thickness should be about one-third or less of the resolution required. If patterns are required with resolutions much smaller than the film thickness, anisotropic etching (i.e., $1 \geq A_f > 0$) must be used. In practice, the value of A_f is chosen to be close to unity. Figure 23*c* shows the limiting case where $A_f = 1$, corresponding to $l = 0$ (or $R_l = 0$).

To achieve a high-fidelity transfer of the resist patterns required for ultralarge-scale integration processing ($A_f = 1$), dry etching methods have been developed. Dry etching is synonymous with plasma-assisted etching, which denotes several techniques that use plasma in the form of low-pressure discharges. Dry-etch methods include plasma etching,

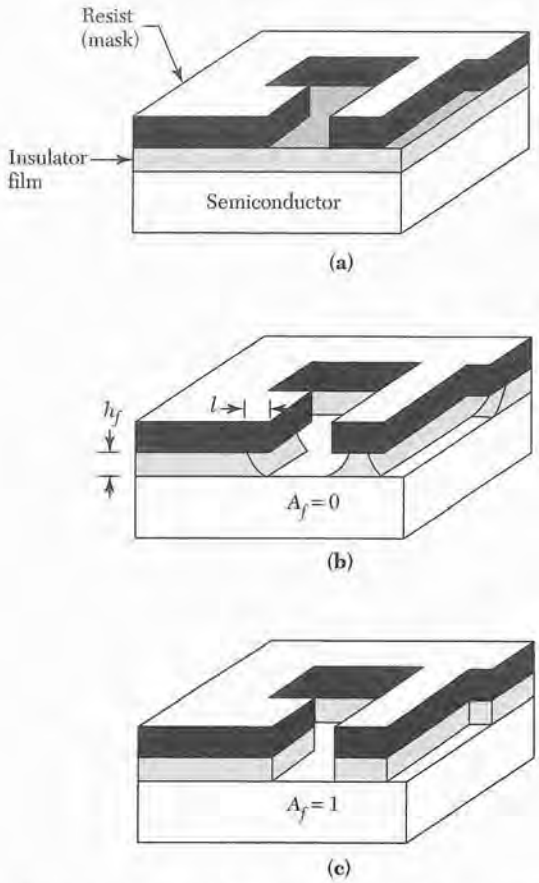


Fig. 23 Comparison of wet chemical etching and dry etching for pattern transfer. [22]

reactive ion etching (RIE), sputter etching, magnetically enhanced RIE (MERIE), reactive ion beam etching, and high-density plasma (HDP) etching.

12.4.1 Plasma Fundamentals

Plasma is a fully or partially ionized gas composed of equal numbers of positive and negative charges and a different number of unionized molecules. Plasma is produced when an electric field of sufficient magnitude is applied to a gas, causing the gas to break down and become ionized. The plasma is initiated by free electrons that are released by some means, such as field emission from a negatively biased electrode. The free electrons gain kinetic energy from the electric field. In the course of their travel through the gas, the electrons collide with gas molecules and lose their energy. The energy transferred in the collision causes the gas molecules to be ionized (i.e., to free electrons). The free electrons gain kinetic energy from the field, and the process continues. Therefore, when the applied voltage is larger than the breakdown potential, a sustained plasma is formed throughout the reaction chamber.

The electron concentrations in the plasma for dry etchings are relatively low, typically on the order of 10^9–10^{12} cm^{-3}. At a pressure of 1 Torr, the concentrations of gas molecules are 10^4–10^7 times higher than the electron concentrations. This results in an

average gas temperature in the range of 50°–100°C. Therefore, the plasma-assisted dry etching is a low-temperature process.

EXAMPLE 4

The electron densities in a RIE system and HDP system range from 10^9–10^{10} and 10^{11}–10^{12} cm^{-3}, respectively. Assuming the RIE chamber pressure is 200 mTorr and HDP chamber pressure is 5 mTorr, calculate the ionization efficiency in RIE reactors and HDP reactors at room temperature. The ionization efficiency is the ratio of the electron density to the density of molecules.

SOLUTION

$$PV = nRT$$

where P is the pressure in atm (1 atm = 760,000 mTorr), V the volume in liters, n the number of moles, R the gas constant (0.082 liter·atm/mol-K), and T the absolute temperature in K, respectively.

For a RIE system,

$$n/V = P/RT = (200/760{,}000)/(0.082 \times 300) = 1.06 \times 10^{-5}\ (\text{mol/liter})$$
$$= 1.06 \times 10^{-5} \times 6.02 \times 10^{23} \div 1000$$
$$= 6.38 \times 10^{15}\ (\text{cm}^{-3}).$$
$$\text{Ionization effeciency} = (10^9 \sim 10^{10})/(6.38 \times 10^{15})$$
$$= 1.56 \times 10^{-7} \sim 1.56 \times 10^{-6}.$$

For a HDP system,

$$n/V = P/RT = (5/760{,}000)/(0.082 \times 300) = 2.66 \times 10^{-7}\ (\text{mol/liter})$$
$$= 2.66 \times 10^{-7} \times 6.02 \times 10^{23} \div 1000$$
$$= 1.6 \times 10^{14}\ (\text{cm}^{-3}).$$
$$\text{Ionization effeciency} = (10^{11} \sim 10^{12})/(1.6 \times 10^{14})$$
$$= 6.25 \times 10^{-4} \sim 6.25 \times 10^{-3}.$$

Therefore, HDP has much higher ionization efficiency than RIE.

12.4.2 Etch Mechanism, Plasma Diagnostics, and End-Point Control

Plasma etching is a process in which a solid film is removed by a chemical reaction with ground-state or excited-state neutral species. Plasma etching is often enhanced or induced by energetic ions generated in a gaseous discharge. The basic etch mechanism, plasma diagnostic, and end-point control are introduced briefly in this section.

Etch Mechanism

Plasma etching proceeds in five steps, as illustrated in Fig. 24: the etchant species is generated in the plasma; the reactant is then transported by diffusion through a stagnant gas layer to the surface. The reactant is adsorbed on the surface; a chemical reaction (along with physical effects such as ion bombardment) follows to form volatile compounds; the compounds are desorbed from the surface, diffused into the bulk gas, and pumped out by the vacuum system.[23]

Plasma etching is based on the generation of plasma in a gas at low pressure. Two basic methods are used—physical methods and chemical methods. The former includes sputter etching and the latter includes pure chemical etching. In physical etching, positive

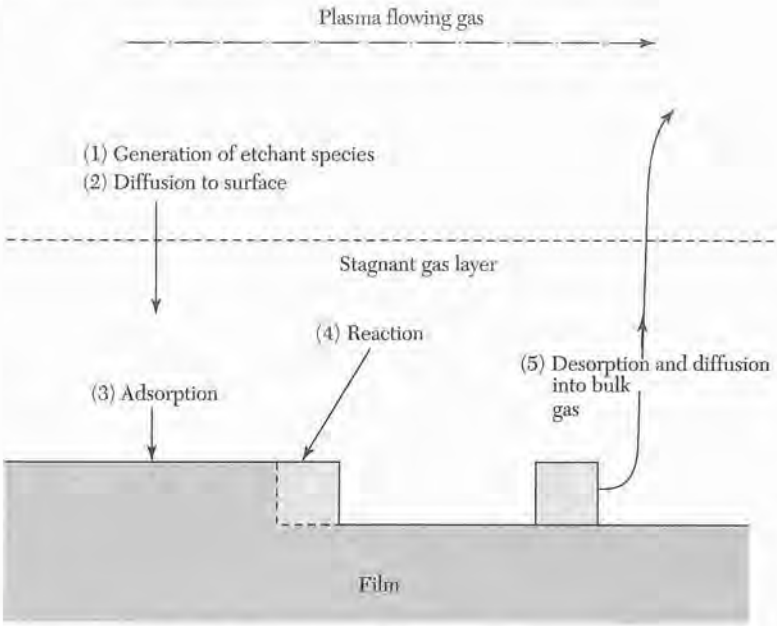


Fig. 24 Basic steps in a dry-etching processing.[23]

ions bombard the surface at high speed; small amounts of negative ions formed in the plasma cannot reach the wafer surface and therefore play no direct role in plasma etching. In chemical etching, neutral reactive species generated by the plasma interact with the material surface to form volatile products. Chemical- and physical-etch mechanisms have different characteristics. Chemical etching exhibits a high etch rate, and good selectivity (i.e., the ratio of etch rates for different materials) produces low ion-bombardment–induced damage but yields isotropic profiles. Physical etching can yield anisotropic profiles, but it is associated with low etch selectivity and high bombardment-induced damage. Combinations of chemical and physical etching give anisotropic etch profiles, reasonably good selectivity, and moderate bombardment-induced damage. An example is the RIE, which uses a physical method to assist chemical etching or creates reactive ions to participate in chemical etching

Plasma Diagnostics

Most processing plasmas emit radiation in that range from infrared to ultraviolet. A simple analytical technique is to measure the intensity of these emissions versus wavelength with the aid of optical emission spectroscopy (OES). Using observed spectral peaks, it is usually possible to determine the presence of neutral and ionic species by correlating these emissions with previously determined spectral series. Relative concentrations of the species can be obtained by correlating changes in intensity with the plasma parameter. The emission signal derived from the primary etchant or by-product begins to rise or fall at the end of the etch cycle.

End-Point Control

Dry etching differs from the wet chemical etching, in that dry etching does not have enough etch selectivity to the underlying layer. Therefore, the plasma reactor must be equipped with a monitor that indicates when the etching process is to be terminated

(i.e., an end point detection system). Laser interferometry of the wafer surface is used to continuously monitor etch rates and to determine the end point. During etching, the intensity of laser light reflected off a thin film surface oscillates. This oscillation occurs because of the phase interference between the light reflected from the outer and inner interfaces of the etching layer. This layer must therefore be optically transparent or semi-transparent to observe the oscillation. Figure 25 shows a typical signal from a silicide/polycrystalline Si gate etch. The period of the oscillation is related to the change in film thickness by

$$\Delta d = \lambda / 2\overline{n} \tag{13}$$

where Δd is the change in film thickness for one period of reflected light, λ is the wavelength of the laser light, and $\overline{n}$ is the refractive index of the etching layer. For example, Δd for polysilicon is 80 nm, measured by using a helium-neon laser source for which $\lambda = 632.8$ nm.

12.4.3 Reactive Plasma-Etching Techniques and Equipment

Plasma reactor technology in the IC industry has changed dramatically since the first application of plasma processing to photoresist stripping. A reactor for plasma etching contains a vacuum chamber, pump system, power supply generators, pressure sensors, gas flow control units, and end-point detector. Table 3 shows the similarities and differences in the types of etching equipment that are commercially available. A comparison of pressure operating ranges and ion energies for different types of reactors is shown in Fig. 26. Each etch tool is designed empirically and uses a particular combination of pressure, electrode configuration and type, and source frequency to control the two primary etch mechanisms – chemical and physical. Higher etch rates and tool automation are required for most etchers used in manufacturing.

Reactive Ion Etcher (RIE)

RIE has been extensively used in the microelectronic industry. In a parallel-plate diode system, a radio frequency (rf) capacitive-coupled bottom electrode holds the wafer. This allows the grounded electrode to have a significantly larger area because it is, in fact, the chamber itself. The larger grounded area combined with the lower operating pressure

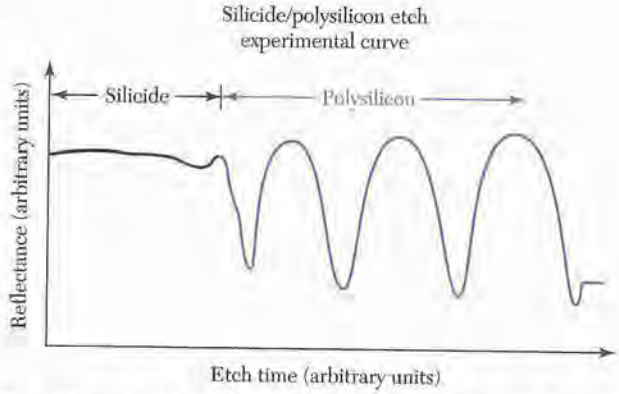


Fig. 25 The relative reflectance of the etching surface of a composite silicide/poly-Si layer. The end point of the etch is indicated by the cessation of the reflectance oscillation.

TABLE 3 Etch Mechanisms, Pressure Ranges of Plasma Reactors

Etch tool configuration	Etch mechanism	Pressure range (Torr)
Barrel etching	Chemical	0.1–10
Downstream plasma etching	Chemical	0.1–10
Reactive ion etching (RIE)	Chemical and physical	0.01–1
Magnetic enhanced RIE	Chemical and physical	0.01–1
Magnetic confinement triode RIE	Chemical and physical	0.001–0.1
Electron cyclotron resonance plasma etch	Chemical and physical	0.001–0.1
Inductively coupled plasma or transformer-coupled plasma	Chemical and physical	0.001–0.1
Surface wave coupled plasma or helicon plasma etching	Chemical and physical	0.001–0.1

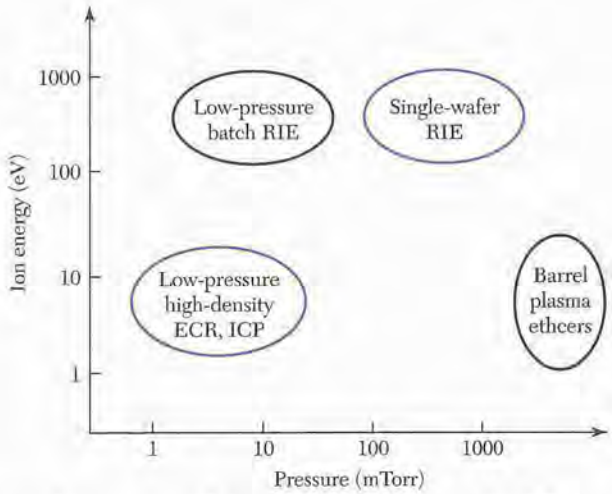


Fig. 26 Comparison of ion energy and operating pressure ranges for different types of plasma reactors.

(< 500 mTorr) causes the wafers to be subjected to a heavy bombardment of energetic ions from the plasma as a result of the large, negative self-bias at the wafer surface.

The etch selectivity of this system is relatively low compared with traditional barrel-etch systems because of strong physical sputtering. However, selectivity can be improved by choosing the proper etch chemistry, for example, by polymerizing the silicon surface with fluorocarbon polymers to obtain selectivity of SiO_2 over silicon. Alternatively, a triode-configuration RIE etch, as shown in Fig. 27, can separate plasma generation from ion transport. Ion energy is controlled through a separate bias on the wafer electrode, thereby minimizing the loss of selectivity and the ion-bombardment–induced damage observed in most traditional RIE systems.

Electron Cyclotron Resonance Plasma Etcher

Most parallel-plate plasma etchers, except triode RIE, do not provide the ability to control plasma parameters, such as electron energy, plasma density, and reactant density inde-

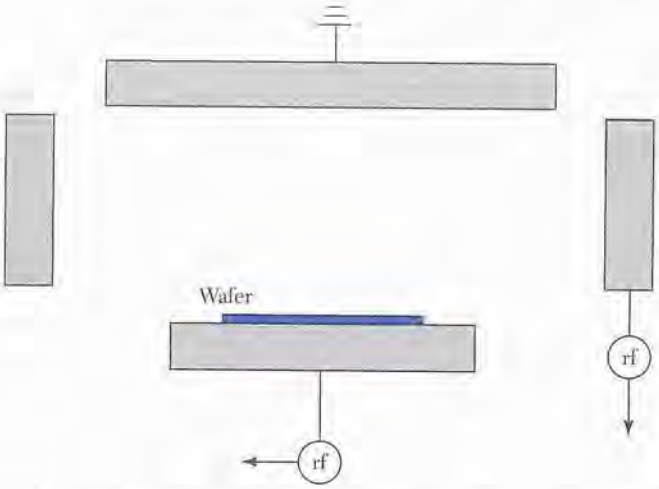


Fig. 27 Schematic of a triode reactive ion etch reactor. The ion energy is separately controlled by a bias voltage on the bottom electrode. rf, radio frequency.

pendently. As a result, ion-bombardment–induced damage becomes a serious problem. The electron cyclotron resonance (ECR) reactor combines microwave power with a static magnetic field to force electrons to circulate around the magnetic field lines at an angular frequency. When this frequency equals to the applied microwave frequency, a resonance coupling occurs between the electron energy and the applied electric field that results in a high degree of dissociation and ionization (10^{-2} for ECR compared with 10^{-6} for RIE). Figure 28 shows an ECR reaction chamber configuration. Microwave power is coupled through a microwave window into the ECR source region. The magnetic field is supplied from the magnetic coils. The ECR plasma systems can also be used in thin film deposition. The high efficiency in exciting the reactants in ECR plasmas allows the deposition of films at room temperature without the need for thermal activation.

Other High-Density Plasma Etchers

As feature sizes for ultralarge-scale integration (ULSI) continue to decrease, the limits of the conventional RIE system are being approached. In addition to the ECR system, other types of high-density plasma (HDP) sources, such as the inductively coupled plasma (ICP) source, the transformer-coupled plasma (TCP) source, and the surface wave–coupled plasma (SWP) sources, have been developed. These etchers have high plasma density (10^{11}–10^{12} cm^{-3}), and low processing pressure (< 20 mTorr). In addition, they allow the wafer platen to be powered independently of the source, providing significant decoupling between the ion energy (wafer bias) and the ion flux (plasma density primarily driven by source power). The primary processing advantages of HDP sources are better critical dimension (CD) control, higher etching rates, and better selectivity.

In addition, HDP sources provide low substrate damage (because of independent biasing of the substrate and the side-electrode potentials) and high anisotropy (because of low pressure yet high active species density). However, because of their complexity and higher cost, the systems may not be used for less critical applications, such as spacer etching or planarization.[24] Figure 29 shows a TCP plasma reactor. A high-density, low-pressure plasma is generated by a flat spiral coil that is separated from the plasma by a dielectric plate on the top of the reactor. The wafer is located away from the coil, so it is not affected by the electromagnetic field generated by the coil. There is little plasma

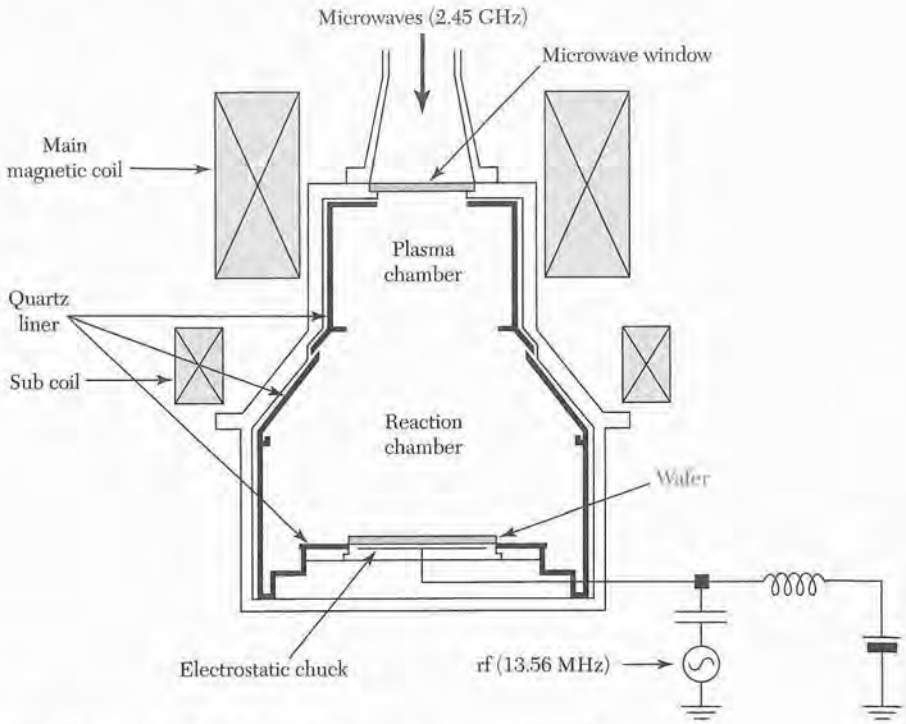


Fig. 28 Schematic of an electron cyclotron resonance reactor.[24]

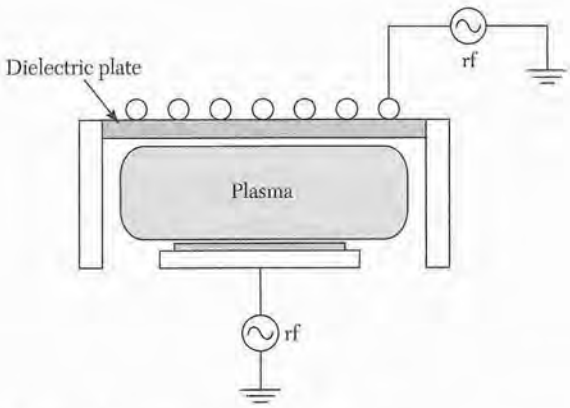


Fig. 29 Schematic of a transformer-coupled plasma reactor.

density loss because plasma is generated only a few mean free paths away from the wafer surface. Therefore, a high-density plasma and high etch rates are achieved.

Clustered Plasma Processing

Semiconductor wafers are processed in clean rooms to minimize exposure to ambient particulate contamination. As device dimensions shrink, particulate contamination becomes a more serious problem. To minimize particulate contamination, clustered plasma tools uses a wafer handler to pass wafers from one process chamber to another in a vacuum

environment. The clustered plasma processing tools can also increase throughput. Figure 30 shows the multilayer metal interconnect (TiW/AlCu/TiW) etching process with clustered tools in an AlCu etch chamber, a TiW etch chamber, and a strip passivation chamber. The clustered tools provide an economic advantage by having a high chip yield because the wafer is exposed to less ambient contamination and is handled less.

12.4.4 Reactive Plasma–Etching Applications

Plasma etching has rapidly evolved from a simple and batch resist stripping to a large and single-wafer processing. The etching system continues to be improved from the conventional RIE tool to the high-density plasma tool for pattern transfer of deep-submicron devices. Besides the etching tool, etch chemistry also plays a critical role in the performance of etch process. Table 4 lists some etch chemistries for different etch processes. Developing an etch process usually means optimizing etch rate, selectivity, profile control, critical dimension, damage, etc., by adjusting a large number of process parameters.

Silicon Trench Etching

As device feature size decreases, a corresponding decrease is needed in the wafer surface area occupied by the isolation between circuit elements and the storage capacitor of a DRAM cell. This surface area can be reduced by etching trenches into the silicon substrate and filling them with suitable dielectric or conductive materials. Deep trenches, usually with a depth larger than 5 μm, are used mainly for forming storage capacitors. Shallow trenches, usually with a depth less than 1 μm, are often used for isolation.

Chlorine-based and bromine-based chemistries have a high silicon etch rate and high etch selectivity to the silicon dioxide mask. The combination of $HBr + NF_3 + SF_6 + O_2$ gas mixtures is used to form a trench capacitor with a depth of ~7 μm. It is also used for shallow trench isolation etching. Aspect ratio dependent etching (i.e., variation in etch rate with aspect ratio) is often observed in submicron-deep silicon trench etching caused

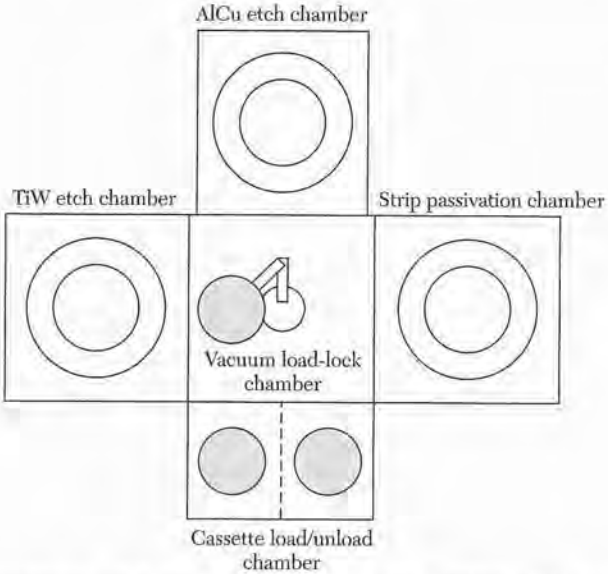


Fig. 30 Cluster reactive ion etch tool for multilayer metal (TiW/AlCu/TiW) interconnect etching.[2]

TABLE 4 Etch Chemistries of Different Etch Processes

Material being etched	Etching chemistry
Deep Si trench	$HBr/NF_3/O_2/SF_6$
Shallow Si trench	$HBr/Cl_2/O_2$
Poly Si	$HBr/Cl_2/O_2$, HBr/O_2, BCl_3/Cl_2, SF_6
Al	BCl_3/Cl_2, $SiCl_4/Cl_2$, HBr/Cl_2
AlSiCu	$BCl_3/Cl_2/N_2$
W	SF_6 only, NF_3/Cl_2
TiW	SF_6 only
WSi_2, $TiSi_2$, $CoSi_2$	CCl_2F_2/NF_3, CF_4/Cl_2, $Cl_2/N_2/C_2F_6$
SiO_2	$CF_4/CHF_3/Ar$, C_2F_6, C_3F_8, C_4F_8/CO, C_5F_8, CH_2F_2
Si_3N_4	CHF_3/O_2, CH_2F_2, CH_2CHF_2

by limited ion and neutral transport within the trench. Figure 31 shows the dependence of average silicon trench etch rate on aspect ratio. Trenches with large aspect ratios are etched more slowly than trenches with small aspect ratios.

Polysilicon and Polycide Gate Etching

Polysilicon or polycide (i.e., low-resistance metal silicides over polysilicon) is usually used as a gate material for MOS devices. Anisotropic etching and high etch selectivity to the gate oxide are the most important requirements for gate etching. For example, the selectivity required in 1G DRAM is more than 150 (i.e., the ratio of etch rates for polycide and gate oxide is 150:1). Achieving high selectivity and etch anisotropy at the same time is difficult for most ion-enhanced etching processes. Therefore, multistep processing is used in which different etch steps in the process are optimized for etch anisotropy and selectivity. On the other hand, the trend in plasma technology for anisotropic etching and high selectivity is to utilize a low-pressure, high-density plasma using a relatively low power. Most chlorine-based and bromine-based chemistries can be used for gate etching to achieve the required etch anisotropy and selectivity.

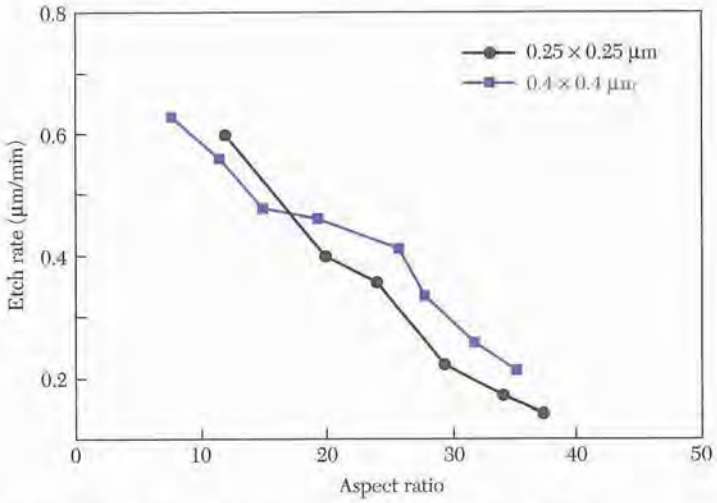


Fig. 31 Dependence of average silicon trench etch rate on aspect ratio.[2]

Dielectric Etching

The patterning of dielectrics, especially silicon dioxide and silicon nitride, is a key process in the manufacture of modern semiconductor devices. Because of their higher bonding energies, dielectric etching requires aggressive ion-enhanced, fluorine-based plasma chemistry. Vertical profiles are achieved by sidewall passivation, typically by introducing a carbon-containing fluorine species to the plasma (e.g., CF_4, CHF_3, C_4F_8). High ion-bombardment energies are required to remove this polymer layer from the oxide, as well as to mix the reactive species into the oxide surface to form SiF_X products.

A low-pressure, high-density plasma is advantageous for aspect-ratio–dependent etching. However, the HDP generates high-temperature electrons and subsequently generates a high degree of dissociation of ions and radicals. It generates far more active radicals and ions than RIE or MERIE plasmas . In particular, a high F concentration worsens the selectivity to silicon. Various methods were tried to enhance the selectivities in the high-density plasma. First, a parent gas with a high C/F ratio, such as C_2F_6, C_4F_8, or C_5F_8, has been tried. Also other methods to scavenge F radicals have been developed.[25]

Interconnect Metal Etching

Etching of a metallization layer is a very important step in IC fabrication. Aluminum, copper, and tungsten are the most popular materials used for interconnection. These materials usually require anisotropic etching. The reaction of aluminum with fluorine results in nonvolatile AlF_3, which has a vapor pressure of only 1 Torr at 1240°C. Chlorine-based (e.g., Cl_2/BCl_3 mixture) chemistry has been widely used for aluminum etching. Chlorine has a very high chemical etch rate with aluminum and tends to produce an undercut during etching. Carbon-containing gas (e.g., CHF_3) or N_2 is added to form sidewall passivation during aluminum etching to obtain anisotropic etching.

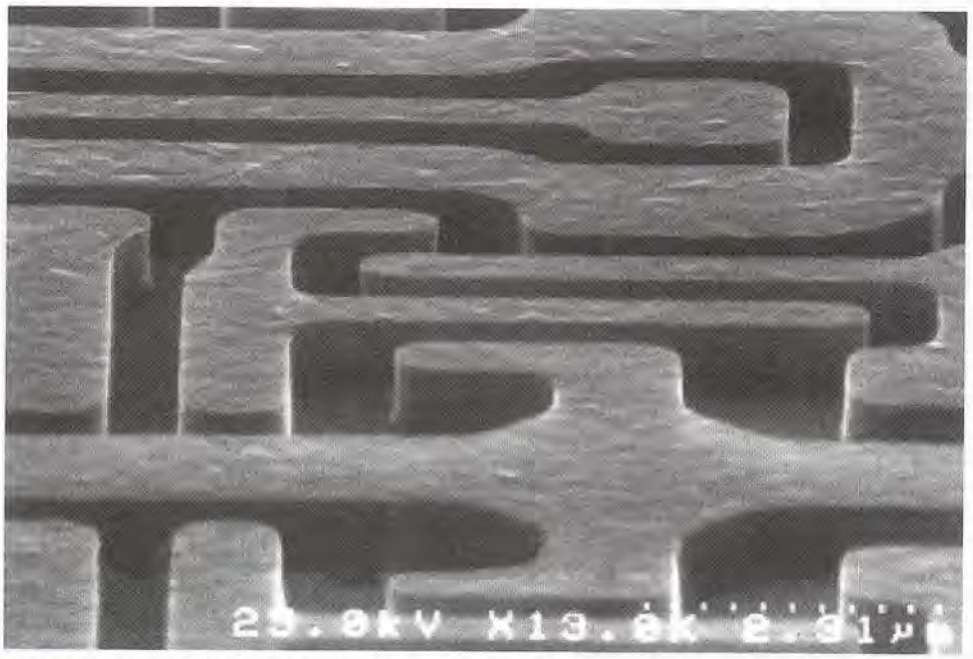


Fig. 32 0.35 μm TiN/Al/TiN lines and spaces on a wafer maintained at ambient for 72 hours after a microwave strip are not corroded.

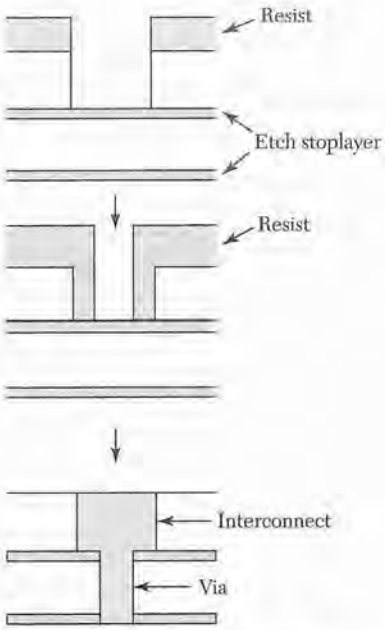


Fig. 33 Various process sequences for dual damascene process.

Exposure to the ambient is another problem in aluminum etching. Residue chlorine on the Al sidewall and the photoresist tends to react with atmosphere water to form HCl, which corrodes aluminum. An in-situ exposure of the wafer to a CF_4 discharge to exchange Cl with F and then to an oxygen discharge to remove the resist followed by immediate immersion in deionized water can eliminate Al corrosion. Figure 32 shows 0.35 μm TiN/Al/Ti lines and spaces on a wafer maintained at ambient for 72 hours. No corrosion is present even after prolonged exposure to the ambient.

Copper has drawn much attention as a new metallization material in ULSI circuits because of its low resistivity (~1.7 μohm-cm) and superior resistance to electromigration compared with Al or Al alloys. However, because of the low volatility of copper halides, plasma etching at room temperature is difficult. Process temperatures higher than 200°C are required to etch coppers films. Therefore, the damascene process is used to form Cu interconnection without dry etching. Damascene processing involves the creation of interconnect lines by first etching a trench or canal in a planar dielectric layer and then filling that trench with metal, such as aluminum or copper. In dual damascene processing (Fig. 33), a second level is involved where a series of holes (i.e., contacts or vias) are etched and filled in addition to the trench. After filling, the metal and dielectric are planarized by chemical-mechanical processing (CMP). The advantage of damascene processing is that it eliminates the need for metal etch. This is an important concern as the industry moves from aluminum to copper interconnections.

Low-pressure CVD (LPCVD) tungsten (W) has been widely used for filling contact holes and first-level metallization because of its excellent deposition conformability. Both fluorine- and chlorine-based chemistries etch W and form volatile etch products. An important tungsten etch processes is the blanket W etchback to form a W plug. The blanket LPCVD W is deposited on top of a TiN barrier layer, as shown in Fig. 34. A two-step process is usually used. First, 90% of the W is etched at a high etch rate, and then the

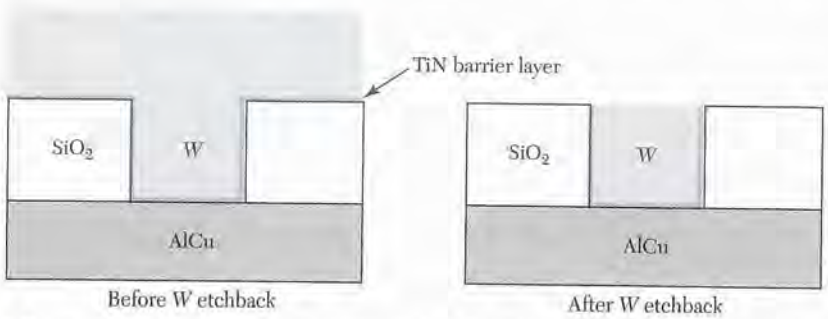


Fig. 34 Formation of tungsten plug in a contact hole by depositing blanket low-pressure chemical-vapor deposition W and then using reaction ion etching etchback.

etch rate is reduced to remove the remaining W with an etchant that has a high W-to-TiN selectivity.

12.5 MICROELECTROMECHANICAL SYSTEMS

There has been rapidly growing interest in microelectromechanical systems (MEMS), since the late 1980s when a spinning micromotor made from polysilicon was fabricated on a silicon chip.[26,27] The manufacture of silicon MEMS adopted many of the highly developed technologies of silicon integrated circuits. This approach has enabled MEMS products to be produced at low cost by using batch fabrication, just as for ICs. In addition to the IC fabrication process, some new specialized techniques have been developed for MEMS. We consider three specialized etching techniques: bulk micromachining, surface micromachining, and the LIGA process.

12.5.1 Bulk Micromachining

In bulk micromachining, the device (e.g., a sensor or an actuator) is shaped by etching a large single-crystal substrate. The films are patterned on the bulk to define the isolation and transducer functions. Orientation-dependent wet chemical etching technique provide a high-resolution etch and tight dimensional control. Often, a bulk-micromachined device uses two-sided processing, creating a self-isolated structure with one side exposed to the measured variables, such as mechamical or chemical signals while the other side is enclosed in a clean package. Two-sided structures are very robust for operation in environments hostile to microelectronic devices. Simple mechanical devices such as diaphragm pressure sensors, membranes, and cantilever-beam piezoresistive acceleration sensors are fabricated commercially by this technique. Figure 35 illustrates a fabrication process of a simple silicone rubber membrane.[28]

12.5.2 Surface Micromachining

Surface-micromachined devices are constructed entirely from thin films. There are several differences and trade-offs between structures made from bulk and thin film materials. Typical dimensions for bulk-micromachined sensors are in the millimeter range, but surface-micromachined devices are of micrometer dimensions. Surface micromachining

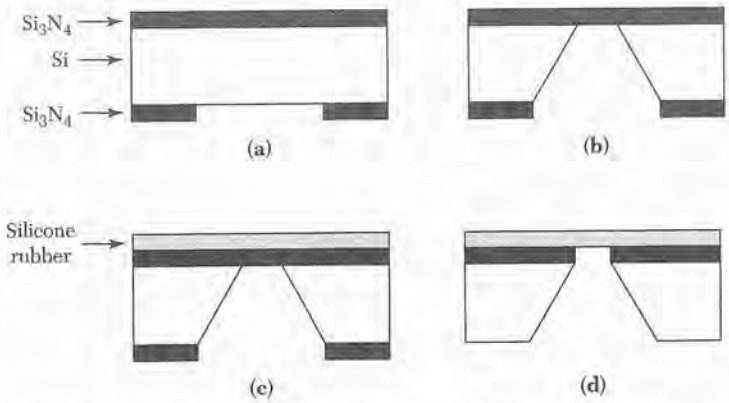


Fig. 35 Fabrication process of simple silicone rubber membrane. (*a*) Nitride deposition and patterning; (*b*) KOH etching; (*c*) silicone rubber spin coating; snf (*d*) nitride removal on back side. [28]

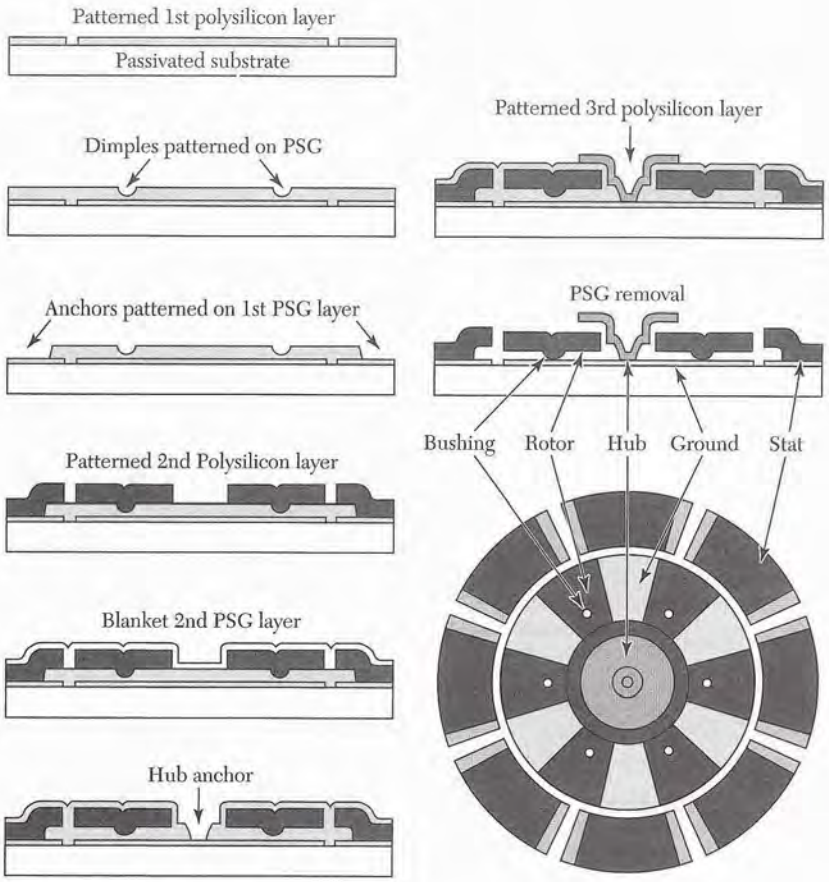
permits the fabrication of structurally complex devices by stacking and patterning layers or "building blocks" of thin films whereas multilayered bulk devices are difficult to construct. Free-standing and movable parts can be fabricated using sacrificial layers. Figure 36 illustrates how sacrificial etching techniques can be used to create an electrostatic micromotor with well-defined, submicron tolerance between the rotor and the center hub.[27]

12.5.3 LIGA Process

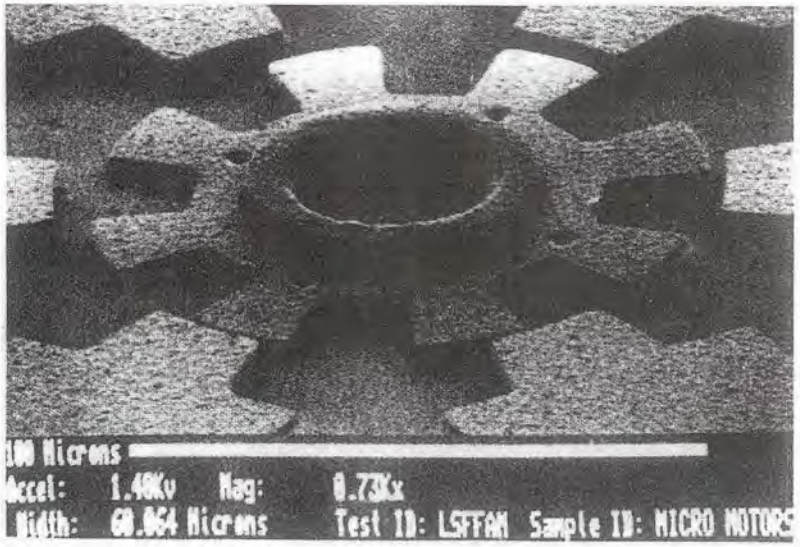
LIGA is a German acronym for lithographic, galvanoformung, abformung.[29] It consists of three basic processing steps: lithography, electroplating, and molding. The LIGA process is based upon X-ray radiation from a synchrotron. The process can produce microstructure with lateral dimensions in the micrometer range and structural heights of several hundred micrometers from a variety of materials. Its potential applications cover microelectronics, sensors, microoptics, micromechanics, and biotechnology.

An example of LIGA process is shown in Fig. 37. A thick X-ray resist ranging from 300 μm to more than 500 μm in thickness is deposited on a substrate with an electrically conductive surface. Lithographic patterning is done with extended exposure from highly collimated X-ray radiation through an X-ray mask, as shown Fig. 37*a*. The flower-shaped trench structure is formed in the thick resist after developer treatment (Fig. 37*b*). Metal is then electroplated on the exposed bottom conductive surface, filling the trench space and covering the top surface of the resist (Fig. 37*c*). The metal structure is formed after removing the resist (Fig. 37*d*). This structure can be used repeatedly as a mold insert for injection molding to form multiple plastic replicas of the original plating base (Fig. 37*e*). The plating base replicas, in turn, can be used to electroplate many metal structures as the final products, as shown in Fig. 37*f* and *g*.

The distinct advantage of the LIGA process is the ability to create three-dimensional structures as thick as bulk-micromachined devices while retaining the same degree of design flexibility as surface micromachining. However, the initial synchnotron radiation process is a very costly step, and the mold-separation steps may result in degradation of the original mold insert.



(a)



(b)

Fig. 36 (*a*) Sacrificial process flow for an electrostatic micromorotor. (*b*) Photograph of a micromotor.[27] PSG, phosphosilicate glass.

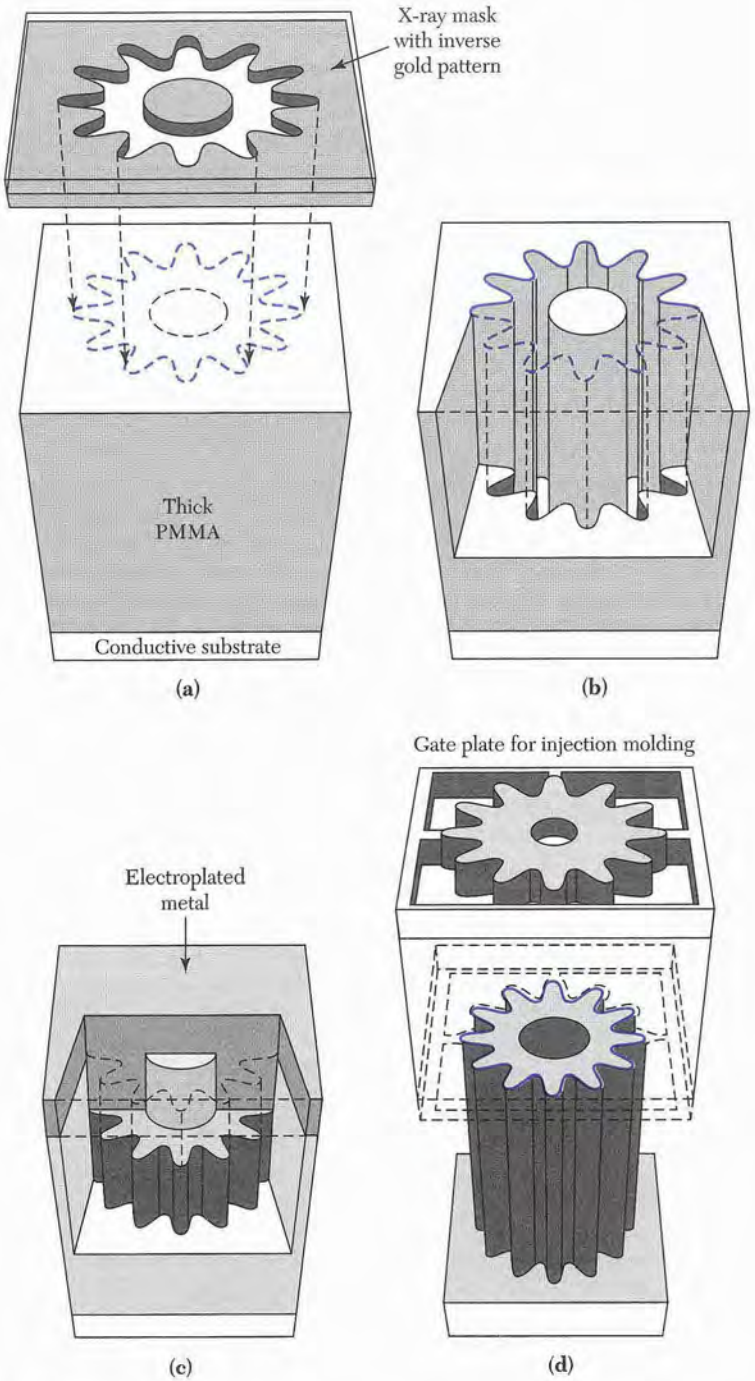


Fig. 37 The LIGA process.[29]

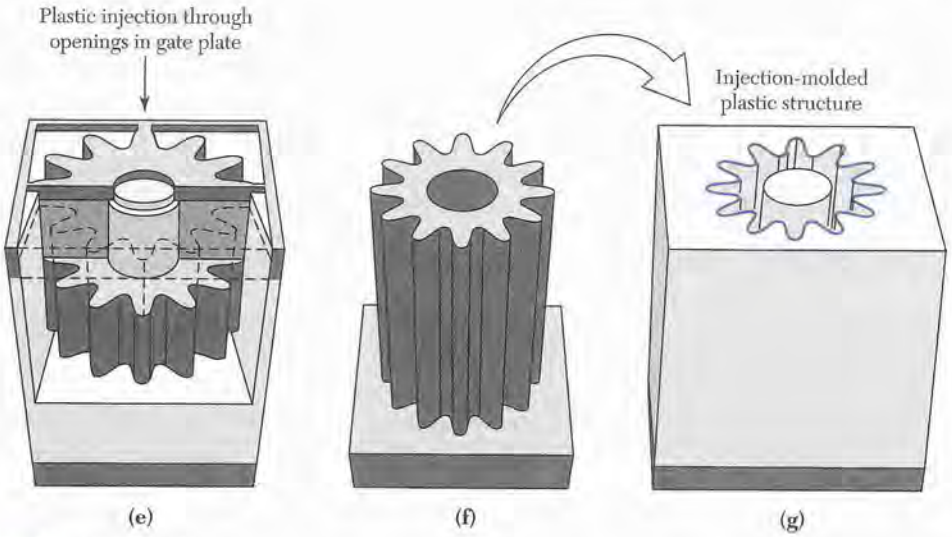


Fig. 37 *(continued)*

SUMMARY

The continued growth of the semiconductor industry is a direct result of the capability of transferring smaller and smaller circuit patterns onto semiconductor wafers. The two major processes to transfer patterns are lithography and etching.

Currently, the vast majority of lithographic equipment are optical systems. We have considered various exposure tools, masks, photoresists, and the clean room for optical lithography. The primary factor limiting resolution in optical lithography is diffraction. However, because of advancements in excimer lasers, photoresist chemistry, and resolution enhancement techniques such as the PSM and OPC, optical lithography will remain the mainstream technology, at least to the 100 nm generation.

Electron-beam lithography is the technology of choice for mask making and nanofabrication, in which new device concepts are explored. Other lithographic processing technologies are EUV, X-ray, and ion-beam lithography. Although all these have 100 nm or better resolution, each process has its own limitation: proximity effect in electron-beam lithography, mask blank production difficulties in EUV lithography, mask-fabrication complexity in X-ray lithography, and stochastic space charge in ion-beam lithography.

At the present time, no obvious successor to optical lithography can be identified unambiguously. However, a mix-and-match approach can take advantage of the unique features of each lithography process to improve resolution and to maximize throughput.

Wet chemical etching is used extensively in semiconductor processing. It is particularly suitable for blanket etching. We have discussed wet chemical etching processes for silicon and gallium arsenide, insulators, and metal interconnections. Wet chemical etching was used for pattern transfer. However the undercutting of the layer underneath the mask has resulted in loss of resolution in the etched pattern.

Dry etching methods are used to achieve high-fidelity pattern transfer. Dry etching is synonymous with plasma-assisted etching. We have considered plasma fundamentals and various dry-etching systems, which have grown from relatively simple, parallel-plate configurations to complex chambers with multiple-frequency generators and a variety of process-control sensors.

The challenges for future etching technology are high etch selectivity, better critical-dimensioned control, low aspect-ratio–dependent etching, and low plasma-induced damage. Low-pressure, high-density plasma reactors are necessary to meet these requirements. As processing evolves from 200 mm to 300 mm and even larger wafers, continued improvements are required for etch uniformity within the wafer. New gas chemistries must be developed to provide the improved selectivity necessary for advanced integration schemes.

Microelectromechanical system (MEMS) is an emerging field. MEMS has adopted the lithographic and etching technologies from IC fabrication. Specialized etching techniques have also been developed for MEMS: bulk micromachining using an orientation-dependent etching process, surface micromachining using sacrifical layers, and the LIGA process using X-ray lithography with highly collimated radiation.

REFERENCES

1. For a more detailed discussion on lithography, see (a) K. Nakamura, "Lithography," in C. Y. Chang and S. M. Sze, Eds., *ULSI Technology*, McGraw-Hill, New York, 1996. (b) P. Rai-Choudhurg, *Handboolk of Microlithography, Micromachining, and Microfabrication*, Vol. 1, SPIE, Washington, DC, 1997. (c) D. A. McGillis, "Lithography," in S. M. Sze. Ed., *VLSI Technology*, McGraw-Hill, New York, 1983.
2. For a more detailed discussion on etching, see Y. J. T. Liu, " Etching," in C. Y. Chang and S. M. Sze, Eds., *ULSI Technology*, McGraw-Hill, New York, 1996.
3. J.M. Duffalo and J. R. Monkowski, " Particulate Contamination and Device Performance," *Solid State Technol.* **27**, 3, 109 (1984).
4. H. P. Tseng and R. Jansen, "Cleanroom Technology," in C. Y. Chang and S. M. Sze, Eds., *ULSI Technology*, McGraw-Hill, New York, 1996.
5. M. C. King, " Principles of Optical Lithography," in N. G. Einspruch, Ed., *VLSI Electronics*, Vol. 1, Academic, New York, 1981.
6. J. H. Bruning, "A Tutorial on Optical Lithography," in D. A. Doane, et al., Eds., *Semiconductor Technology*, Electrochemical Soc., Penningston, 1982.
7. R. K. Watts and J. H. Bruning, "A Review of Fine-Line Lithographic Techniques: Present and Future," *Solid State Technol.*, **24**, 5,99 (1981).
8. W. C. Till and J. T. Luxon, *Integrated Circuits, Materials, Devices, and Fabrication*, Princeton-Hall, Englewood Cliffs, NJ, 1982.
9. M. D. Levenson, N. S. Viswanathan, and R. A. Simpson, "Improving Resolution in Photolithography with a Phase-Shift Mask," *IEEE Trans. Electron Devices*, **ED-29**, 18–28 (1982).
10. D. P. Kern, et al., "Practical Aspects of Microfabrication in the 100-nm Region," *Solid State Technol.*, **27**, 2, 127 (1984).
11. J. A. Reynolds, "An Overview of e-Beam Mask-Making, "*Solid State Technol.*, **22**, 8, 87 (1979).
12. Y. Someda, et al. "Electron-Beam Cell Projection Lithography: Its Accuracy and Its Throughput," *J. Vac. Sci. Technol.*, **B12** (6), 3399 (1994).
13. J. A. Liddle, et al., "The Scattering with Angular Limitation in Projection Electron-Beam Lithography (SCALPEL) System," *Jpn. J. Appl. Phys.*, **34**, 6663 (1995).
14. W. L. Brown, T. Venkatesan, and A. Wagner, "Ion Beam Lithography," *Solid State Technol.*, **24**, 8, 60 (1981).
15. D. S. Kyser and N. W. Viswanathan, "Monte Carlo Simulation of Spacially Distributed Beams in Election-Beam Lithography", *J. Vac. Sci. Technol.*, **12**, 1305 (1975).
16. Charles Gwyn, et al., *Extreme Ultraviolet Lithography-White Paper*, Sematech, Next Generation Lithography Workshop, Colorado Springs, Dec. 7–10, 1998.
17. J. P. Silverman, *Proximity X-Ray Lithography-White Paper*, Sematech, Next Generation Lithography Workshop, Colorado Spring, Dec. 7-10. 1998.

18. L. Karapiperis, et al., "Ion Beam Exposure Profiles in PMMA-Computer Simulation," *J. Vac. Sci. Technol.*, **19**, 1259 (1981).

19. H. Robbins and B. Schwartz, "Chemical Etching of Silicon II, the System HF, HNO_3, H_2O and $HC_2H_3O_2$, " *J. Electrochem. Soc.*, **107**, 108 (1960).

20. K. E. Bean, "Anisotropic Etching in Silicon," *IEEE Trans. Electron Devices*, **ED-25**, 1185 (1978).

21. S. Iida and K. Ito, "Selective Etching of Gallium Arsenide Crystal in H_2SO_4-H_2O_2-H_2O System," *J. Electrochem. Soc.*, **118**, 768 (1971).

22. E. C. Douglas, "Advanced Process Technology for VLSI Circuits," *Solid State Technol.*, **24**, 5, 65 (1981).

23. J. A. Mucha and D. W. Hess, " Plasma Etching," in L. F. Thompson and C. G. Willson, Eds., *Microcircuit Processing: Lithography and Dry Etching*, American Chemical Society, Washington, D. C., 1984.

24. M. Armacost et. al., "Plasma-Etching Processes for ULSI Semiconductor Circuits," *IBM J. Res. Dev.*, **43**, 39 (1999).

25. C. O. Jung, et al., "Advanced Plasma Technology in Microelectronics," *Thin Solid Films*, **341**, 112, (1999).

26. C. H. Mastrangelo and W. C. Tang, " Semiconductor Sensor Technology," in S. M. Sze, Ed., *Semiconductor Sensors*, Wiley, New York, 1994.

27. L. S. Fan, Y. C. Tai, and R. S. Muller, "IC-Processed Electrostatic Micromotors," in *IEEE Int. Electron Devices Meeting* (IEDM), p. 666, 1988.

28. X. Yang, et al., "A MEMS Thermopenumatic Silicone Rubber Membrane Valve, " *Sens. Actuators*, **A64**, 101, (1998).

29. W. Ehrfeld, et al. "Fabrication of Microstructures Using the LIGA Process," *Proc. IEEE Micro Robots and Teleoperators Workshop*, Hyannis, MA, Nov. 1987.

PROBLEMS (* DENOTES DIFFICULT PROBLEMS)

FOR SECTION 12.1 OPTICAL LITHOGRAPHY

1. For a class 100 clean room, find the number of dust particles per cubic meter with particle sizes (a) between 0.5 and 1 μm, (b) between 1 and 2 μm, and (c) above 2 μm.

2. Find the final yield for a nine-mask–level process in which the average fatal defect density per cm^2 is 0.1 for four levels, 0.25 for four levels, and 1.0 for one level. The chip area is 50 mm^2.

3. An optical lithographic system has an exposure power of 0.3 mW/cm^2. The required exposure energy for a positive photoresist is 140 mJ/cm^2 and for a negative photoresist is 9 mJ/cm^2. Assuming negligible times for loading and unloading wafers, compare the wafer throughput for positive photoresist and negative photoresist.

4. (a) For an ArF-excimer laser 193 nm optical lithographic system with NA = 0.65, k_1 = 0.60, and k_2 = 0.50, what are the theoretical resolution and depth-of-focus for this tool? (b) What can we do in practice to adjust NA, k_1, and k_2 parameters to improve resolution? (c) What parameter does the phase-shift mask (PSM) technique change to improve resolution?

5. The plots in Fig. 9 are called *response curves* in microlithography. (a) What are the advantages and disadvantages of using resists with high γ values? (b) Conventional resists cannot be used for 248 nm or 193 nm lithography. Why not?

FOR SECTION 12.2 NEXT-GENERATION LITHOGRAPHIC METHODS

6. (a) Explain why a shaped beam promises higher throughput than a Gaussian beam in e-beam lithography. (b) How can alignment be performed for e-beam lithography? Why is alignment in X-ray lithography so difficult? (c) What are the potential advantages of X-ray lithography over e-beam lithography?

7. Why has the operating mode of optical lithographic systems evolved from proximity printing to 1:1 projection printing and finally to 5:1 projection step-and-repeat? (b) Is it possible to build a step-and-scan X-ray lithographic system? Why or why not?

FOR SECTION 12.3 WET CHEMICAL ETCHING

8. If the mask and the substrate can not be etched by a particular etchant, sketch the edge profile of an isotropically etched feature in a film of thickness h_f for (a) etching just to completion, (b) 100% overetch, and (c) 200% overetch.

9. A <100>-oriented silicon crystal is etched in a KOH solution through a 1.5 μm × 1.5 μm window defined in silicon dioxide. The etch rate normal to (100)-planes is 0.6 μm/min. The etch rate ratios are 100:16:1 for the (100):(110):(111)-planes. Show the etched profile after 20 seconds, 40 seconds, and 60 seconds.

10. Repeat the previous problem, a <$\bar{1}$10>-oriented silicon is etched with a thin SiO_2 mask in KOH solution. Show the etched pattern profiles on <$\bar{1}$10>-Si.

11. A <100>-oriented silicon wafer 150 mm in diameter is 625 μm thick. The wafer has 1000 μm × 1000 μm ICs on it. The IC chips are to be separated by orientation-dependent etching. Describe two methods for doing this and calculate the fraction of the surface area that is lost in these process.

FOR SECTION 12.4 DRY ETCHING

*12. The average distance traveled by particles between collisions is called the mean free path (λ), $\lambda \cong 5 \times 10^{-3}/P$(cm), where P is pressure in Torr. In typical plasmas of interest, the chamber pressure ranges from 1 Pa to 150 Pa. What are the corresponding density of gas molecules (cm^{-3}) and the mean free path?

13. Fluorine (F) atoms etch Si at a rate given by

$$\text{Etch Rate (nm/min)} = 2.86 \times 10^{-13} n_F \times T^{1/2} e^{-E_a/RT}$$

where n_F is the concentration of F atoms (cm^{-3}), T the temperature (K), and E_a and R the activation energy (2.48 kcal/mol) and gas constant (1.987 cal-K), respectively. If n_F is 3×10^{15}, calculate the etch rate of Si at room temperature.

14. Repeat the previous problem, SiO_2 etched by F atoms could also be expressed by

$$\text{Etch rate (nm/min)} = 0.614 \times 10^{-13} n_F \times T^{1/2} e^{-E_a/RT}$$

where n_F is 3×10^{15}(cm^{-3}) and E_a is 3.76 kcal/mol. Calculate the etch rate of SiO_2 and etch selectivity of SiO_2 over Si at room temperature.

15. A multiple-step etch process is required for etching a polysilicon gate with thin gate oxide. How do you design an etch process that has no micromasking, has an anisotropic etch profile, and is selective to thin gate oxide?

16. Find the etch selectivity required to etch a 400-nm polysilicon layer without removing more than 1 nm of its underlying gate oxide, assuming that the polysilicon is etched with a process having a 10% etch-rate uniformity.

17. A 1 μm Al film is deposited over a flat field oxide region and patterned with photoresist. The metal is then etched with a mixture of BCl_3/Cl_2 gases at a temperature of 70°C in a Helicon etcher. The selectivity of Al over photoresist is maintained at 3. Assuming a 30% overetch, what is the minimum photoresist thickness required to ensure that the top metal surface is not attacked?

18. In an ECR plasma, a static magnetic field B forces electrons to circulate around the magnetic field lines at an angular frequency, ω_e, which is given by

$$\omega_e = qB/m_e.$$

where q is the electronic charge, and m_e the electron mass. If the microwave frequency is 2.45 GHz, what is the required magnetic field.

19. What are the major distinctions between the traditional reactive ion etching and high density plasma etching (ECR, ICP, etc.)?

20. Describe how to eliminate the corrosion issues in Al lines after etching with chlorine-based plasma.

CHAPTER 13

Impurity Doping

- 13.1 BASIC DIFFUSION PROCESS
- 13.2 EXTRINSIC DIFFUSION
- 13.3 DIFFUSION-RELATED PROCESSES
- 13.4 RANGE OF IMPLANTED IONS
- 13.5 IMPLANT DAMAGE AND ANNEALING
- 13.6 IMPLANTATION-RELATED PROCESSES
- SUMMARY

Impurity doping is the introduction of controlled amounts of impurity dopants into semiconductors. The practical use of impurity doping has been mainly to change the electrical properties of the semiconductors. *Diffusion* and *ion implantation* are the two key methods of impurity doping. Until the early 1970s, impurity doping was done mainly by diffusion at elevated temperatures, as shown in Fig. 1*a*. In this method the dopant atoms are placed on or near the surface of the wafer by deposition from the gas phase of the dopant or by using doped-oxide sources. The doping concentration decreases monotonically from the surface, and the profile of the dopant distribution is determined mainly by the temperature and diffusion time.

Since the early 1970s, many doping operations have been performed by ion implantation, as shown in Fig. 1*b*. In this process the dopant ions are implanted into the semiconductor by means of an ion beam. The doping concentration has a peak distribution inside the semiconductor and the profile of the dopant distribution is determined mainly by the ion mass and the implanted-ion energy. Both diffusion and ion implantation are used for fabricating discrete devices and integrated circuits because these processes generally complement each other.[1,2] For example, diffusion is used to form a deep junction (e.g., a twin well in CMOS) whereas ion implantation is used to form a shallow junction (e.g., a source/drain junction of a MOSFET).

Specifically, we cover the following topics:

- The movement of impurity atoms in the crystal lattice under high temperature and high concentration-gradient conditions.
- Impurity profiles for constant diffusivity and concentration-dependent diffusivity.
- The impact of lateral diffusion and impurity redistribution on device characteristics.
- The process and advantages of ion implantation.
- Ion distributions in the crystal lattice and how to remove lattice damages caused by ion implantation.

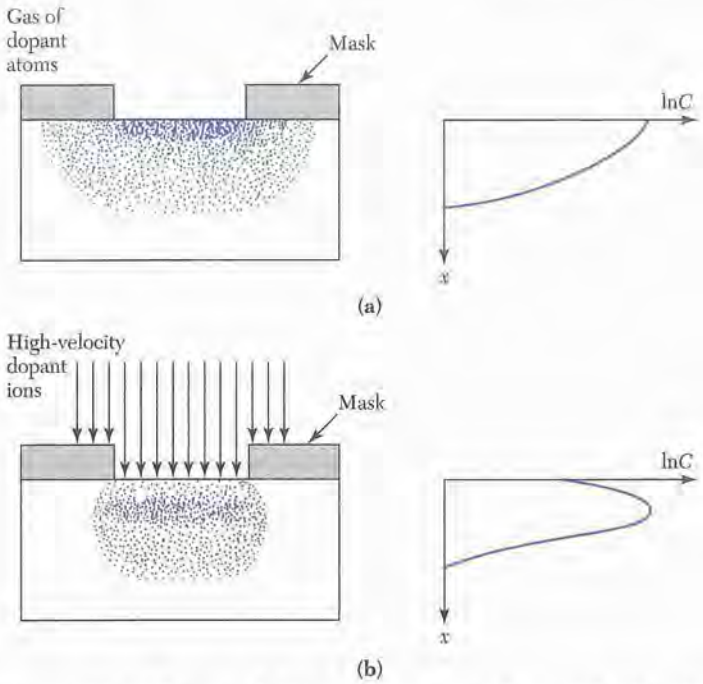


Fig. 1 Comparison of (*a*) diffusion and (*b*) ion-implantation techniques for the selective introduction of dopants into the semiconductor substrate.

- Implantation-related processes such as masking, high-energy implantation, and high-current implantation.

13.1 BASIC DIFFUSION PROCESS

Diffusion of impurities is typically done by placing semiconductor wafers in a carefully controlled high-temperature quartz-tube furnace and passing a gas mixture that contains the desired dopant through it. The temperature usually ranges between 800° and 1200°C for silicon and 600° and 1000°C for gallium arsenide. The number of dopant atoms that diffuse into the semiconductor is related to the partial pressure of the dopant impurity in the gas mixture.

For diffusion in silicon, boron is the most popular dopant for introducing a *p*-type impurity, whereas arsenic and phosphorus are used extensively as *n*-type dopants. These three elements are highly soluble in silicon, as they have solubilities above 5×10^{20} cm^{-3} in the diffusion temperature range. These dopants can be introduced in several ways, including solid sources (e.g., BN for boron, As_2O_3 for arsenic, and P_2O_5 for phosphorus), liquid sources (BBr_3, $AsCl_3$, and $POCl_3$), and gaseous sources (B_2H_6, AsH_3, and PH_3). However, liquid sources are most commonly used. A schematic diagram of the furnace and gas flow arrangement for a liquid source is shown in Fig. 2. This arrangement is similar to that used for thermal oxidation. An example of the chemical reaction for phosphorus diffusion using a liquid source is

$$4POCl_3 + 3O_2 \rightarrow 2P_2O_5 + 6Cl_2 \uparrow . \qquad (1)$$

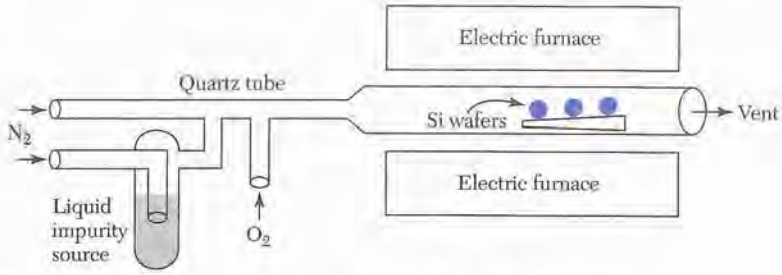


Fig. 2 The schematic diagram of a typical open-tube diffusion system.

The P_2O_5 forms a glass on silicon wafer and is then reduced to phosphorus by silicon,

$$2P_2O_5 + 5Si \rightarrow 4P + 5SiO_2, \tag{2}$$

and the phosphorus is released and diffuses into the silicon and Cl_2 is vented.

For diffusion in gallium arsenide, the high vapor pressure of arsenic requires special methods to prevent the loss of arsenic by decomposition or evaporation.[2] These methods include diffusion in sealed ampules with an overpressure of arsenic and diffusion in an open-tube furnace with a doped-oxide capping layer (e.g., silicon nitride). Most of the studies on p-type diffusion have been confined to the use of zinc in the forms of Zn-Ga-As alloys and $ZnAs_2$ for the sealed-ampule approach or ZnO-SiO_2 for the open-tube approach. The n-type dopants in gallium arsenide include selenium and tellurium.

13.1.1 Diffusion Equation

Diffusion in a semiconductor can be visualized as atomic movement of the diffusant (dopant atoms) in the crystal lattice by vacancies or interstitials. Figure 3 shows the two basic atomic diffusion models in a solid.[1,3] The open circles represent the host atoms occupying the equilibrium lattice positions. The solid dots represent impurity atoms. At elevated temperatures, the lattice atoms vibrate around the equilibrium lattice sites. There is a finite probability that a host atom acquires sufficient energy to leave the lattice site and to become an interstitial atom, thereby creating a vacancy. When a neighboring impurity atom migrates

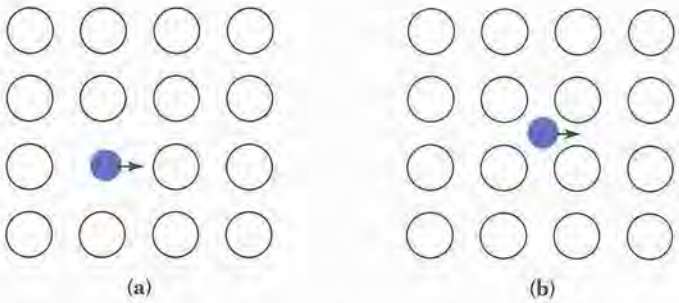


Fig. 3 Atomic diffusion mechanisms for a two-dimensional lattice.[1,3] (*a*) Vacancy mechanism; (*b*) interstitial mechanism.

to the vacancy site, as illustrated in Fig. 3*a*, the mechanism is called *vacancy diffusion*. If an interstitial atom moves from one place to another without occupying a lattice site (Fig. 3*b*), the mechanism is *interstitial diffusion*. An atom smaller than the host atom often moves interstitially.

The basic diffusion process of impurity atoms is similar to that of charge carriers (electrons and holes), discussed in Chapter 3. Accordingly, we define a flux F as the number of dopant atoms passing through a unit area in a unit time and C as the dopant concentration per unit volume. From Eq. 27 in Chapter 3, we have

$$F = -D\frac{\partial C}{\partial x}, \tag{3}$$

where we have substituted C for the carrier concentration and the proportionality constant D is the diffusion coefficient or diffusivity. Note that the basic driving force of the diffusion process is the concentration gradient dC/dx. The flux is proportional to the concentration gradient, and the dopant atoms will move (diffuse) away from a high-concentration region toward a lower-concentration region.

If we substitute Eq. 3 into the one-dimensional continuity equation, Eq. 56, given in Chapter 3 under the condition that no materials are formed or consumed in the host semiconductor (i.e., $G_n = R_n = 0$), we obtain

$$\frac{\partial C}{\partial t} = -\frac{\partial F}{\partial x} = \frac{\partial}{\partial x}\left(D\frac{\partial C}{\partial x}\right). \tag{4}$$

When the concentration of the dopant atoms is low, the diffusion coefficient can be considered to be independent of doping concentration, and Eq. 4 becomes

$$\boxed{\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2}.} \tag{5}$$

Equation 5 is often referred to as *Fick's diffusion equation*.

Figure 4 shows the measured diffusion coefficients for low concentrations of various dopant impurities in silicon and gallium arsenide.[4,5] The logarithm of the diffusion coefficient plotted against the reciprocal of the absolute temperature gives a straight line in most of the cases. This implies that over the temperature range, the diffusion coefficients can be expressed as

$$\boxed{D = D_0 \exp\left(\frac{-E_a}{kT}\right),} \tag{6}$$

where D_0 is the diffusion coefficient in cm^2/s extrapolated to infinite temperature and E_a is the activation energy in eV.

For the interstitial diffusion model, E_a is related to the energies required to move dopant atoms from one interstitial site to another. The values of E_a are found to be between 0.5 and 2 eV in both silicon and gallium arsenide. For the vacancy diffusion model, E_a is related to both the energies of motion and the energies of formation of vacancies. Thus,

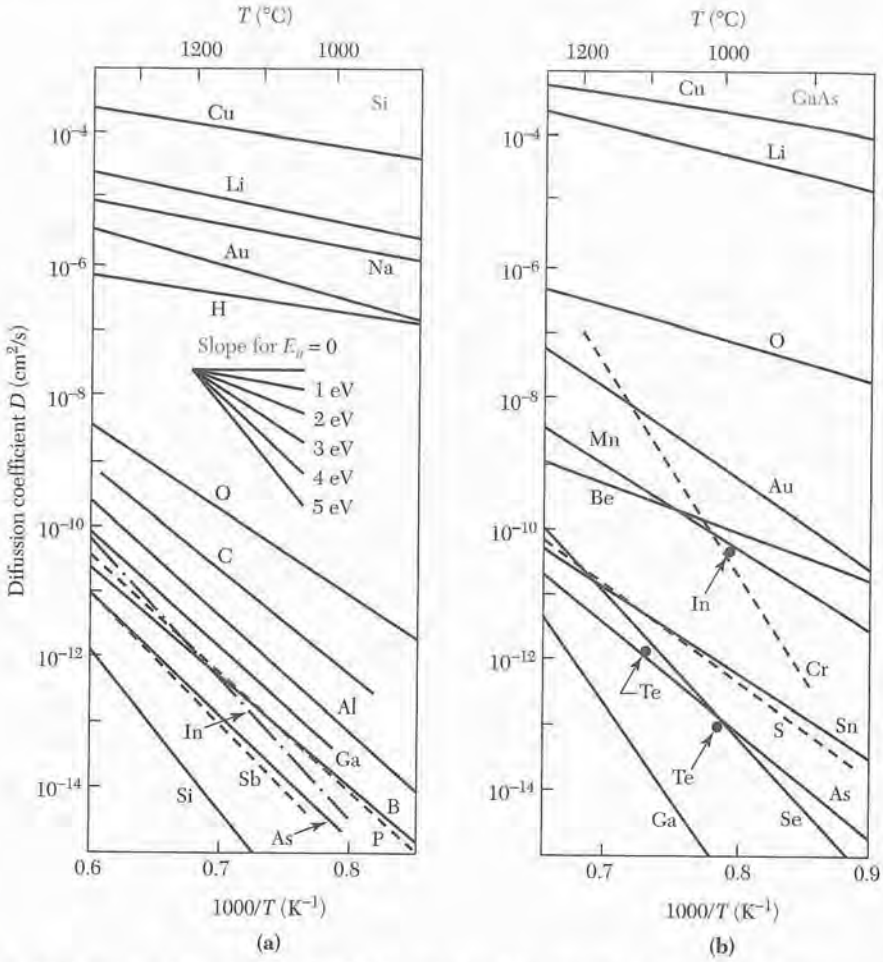


Fig. 4 Diffusion coefficient (also called diffusivity) as a function of the reciprocal of temperature for (*a*) silicon and (*b*) gallium arsenide.[4,5]

E_a for vacancy diffusion is larger than that for interstitial diffusion, usually between 3 and 5 eV.

For fast diffusants, such as Cu in Si and GaAs, shown in the upper portion of Fig. 4*a* and 4*b*, the measured activation energies are less than 2 eV, and interstitial atomic movement is the dominant diffusion mechanism. For slow diffusants, such as As in Si and GaAs, shown in the lower portion of Fig. 4*a* and 4*b*, E_a is larger than 3 eV, and vacancy diffusion is the dominant mechanism.

13.1.2 Diffusion Profiles

The diffusion profile of the dopant atoms is dependent on the initial and boundary conditions. In this subsection we consider two important cases, namely, constant–surface-concentration diffusion and constant–total-dopant diffusion. In the first case, impurity atoms are transported from a vapor source onto the semiconductor surface and diffused into the semiconductor wafers. The vapor source maintains a constant level of surface

concentration during the entire diffusion period. In the second case, a fixed amount of dopant is deposited onto the semiconductor surface and is subsequently diffused into the wafers.

Constant–Surface-Concentration Diffusion

The initial condition at $t = 0$ is

$$C(x, 0), \tag{7}$$

which states that the dopant concentration in the host semiconductor is initially zero. The boundary conditions are

$$C(0, t) = C_s \tag{8a}$$

and

$$C(\infty, t) = 0 \tag{8b}$$

where C_s is the surface concentration (at $x = 0$), which is independent of time. The second boundary condition states that at large distances from the surface there are no impurity atoms.

The solution of the diffusion equation (Eq. 5) that satisfies the initial and boundary conditions is given by [6]

$$\boxed{C(x,t) = C_s \operatorname{erfc}\left(\frac{x}{2\sqrt{Dt}}\right),} \tag{9}$$

where erfc is the complementary error function and $\sqrt{Dt}$ is the diffusion length. The definition of erfc and some properties of the function are summarized in Table 1. The diffusion profile for the constant–surface-concentration condition is shown in Fig. 5*a*, where we plot, on both linear (upper) and logarithmic (lower) scales, the normalized concentration as a function of depth for three values of the diffusion length $\sqrt{Dt}$ corresponding to three consecutive diffusion times and a fixed D for a given diffusion temperature. Note that as the time progresses, the dopant penetrates deeper into the semiconductor.

The total number of dopant atoms per unit area of the semiconductor is given by

$$Q(t) = \int_0^\infty C(x,t)dx. \tag{10}$$

Substituting Eq. 9 into Eq. 10 yields

$$Q(t) = \frac{2}{\sqrt{\pi}} C_s \sqrt{Dt} \cong 1.13\, C_s \sqrt{Dt}. \tag{11}$$

This expression can be interpreted as follows. The quantity $Q(t)$ represents the area under one of the diffusion profiles of the linear plot in Fig. 5*a*. These profiles can be approximated by triangles with height C_s and base $2\sqrt{Dt}$. This leads to $Q(t) \cong C_s\sqrt{Dt}$, which is close to the exact result obtained from Eq. 11.

A related quantity is the gradient of the diffusion profile dC/dx. The gradient can be obtained by differentiating Eq. 9:

$$\left.\frac{dC}{dx}\right|_{x,t} = -\frac{C_s}{\sqrt{\pi Dt}} e^{-x^2/4Dt}. \tag{12}$$

TABLE 1 Error Function Algebra

$$\mathrm{erf}(x) \equiv \frac{2}{\sqrt{\pi}}\int_0^x e^{-y^2}dy$$

$$\mathrm{erfc}(x) \equiv 1-\mathrm{erf}(x)$$

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

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

$$\mathrm{erf}(x) \cong \frac{2}{\sqrt{\pi}}x \quad \text{for } x \ll 1$$

$$\mathrm{erfc}(x) \cong \frac{1}{\sqrt{\pi}}\frac{e^{-x^2}}{x} \quad \text{for } x \gg 1.$$

$$\frac{d}{dx}\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}e^{-x^2}$$

$$\frac{d^2}{dx^2}\mathrm{erf}(x) = -\frac{4}{\sqrt{\pi}}xe^{-x^2}$$

$$\int_0^x \mathrm{erfc}(y')dy' = x\,\mathrm{erfc}(x)+\frac{1}{\sqrt{\pi}}(1-e^{-x^2})$$

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

EXAMPLE 1

For a boron diffusion in silicon at 1000°C, the surface concentration is maintained at $10^{19}\,\mathrm{cm}^{-3}$ and the diffusion time is 1 hour. Find $Q(t)$ and the gradient at $x = 0$ and at a location where the dopant concentration reaches $10^{15}\,\mathrm{cm}^{-3}$.

SOLUTION The diffusion coefficient of boron at 1000°C, as obtained from Fig. 4, is about $2\times10^{14}\,\mathrm{cm}^2/\mathrm{s}$, so that the diffusion length is

$$\sqrt{Dt} = \sqrt{2\times10^{-14}\times3600} = 8.48\times10^{-6}\,\mathrm{cm},$$

$$Q(t) = 1.13C_s\sqrt{Dt} = 1.13\times10^{19}\times8.48\times10^{-6} = 9.5\times10^{13}\ \text{atoms}/\mathrm{cm}^2.$$

$$\left.\frac{dC}{dx}\right|_{x=0} = -\frac{C_s}{\sqrt{\pi Dt}} = \frac{-10^{19}}{\sqrt{\pi}\times8.48\times10^{-6}} = -6.7\times10^{23}\,\mathrm{cm}^{-4}.$$

When $C = 10^{15}\,\mathrm{cm}^{-3}$, the corresponding distance x_j is given by Eq. 9, or

$$x_j = 2\sqrt{Dt}\ \mathrm{erfc}^{-1}\left(\frac{10^{15}}{10^{19}}\right) = 2\sqrt{Dt}\ (2.75) = 4.66\times10^{-5}\,\mathrm{cm} = 0.466\ \mu\mathrm{m},$$

$$\left.\frac{dC}{dx}\right|_{x=0.466\ \mu m} = -\frac{C_s}{\sqrt{\pi Dt}}e^{-x_j^2/4Dt} = -3.5\times10^{20}\,\mathrm{cm}^{-4}.$$

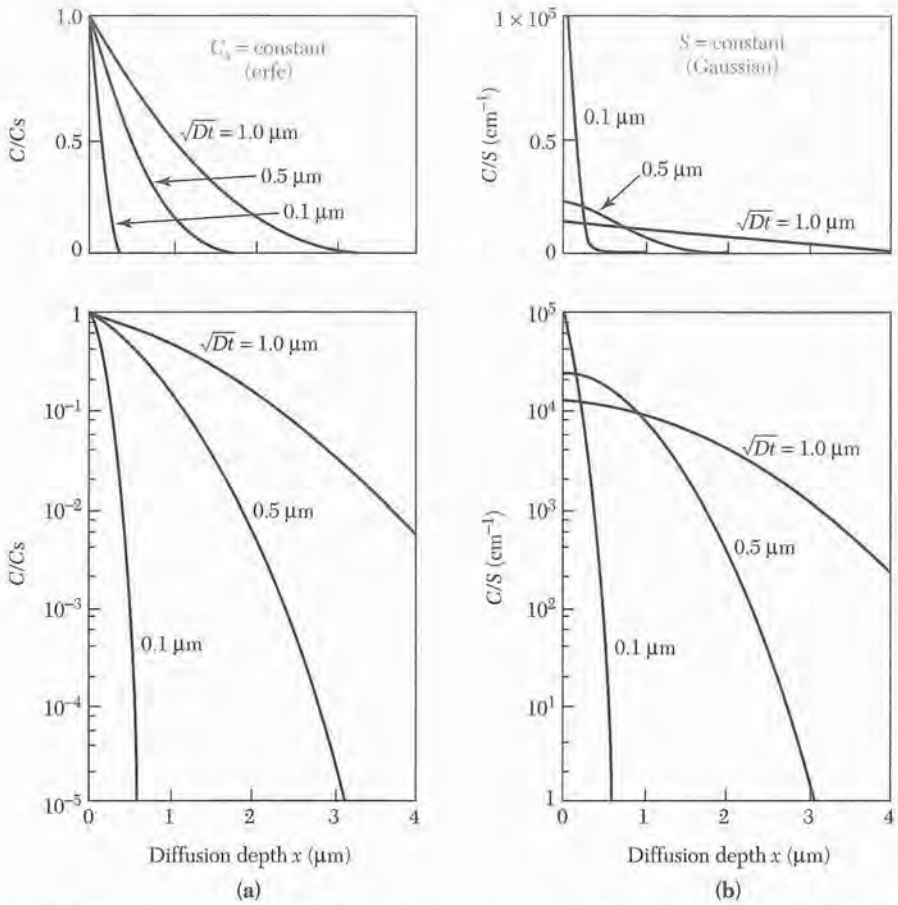


Fig. 5 Diffusion profiles. (*a*) Normalized complementary error function versus distance for successive diffusion times. (*b*) Normalized Gaussian function versus distance.

Constant–Total-Dopant Diffusion

For this case, a fixed (or constant) amount of dopant is deposited onto the semiconductor surface in a thin layer, and the dopant subsequently diffuses into the semiconductor. The initial condition is the same as in Eq. 7. The boundary conditions are

$$\int_0^\infty C(x,t)dx = S \tag{13a}$$

and

$$C(\infty, t) = 0 \tag{13b}$$

where S is the total amount of dopant per unit area.

The solution of the diffusion equation, Eq. 5, that satisfies the above conditions is

$$\boxed{C(x,t) = \frac{S}{\sqrt{\pi Dt}} \exp\left(-\frac{x^2}{4Dt}\right).} \tag{14}$$

This expression is the Gaussian distribution. Since the dopant will move into the semiconductor as time increases, to keep the total dopant S constant, the surface concentration must decrease. This is indeed the case, since the surface concentration is given by Eq. 14 with $x = 0$:

$$C_s(t) = \frac{S}{\sqrt{\pi D t}}. \tag{15}$$

Figure 5*b* shows the dopant profile for a Gaussian distribution where we plot the normalized concentration (C/S) as a function of the distance for three increasing diffusion lengths. Note the reduction of the surface concentration as the diffusion time increases. The gradient of the diffusion profile is obtained by differentiating Eq. 14 and is

$$\left.\frac{dC}{dx}\right|_{x,t} = -\frac{xS}{2\sqrt{\pi}(Dt)^{3/2}} e^{-x^2/4Dt} = -\frac{x}{2Dt} C(x,t). \tag{16}$$

The gradient (or slope) is zero at $x = 0$ and at $x = \infty$ and the maximum gradient occurs at $x = \sqrt{2Dt}$.

In integrated-circuit processing, a two-step diffusion process is commonly used, in which a *predeposition* diffused layer is first formed under a constant–surface-concentration condition. This step is followed by a *drive-in* diffusion (also called *redistribution* diffusion) under a constant–total-dopant condition. For most practical cases, the diffusion length $\sqrt{Dt}$ for the predeposition diffusion is much smaller than the diffusion length for the drive-in diffusion. Therefore, we can consider the predeposition profile as a delta function at the surface, and we can regard the extent of the penetration of the predeposition profile to be negligibly small compared with that of the final profile that results from the drive-in step.

EXAMPLE 2

Arsenic was predeposited by arsine gas and the resulting total amount of dopant per unit area is 1×10^{14} atoms/cm^2. How long would it take to drive the arsenic in to a junction depth of 1 μm? Assume a background doping of 1×10^{15} atoms/cm^3, and a drive-in temperature of 1200°C. For As diffusion, $D_0 = 24$ cm^2/s, and $E_a = 4.08$ eV.

SOLUTION

$$D = D_0 \exp\left(\frac{-E_a}{kT}\right) = 24 \exp\left(\frac{-4.08}{8.614\times10^{-5}\times1473}\right) = 2.602\times10^{-13}\ \text{cm}^2/\text{s},$$

$$x_j^2 = 10^{-8} = 4Dt \ln\left(\frac{S}{C_B\sqrt{\pi D t}}\right) = 1.04\times10^{-12}\, t \ln\left(\frac{1.106\times10^5}{\sqrt{t}}\right),$$

$t \cdot \log t - 10.09t + 8350 = 0.$

The solution to the above equation can be determined by the cross point of equation $y = t \cdot \log t$ and $y = 10.09t - 8350$.

Therefore, $t = 1190$ seconds $\cong 20$ minutes.

13.1.3 Evaluation of Diffused Layers

The results of a diffusion process can be evaluated by three measurements—the junction depth, the sheet resistance, and the dopant profile of the diffused layer. The junction depth can be delineated by cutting a groove into the semiconductor and etching the surface with a solution (e.g., 100 cm^3 HF and a few drops of HNO_3 for silicon) that stains the p-type region darker than the n-type region, as illustrated in Fig. 6a. If R_0 is the radius of the tool used to form the groove, then the junction depth x_j is given by

$$x_j = \sqrt{R_0^2 - b^2} - \sqrt{R_0^2 - a^2}, \tag{17}$$

where a and b are indicated in the figure. In addition, if R_0 is much larger than a and b, then

$$x_j \cong \frac{a^2 - b^2}{2R_0}. \tag{18}$$

The junction depth x_j as illustrated in Fig. 6b is the position where the dopant concentration equals the substrate concentration C_B, or

$$C(x_j) = C_B. \tag{19}$$

Thus, if the junction depth and C_B are known, the surface concentration C_s and the impurity distribution can be calculated, provided the diffusion profile follows one or the other simple equation derived in Section 13.1.2.

The resistance of a diffused layer can be measured by the four-point probe technique described in Chapter 3. The *sheet resistance* R is related to the junction depth x_j, the carrier mobility μ (which is a function of the total impurity concentration), and the impurity distribution $C(x)$ by the following expression:[7]

$$\boxed{R = \frac{1}{q\int_0^{x_j} \mu C(x)dx}}. \tag{20}$$

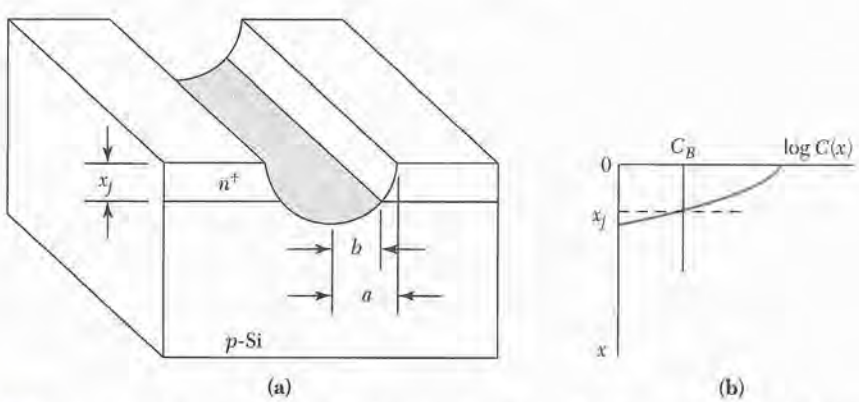


Fig. 6 Junction-depth measurement. (a) Grooving and staining; (b) position in which dopant and substrate concentration are equal.

For a given diffusion profile, the average resistivity $\overline{\rho} = Rx_j$ is uniquely related to the surface concentration C_s and the substrate-doping concentration for an assumed diffusion profile. Design curves relation C_s and $\overline{\rho}$ have been calculated for simple diffusion profiles, such as the erfc or Gaussian distribution.[8] To use these curves correctly we must be sure that the diffusion profiles agree with the assumed profiles. For low concentration and deep diffusions, the diffusion profiles generally can be represented by the aforementioned simple functions. However, as we discuss in the next section, for high concentration and shallow diffusions, the diffusion profiles cannot be represented by these simple functions.

The diffusion profile can be measured using the capacitance-voltage technique described in Chapter 7. The majority carrier profile, which is equal to the impurity profile if impurities are fully ionized, can be determined by measuring the reverse-bias capacitance of a *p-n* junction or a Schottky barrier diode as a function of the applied voltage. A more elaborate method is the secondary-ion–mass spectroscope (SIMS) technique, which measures the total impurity profile. In the SIMS technique, an ion beam sputters material off the surface of a semiconductor, and the ion component is detected and mass analyzed. This technique has high sensitivity to many elements, such as boron and arsenic, and it is an ideal tool for providing the precision needed for profile measurements in high-concentration or shallow-junction diffusions.[9]

13.2 EXTRINSIC DIFFUSION

The diffusion profiles described in Section 13.1 are for constant diffusivities. These profiles occur when the doping concentration is lower than the intrinsic-carrier concentration $n_i(T)$ at the diffusion temperature. For example, at $T = 1000°C$, n_i equals 5×10^{18} cm^{-3} for silicon and 5×10^{17} cm^{-3} for gallium arsenide. The diffusivity at low concentrations is often referred to as the intrinsic diffusivity. Doping profiles that have concentrations less than n_i (T) are in the *intrinsic* diffusion region as indicated in the left side of Fig. 7. In this region, the resulting dopant profiles of sequential or simultaneous diffusions of *n*- and *p*-type impurities can be determined by superposition, that is, the diffusions can be treated independently. However, when the impurity concentration, including both the substrate and the dopant, is greater than $n_i(T)$, the semiconductor becomes extrinsic and the diffusivity is considered to be extrinsic. In the extrinsic diffusion region the diffusivity becomes concentration dependent.[10] In the extrinsic diffusion region the diffusion profiles are more complicated, and there are interactions and cooperative effects among the sequential or simultaneous diffusions.

13.2.1 Concentration-Dependent Diffusivity

As mentioned previously, when a host atom acquires sufficient energy from the lattice vibration to leave its lattice site, a vacancy is created. Depending on the charges associated with a vacancy, we can have a neutral vacancy V^0, an acceptor vacancy V^-, a double-charged acceptor vacancy V^{2-}, a donor vacancy V^+, and so forth. We expect that the vacancy density of a given charge state (i.e., the number of vacancies per unit volume, C_V) has a temperature dependence similar to that of the carrier density (see to Eq. 28 in Chapter 2), that is ,

$$C_V = C_i \exp\left(\frac{E_F - E_i}{kT}\right), \tag{21}$$

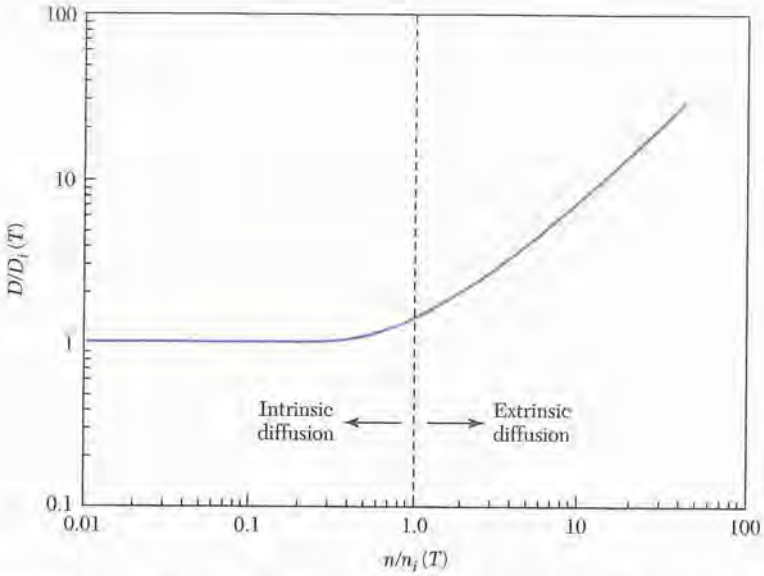


Fig. 7 Donor-impurity diffusivity versus electron concentration showing regions of intrinsic and extrinsic diffusion.[10]

where C_i is the intrinsic vacancy density, E_F is the Fermi level, and E_i is the intrinsic Fermi level.

If the dopant diffusion is dominated by the vacancy mechanism, the diffusion coefficient is expected to be proportional to the vacancy density. At low doping concentrations ($n < n_i$), the Fermi level coincides with the intrinsic Fermi level ($E_F = E_i$). The vacancy density is equal to C_i and is independent of doping concentration. The diffusion coefficient, which is proportional to C_i, also is independent of doping concentration. At high concentrations ($n > n_i$), the Fermi level will move toward the conduction band edge (for donor-type vacancies), and the term $[\exp(E_F - E_i)/kT]$ becomes larger than unity. This causes C_V to increase, which in turn causes the diffusion coefficient to increase, as shown in the right side of Fig. 7.

When the diffusion coefficient varies with dopant concentration, Eq. 4 should be used as the diffusion equation instead of Eq. 5, in which D is independent of C. We consider the case where the diffusion coefficient can be written as

$$D = D_s\left(\frac{C}{C_s}\right)^{\gamma}, \tag{22}$$

where C_s is the surface concentration, D_s is the diffusion coefficient at the surface, and γ is a parameter to describe the concentration dependence. For such a case, we can write the diffusion equation, Eq. 4, as an ordinary differential equation and solve it numerically.

Figure 8 shows the solutions[11] for a constant–surface-concentration diffusion with different values of γ. For $\gamma = 0$, we have the case of constant diffusivity and the profile is the same as that shown in Fig. 5*a*. For $\gamma > 0$ the diffusivity decreases as the concentration decreases, and increasingly steep and box-like concentration profiles result for increasing γ. Therefore, highly abrupt junctions are formed when diffusions are made into a background of an opposite impurity type. The abruptness of the doping profile

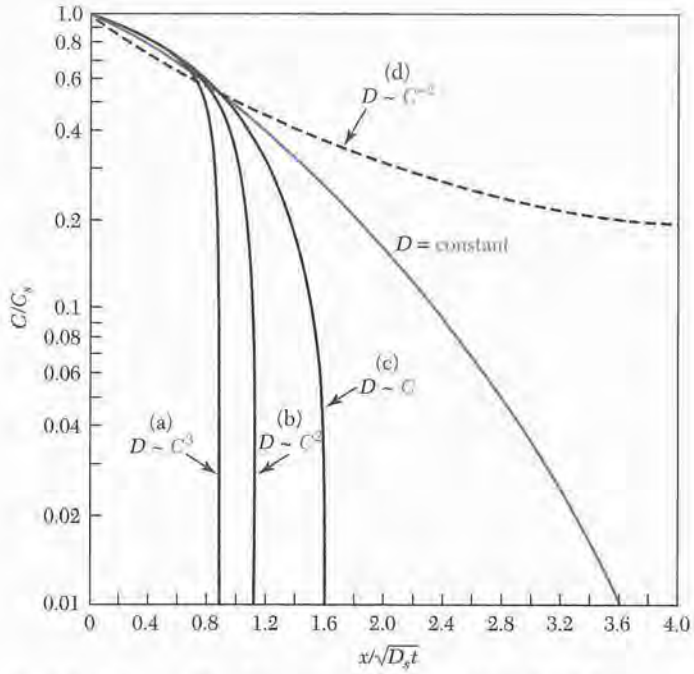


Fig. 8 Normalized diffusion profiles for extrinsic diffusion where the diffusion coefficient becomes concentration dependent.[10,11]

results in a junction depth virtually independent of the background concentration. Note that the junction depth (see Fig. 8) is given by

$$x_j = 1.6\sqrt{D_s t} \quad \text{for } D \sim C(\gamma = 1),$$
$$x_j = 1.1\sqrt{D_s t} \quad \text{for } D \sim C^2(\gamma = 2), \tag{23}$$
$$x_j = 0.87\sqrt{D_s t} \quad \text{for } D \sim C^3(\gamma = 3).$$

In the case of $\gamma = -2$, the diffusivity increases with decreasing concentration, which leads to a concave profile, as opposed to to the convex profiles for other cases.

13.2.2 Diffusion Profiles

Diffusion in Silicon

The measured diffusion coefficients of boron and arsenic in silicon have a concentration dependence with $\gamma \cong 1$. Their concentration profiles are abrupt, as depicted in curve *c* of Fig. 8. For gold and platinum diffusion in silicon, γ is close to –2, and their concentration profiles have the concave shape shown in curve *d* of Fig. 8.

The diffusion of phosphorus in silicon is associated with the doubly charged acceptor vacancy V^{2-}, and the diffusion coefficient at high concentration varies as C^2. We would expect that the diffusion profile of phosphorus resembles that shown in curve *b* of Fig. 8. However, because of a *dissociation effect*, the diffusion profile exhibits anomalous behavior.

Figure 9 shows phosphorus diffusion profiles for various surface concentrations after diffusion into silicon for 1 hour at 1000°C.[12] When the surface concentration is low, cor-

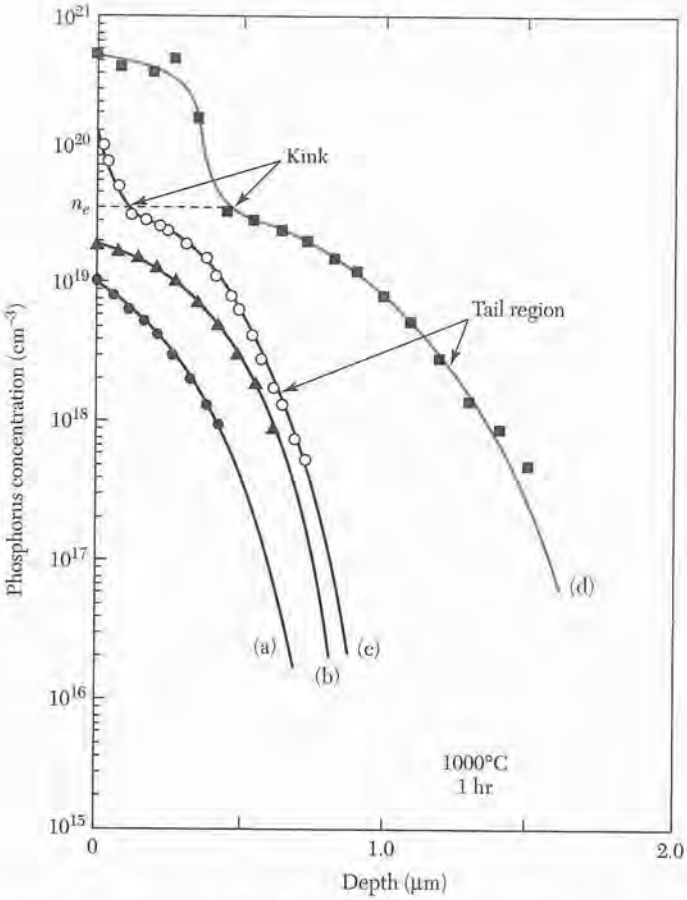


Fig. 9 Phosphorus diffusion profiles[12] for various surface concentrations after diffusion into silicon for 1 hour at 1000°C.

responding to the intrinsic diffusion region, the diffusion profile is given by an erfc (curve *a*). As the concentration increases, the profile begins to deviate from the simple expression (curves *b* and *c*). At very high concentration (curve *d*), the profile near the surface is indeed similar to that shown in curve *b* of Fig. 8. However, at concentration n_e, a kink occurs and is followed by a rapid diffusion in the tail region. The concentration n_e corresponds to a Fermi level 0.11 eV below the conduction band. At this energy level, the coupled impurity-vacancy pair (P^+V^{2-}) dissociates to P^+, V^-, and an electron. Thus, the dissociation generates a large number of singly charged acceptor vacancies V^-, which in turn enhances the diffusion in the tail region of the profile. The diffusivity in the tail region is over 10^{-12} cm^2/s, which is about two orders of magnitude larger than the intrinsic diffusivity at 1000°C. Because of its high diffusivity, phosphorus is commonly used to form deep junctions, such as the *n*-tubs in CMOS.

Zinc Diffusion in Gallium Arsenide

We expect diffusion in gallium arsenide to be more complicated than that in silicon because the diffusion of impurities may involve atomic movements on both the gallium and arsenic sublattices. Vacancies play a dominant role in diffusion processes in gallium arsenide

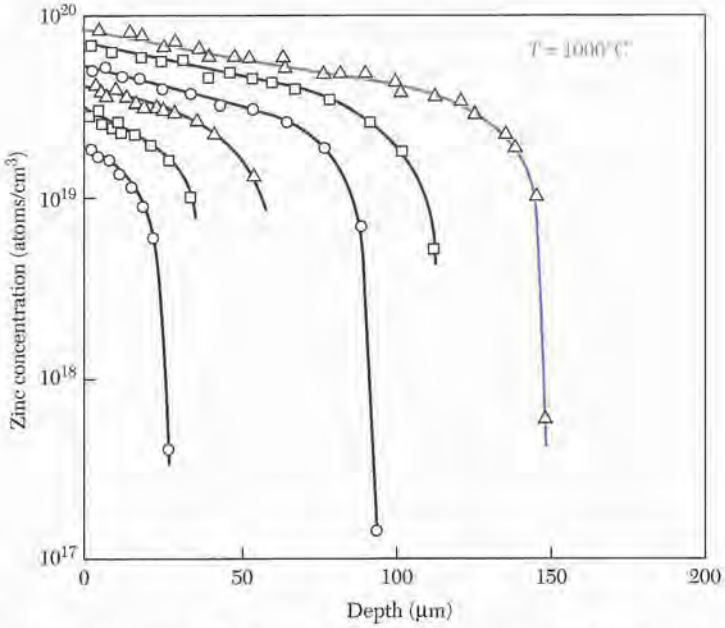


Fig. 10 Diffusion profiles[13] of zinc in GaAs after annealing at 1000°C for 2.7 hours. The different surface concentrations are obtained by maintaining the Zn source at temperatures in the range 600°–800°C.

because both p- and n-type impurities must ultimately reside in lattice sites. However, the charge states of the vacancies have not been established.

Zinc is the most extensively studied diffusant in gallium arsenide. Its diffusion coefficient is found to vary as C^2. Therefore, the diffusion profiles are steep, as shown [13] in Fig. 10, and resemble curve b of Fig. 8. Note that even for the case of the lowest surface concentration, the diffusion is in the extrinsic-diffusion region, because n_i for GaAs at 1000°C is less than 10^{18} cm^{-3}. As can be seen in Fig. 10, the surface concentration has a profound effect on the junction depth. The diffusivity varies linearly with the partial pressure of the zinc vapor, and the surface concentration is proportional to the square root of the partial pressure. Therefore, from Eq. 23, the junction depth is linearly proportional to the surface concentration.

13.3 DIFFUSION-RELATED PROCESSES

In this section we consider two processes in which diffusion plays an important role and the impact of these processes on device performances.

13.3.1 Lateral Diffusion

The one-dimensional diffusion equation discussed previously can describe satisfactorily the diffusion process, except at the edge of the mask window. Here the impurities will diffuse downward and sideways (i.e., laterally). In this case, we must consider a two-dimensional diffusion equation and use a numerical technique to obtain the diffusion profiles under different initial and boundary conditions.

Figure 11 shows the contours of constant doping concentration for a constant–surface-concentration diffusion condition assuming that the diffusivity is independent of concentration.[14] At the far right of the figure, the variation of the dopant concentration from $0.5\ C_s$ to $10^{-4}\ C_s$ (where C_s is the surface concentration) corresponds to the erfc distribution given by Eq. 9. The contours are in effect a map of the location of the junctions created by diffusing into various background concentrations. For example, at $C/C_s = 10^{-4}$ (i.e., the background doping is 10^4 times lower than the surface concentration), we see from this constant-concentration curve that the vertical penetration is about 2.8 μm, whereas the lateral penetration is about 2.3 μm (i.e., the penetration along the diffusion mask-semiconductor interface). Therefore, the lateral penetration is about 80% of the penetration in the vertical direction for concentrations three or more orders of magnitude below the surface concentration. Similar results are obtained for a constant–total-dopant diffusion condition. The ratio of lateral to vertical penetration is about 75%. For concentration-dependent diffusivities, the ratio is found to be reduced slightly, to about 65%–70%.

Because of the lateral-diffusion effect, the junction consists of a central plane (or flat) region with approximately cylindrical edges with a radius of curvature r_j, as shown in Fig. 11. In addition, if the diffusion mask contains sharp corners, the shape of the junction near the corner will be roughly spherical because of lateral diffusion. Since the electric-field intensities are higher for cylindrical and spherical junction regions, the avalanche breakdown voltages of such regions can be substantially lower than that of a plane junction having the same background doping. This junction " curvature effect" was discussed in Chapter 4.

13.3.2 Impurity Redistribution During Oxidation

Dopant impurities near the silicon surface will be redistributed during thermal oxidation. The redistribution depends on several factors. When two solid phases are brought together, an impurity in one solid will redistribute between the two solids until it reaches equilibrium. This is similar to our previous discussion on impurity redistribution in crystal growth

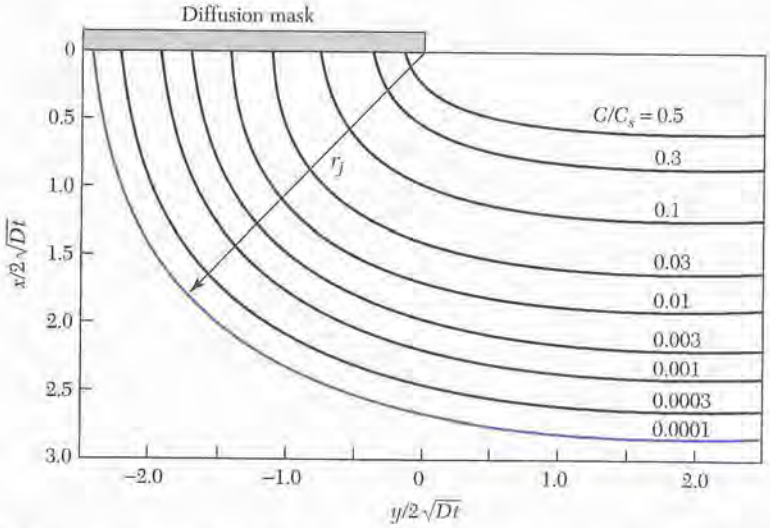


Fig. 11 Diffusion contours at the edge of an oxide window, where r_j is the radius of curvature.[14]

from the melt. The ratio of the equilibrium concentration of the impurity in the silicon to that in the silicon dioxide is called the *segregation coefficient* and is defined as

$$k = \frac{\text{equilibrium concentration of impurity in silicon}}{\text{equilibrium concentration of impurity in } SiO_2}. \tag{24}$$

A second factor that influences impurity distribution is that the impurity may diffuse rapidly through the silicon dioxide and escape to the gaseous ambient. If the diffusivity of the impurity in silicon dioxide is large, this factor will be important. A third factor in the redistribution process is that the oxide is growing, and thus, the boundary between the silicon and the oxide is advancing into the silicon as a function of time. The relative rate of this advance compared with the diffusion rate of the impurity through the oxide is important in determining the extent of the redistribution. Note that even if the segregation coefficient of an impurity k equals unity, some redistribution of the impurity in the silicon will still take place. As indicated in Fig. 3 of Chapter 11, the oxide layer will be about twice as thick as the silicon layer it replaced. Therefore, the same amount of impurity will now be distributed in a larger volume, resulting in a depletion of the impurity from the silicon.

Four possible redistribution processes are illustrated[6] in Fig. 12. These processes can be classified into two groups. In one group the oxide takes up the impurity (Fig. 12*a* and

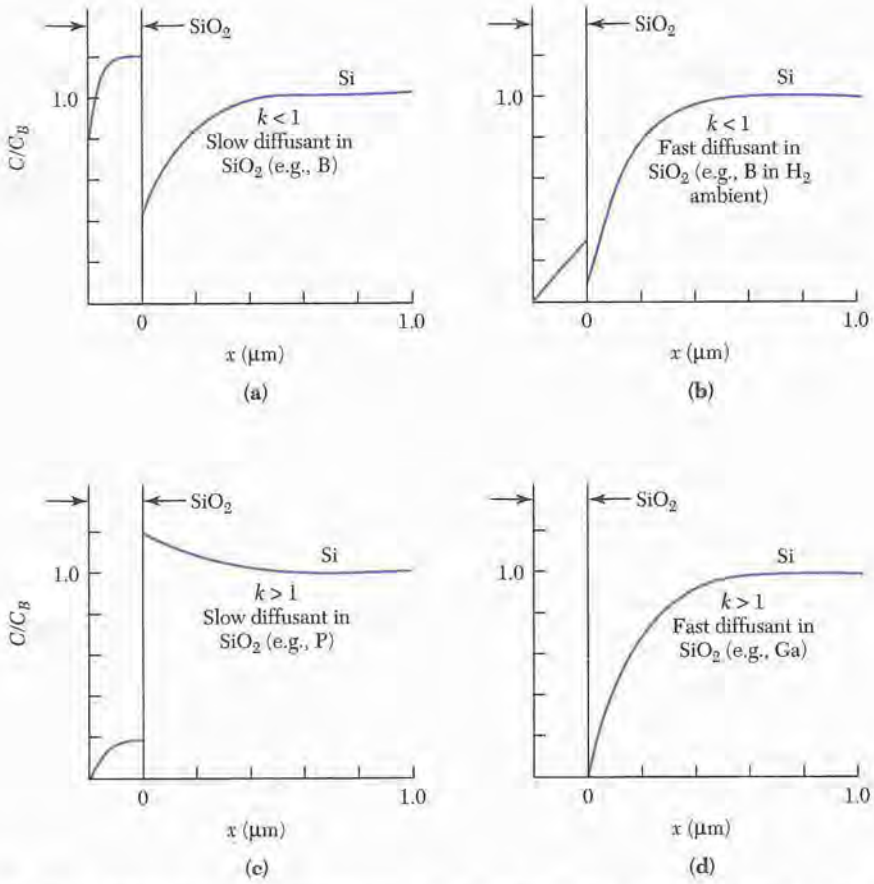


Fig. 12 Four different cases of impurity redistribution in silicon due to thermal oxidation.[6]

b for $k < 1$), and in the other the oxide rejects the impurity (Fig. 12*c* and *d* for $k > 1$). In each case, what happens depends on how rapidly the impurity can diffuse through the oxide. In group 1, the silicon surface is depleted of impurities; an example is boron with k approximately equal to 0.3. Rapid diffusion of the impurity through the silicon dioxide increases the amount of depletion; an example is boron-doped silicon heated in a hydrogen ambient because hydrogen in silicon dioxide enhances the diffusivity of boron. In group 2, k is greater than unity, so that the oxide rejects the impurity. If diffusion of the impurity through the silicon dioxide is relatively slow, the impurity piles up near the silicon surface; an example is phosphorus, with k approximately equal to 10. When diffusion through the silicon dioxide is rapid, so much impurity may escape from the solid to the gaseous ambient that the overall effect will be a depletion of the impurity; an example is gallium, with k approximately equal to 20.

The redistributed dopant impurities in silicon dioxide are seldom electrically active. However, redistribution in silicon has an important effect on processing and device performance. For example, nonuniform dopant distribution will modify the interpretation of the measurements of interface-trap properties (see Chapter 6), and the change of the surface concentration will modify the threshold voltage and device contact resistance (see Chapter 7).

13.4 RANGE OF IMPLANTED IONS

Ion implantation is the introduction of energetic, charged particles into a substrate such as silicon. Implantation energies are between 1 keV and 1 MeV, resulting in ion distributions with average depths ranging from 10 nm to 10 μm. Ion doses vary from 10^{12} ions/cm^2 for threshold voltage adjustment to 10^{18} ions/cm^2 for the formation of buried insulating layer. Note that the dose is expressed as the number of ions implanted into 1 cm^2 of the semiconductor surface area. The main advantages of ion implantation are its more precise control and reproducibility of impurity dopings and its lower processing temperature compared with those of the diffusion process.

Figure 13 shows schematically a medium-energy ion implantor.[15] The ion source has a heated filament to break up source gas such as BF_3 or AsH_3 into charged ions (B^+ or As^+). An extraction voltage, around 40 kV, causes the charged ions to move out of the ion-source chamber into a mass analyzer. The magnetic field of the analyzer is chosen such that only ions with the desired mass-to-charge ratio can travel through it without being filtered. The selected ions then enter the acceleration tube, where they are accelerated to the implantation energy as they move from high voltage to ground. Apertures

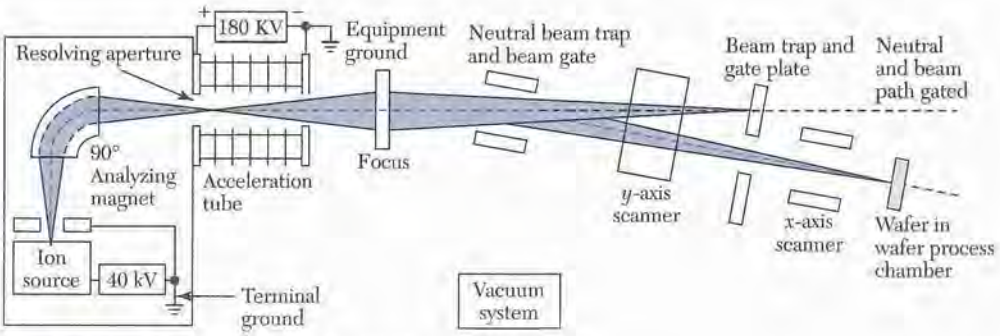


Fig. 13 Schematic of a medium-current ion implantor.

ensure that the ion beam is well collimated. The pressure in the implantor is kept below 10^{-4} Pa to minimize ion scattering by gas molecules. The ion beam is then scanned over the wafer surface using electrostatic deflection plates and is implanted into the semiconductor substrate.

The energetic ions lose their energies through collision with electrons and nuclei in the substrate and finally come to rest at some depth within the lattice. The average depth can be controlled by adjusting the acceleration energy. The dopant dose can be controlled by monitoring the ion current during implantation. The principle side effect is the disruption or damage of the semiconductor lattice due to ion collisions. Therefore, a subsequent annealing treatment is needed to remove these damages.

13.4.1 Ion Distribution

The total distance that an ion travels in coming to rest is called its *range* R and is illustrated[16] in Fig. 14*a*. The projection of this distance along the axis of incidence is called the *projected range* R_p. Because the number of collisions per unit distance and the energy lost per collision are random variables, there will be a spatial distribution of ions having

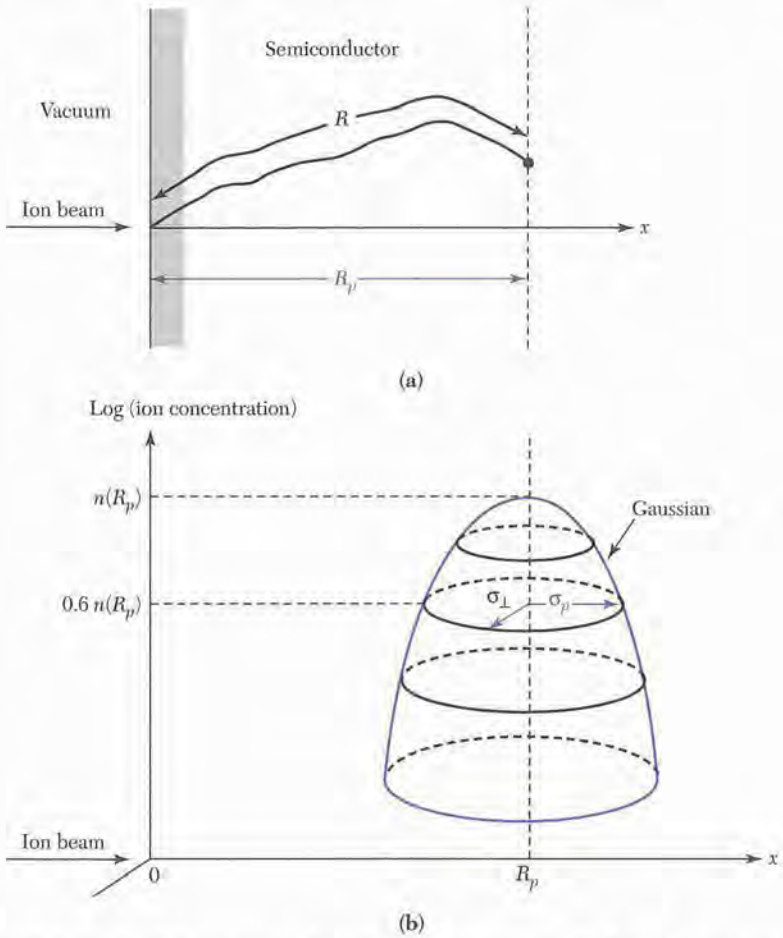


Fig. 14 (*a*) Schematic of the ion range R and projected range σ_p. (*b*) Two-dimensional distribution of the implanted ions.[16]

the same mass and the same initial energy. The statistical fluctuations in the projected range are called the *projected straggle* σ_p. There is also a statistical fluctuation along an axis perpendicular to the axis of incidence, called the *lateral straggle* $\sigma_\perp$.

Figure 14*b* shows the ion distribution. Along the axis of incidence, the implanted impurity profile can be approximated by a Gaussian distribution function:

$$\boxed{n(x) = \frac{S}{\sqrt{2\pi}\sigma_p} \exp\left[-\frac{(x-R_p)^2}{2\sigma_p^2}\right],} \tag{25}$$

where S is the ion dose per unit area. This equation is similar to Eq. 14 for constant–total-dopant diffusion, except that the quantity $4Dt$ is replaced by $2\sigma_p^2$ and the distribution is shifted along the x-axis by R_p. Thus, for diffusion, the maximum concentration is at $x = 0$, whereas for ion implantation the maximum concentration is at the projected range R_p. The ion concentration is reduced by 40% from its peak value at $(x - R_p) = \pm\sigma_p$, by one decade at $\pm 2\sigma_p$, by two decades at $\pm 3\sigma_p$, and by five decades at $\pm 4.8\sigma_p$.

Along the axis perpendicular to the axis of incidence, the distribution is also a Gaussian function of the form $\exp(-y^2/2\sigma_\perp^2)$. Because of this distribution, there will be some lateral implantation.[17] However, the lateral penetration from the mask edge (on the order of $\sigma_\perp$) is considerably smaller than that from the thermal diffusion process discussed in Section 13.3.

13.4.2 Ion Stopping

There are two stopping mechanisms by which an energetic ion, on entering a semiconductor substrate (also called the target), can be brought to rest. The first is by transferring its energy to the target nuclei. This causes deflection of the incident ion and also dislodges many target nuclei from their original lattice sites. If E is the energy of the ion at any point x along its path, we can define a nuclear stopping power $S_n(E) \equiv (dE/dx)_n$ to characterize this process. The second stopping mechanism is by the interaction of the incident ion with the cloud of electrons surrounding the target's atoms. The ion loses energy in collisions with electrons through Coulombic interaction. The electrons can be excited to higher energy levels (excitation), or they can be ejected from the atom (ionization). We can define an electronic stopping power $S_e(E) \equiv (dE/dx)_e$ to characterize this process.

The average rate of energy loss with distance is given by a superposition of the above two stopping mechanisms:

$$\frac{dE}{dx} = S_n(E) + S_e(E). \tag{26}$$

If the total distance traveled by the ion before coming to rest is R, then

$$R = \int_0^R dx = \int_0^{E_0} \frac{dE}{S_n(E) + S_e(E)}, \tag{27}$$

where E_0 is the initial ion energy. The quantity R has been defined previously as the range.

We can visualize the nuclear stopping process by considering the elastic collision between an incoming hard sphere (energy E_0 and mass M_1) and a target hard sphere (initial energy zero and mass M_2), as illustrated in Fig. 15. When the spheres collide, momentum is transferred along the centers of the spheres. The deflection angle θ and the velocities v_1 and v_2 can be obtained from the requirements for conservation of momentum and

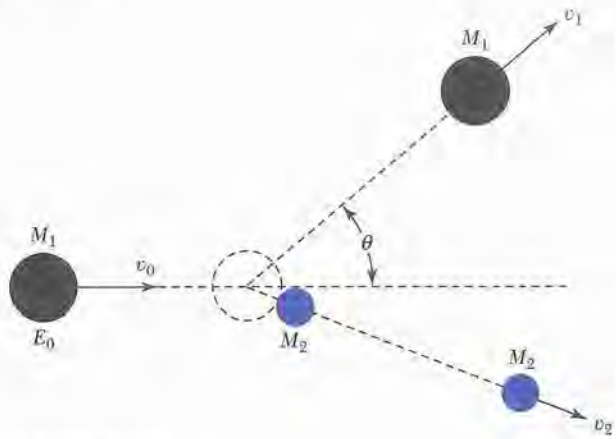


Fig. 15 Collision of hard spheres.

energy. The maximum energy loss is in a head-on collision. For this case, the energy loss by the incident particle M_1 or the energy transferred to M_2 is

$$\frac{1}{2}M_2 v_2^2 = \left[\frac{4M_1M_2}{(M_1+M_2)^2}\right]E_0. \tag{28}$$

Since M_2 is usually of the same order of magnitude as M_1, a large amount of energy can be transferred in nuclear stopping process.

Detailed calculations show that the nuclear stopping power increases linearly with energy at low energies (similar to Eq. 28), and $S_n(E)$ reaches a maximum at some intermediate energy. At high energies, $S_n(E)$ becomes smaller because fast particles may not have sufficient interaction time with the target atoms to achieve effective energy transfer. The calculated values of $S_n(E)$ for arsenic, phosphorus, and boron in silicon at various energies are shown in Fig. 16 (solid lines, where the superscript indicates the atomic weight).[18] Note that heavier atoms, such as arsenic, have larger nuclear stopping power, that is, larger energy loss per unit distance.

The electronic stopping power is found to be proportional to the velocity of the incident ion, or

$$S_e(E) = k_e\sqrt{E} \tag{29}$$

where the coefficient k_e is a relatively weak function of atomic mass and atomic number. The value of k_e is approximately 10^7 (eV)$^{1/2}$/cm for silicon and 3×10^7 (eV)$^{1/2}$/cm for gallium arsenide. The electronic stopping power in silicon is plotted in Fig. 16 (dotted line). Also shown in the figure are the crossover energies, at which $S_e(E)$ equals $S_n(E)$. For boron, which has a relatively low ion mass compared with the target silicon atom, the crossover energy is only 10 keV. This means that over most of the implantation energy range of 1 keV to 1 MeV, the main energy loss mechanism is due to electronic stopping. On the other hand, for arsenic with relatively high ion mass, the crossover energy is 700 keV. Thus, nuclear stopping dominates over most of the energy range. For phosphorus, the crossover energy is 130 keV. For an E_0 less than 130 keV, nuclear stopping will dominate; for higher energies, electronic stopping will take over.

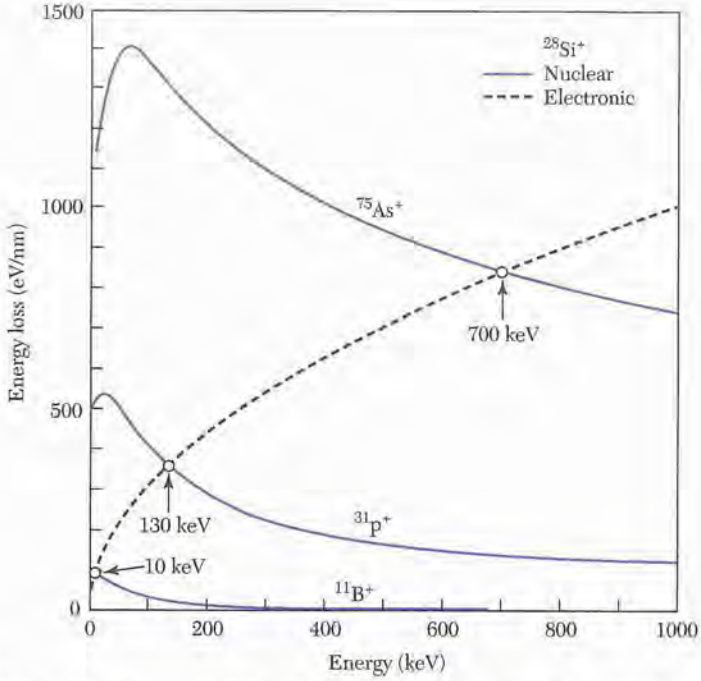


Fig. 16 Nuclear stopping power $S_n(E)$, and electronic stopping power $S_e(E)$ for As, P, and B in Si. The points of intersection of the curves correspond to the energy at which nuclear and electronic stopping are equal.[18]

Once $S_n(E)$ and $S_e(E)$ are known, we can calculate the range from Eq. 27. This in turn can give us the projected range and projected straggle with the help of the following approximate equations[15]:

$$R_p \cong \frac{R}{1+(M_2/3M_1)}, \tag{30}$$

$$\sigma_p \cong \frac{2}{3}\left[\frac{\sqrt{M_1 M_2}}{M_1+M_2}\right]R_p. \tag{31}$$

Figure 17*a* shows the projected range (R_p), the projected straggle (σ_p), and the lateral straggle ($\sigma_\perp$) for arsenic, boron, and phosphorus in silicon.[19] As expected, the larger the energy loss, the smaller the range. Also, the projected range and straggles increase with ion energy. For a given element at a specific incident energy, σ_p and $\sigma_\perp$ are comparable and usually within ±20%. Figure 17*b* shows the corresponding values for hydrogen, zinc, and tellurium in gallium arsenide.[17] If we compare Fig. 17*a* with Fig. 17*b*, we see that most of the popular dopants (except hydrogen) have larger projected ranges in silicon than they have in gallium arsenide.

EXAMPLE 3

Assume 100 keV boron implants on a 200 mm silicon wafer at a dose of 5×10^{14} ions/cm^2. Calculate the peak concentration and the required ion beam current for 1 minute of implantation.

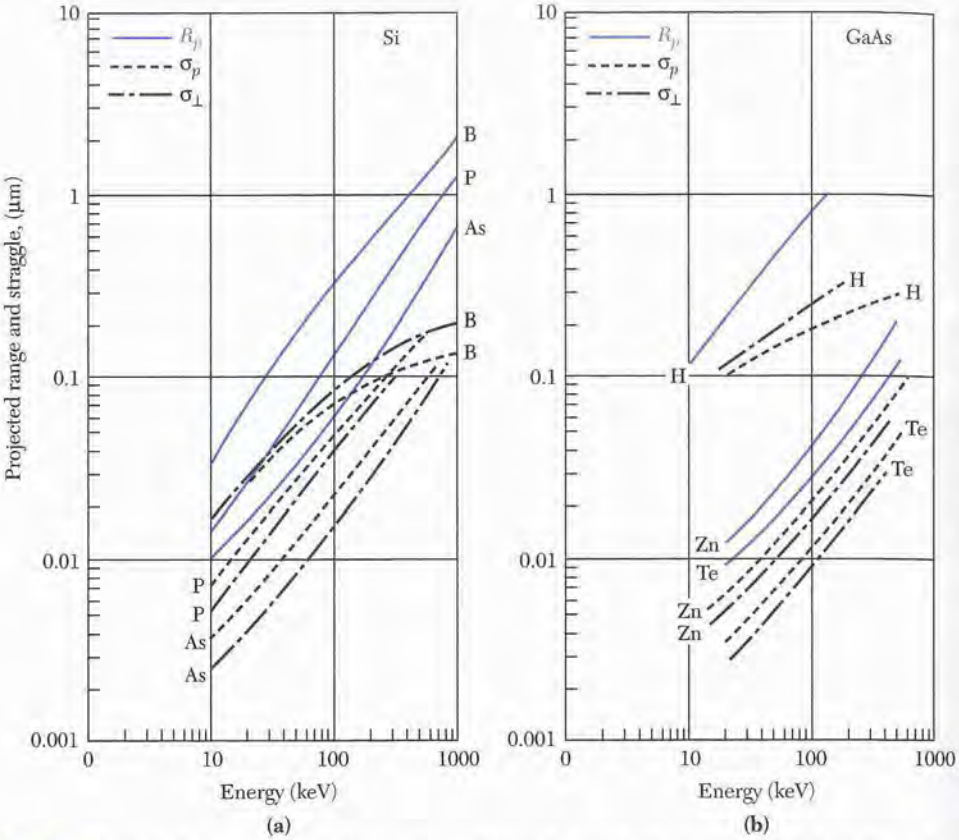


Fig. 17 Projected range, projected straggle, and lateral straggle for (*a*) B, P, and As in Si, and (*b*) H, Zn, and Te in GaAs.[17,19]

SOLUTION From Fig. 17*a*, we obtain 0.31 and 0.07 μm for the projected range and project straggle, respectively.

From Eq.25, $n(x) = \dfrac{S}{\sqrt{2\pi}\sigma_p} \exp\left[\dfrac{-(x-R_p)^2}{2\sigma_p^2}\right]$,

$$\frac{dn}{dx} = -\frac{S}{\sqrt{2\pi}\sigma_p}\frac{2(x-R_p)}{2\sigma_p^2}\exp\left[\frac{-(x-R_p)^2}{2\sigma_p^2}\right] = 0.$$

The peak concentration is at $x = R_p$, and $n(x) = 2.85 \times 10^{19}$ ions/cm^3.

The total number of implanted ions $= Q = 5\times10^{14}\times\pi\times\left(\dfrac{20}{2}\right)^2 = 1.57\times10^{17}$ ions.

The required ion current $= I = \dfrac{qQ}{t} = \dfrac{1.6\times10^{-19}\times1.57\times10^{17}}{60} = 4.19\times10^{-4}$ A.

$= 0.42$ mA.

13.4.3 Ion Channeling

The projected range and straggle of the Gaussian distribution discussed previously give a good description of the implanted ions in amorphous or fine-grain polycrystalline substrates. Both silicon and gallium arsenide behave as if they were amorphous semiconductors, provided the ion beam is misoriented from the low-index crystallographic direction (e.g.,<111>). In this situation, the doping profile described by Eq. 25 is followed closely near the peak and extended to one or two decades below the peak value. This is illustrated [16] in Fig. 18. However, even for a misorientation of 7° from the <111>-axis, there still is a tail that varies exponentially with distance as exp $(-x/\lambda)$, where λ is typically on the order of 0.1 μm.

The exponential tail is related to the ion-channeling effect. Channeling occurs when incident ions align with a major crystallographic direction and are guided between rows of atoms in a crystal. Figure 19 illustrates a diamond lattice viewed along a <110>-direction.[20] Ions implanted in the <110>-direction will follow trajectories that will not bring them close enough to a target atom to lose significant amounts of energy in nuclear collisions. Thus, for channeled ions, the only energy loss mechanism is electronic stopping, and the range of channeled ions can be significantly larger than it would be in an amorphous target. Ion channeling is particularly critical for low-energy implant and heavy ions.

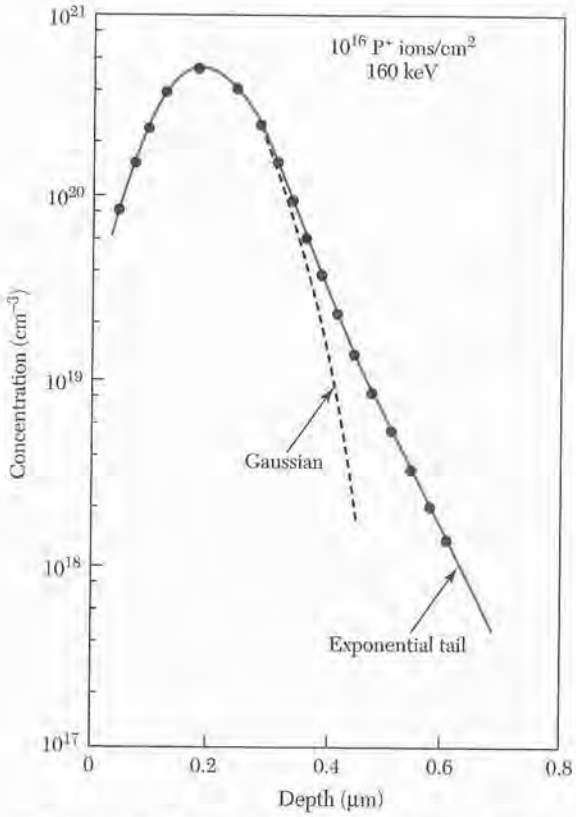


Fig. 18 Impurity profile obtained in a purposely misoriental target. Ion beam is incident 7° from the <111> -axis.[16]

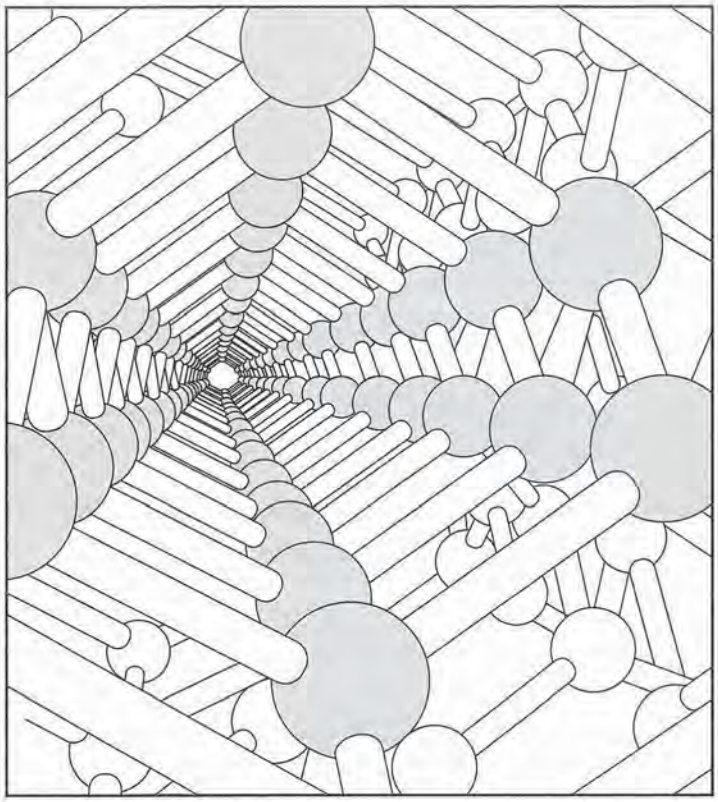


Fig. 19 Model for a diamond structure, viewed along a ⟨110⟩-axis.[20]

Channeling can be minimized by several techniques: a blocking amorphous surface layer, misorientation of the wafer, and creating a damage layer in the wafer surface. The usual blocking amorphous layer is simply a thin layer of grown silicon dioxide (Fig. 20*a*). The layer randomizes the direction of the ion beam so that the ions enter the wafer at different angles and not directly down the crystal channels. Misorientation of the wafers 5°–10 ° off the major plane also has the effect of preventing the ions from entering the channels (Fig. 20*b*). With this method, most implantation machines tilt the wafer by 7° and then apply a 22° twist from the flat to prevent channeling. Predamaging the wafer surface with a heavy silicon or germanium implant creates a randomizing layer in the wafer surface (Fig. 20*c*). This method, however, increases the use of the expensive ion implantor.

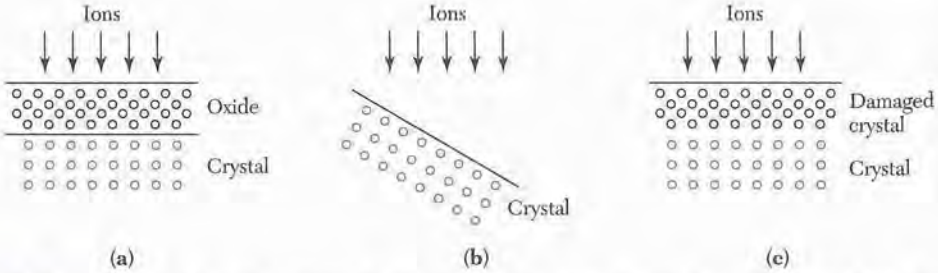


Fig. 20 (*a*) Implant through an amorphous oxide layer, (*b*) misorient the beam direction to all crystal axes, and (*c*) predamage on the crystal surface.

13.5 IMPLANT DAMAGE AND ANNEALING

13.5.1 Implant Damage

When energetic ions enter a semiconductor substrate, they lose their energy in a series of nuclear and electronic collisions and finally come to rest. The electronic-energy loss can be accounted for in terms of electronic excitations to higher energy levels or in the generation of electron-hole pairs. However, electronic collisions do not displace semiconductor atoms from their lattice positions. Only nuclear collisions can transfer sufficient energy to the lattice so that host atoms are displaced resulting in implant damage (also called lattice disorder).[21] These displaced atoms may possess large fractions of the incident energy, and they in turn cause cascades of secondary displacements of nearby atoms to form a *tree of disorder* along the ion path. When the displaced atoms per unit volume approach the atomic density of the semiconductor, the material becomes amorphous.

The tree of disorder for light ions is quite different from that for heavy ions. Much of the energy loss for light ions (e.g., $^{11}B^+$ in silicon) is due to electronic collisions (see Fig. 16), which do not cause lattice damage. The ions lose their energies as they penetrate deeper into the substrate. Eventually, the ion energy is reduced below the crossover energy (10 keV for boron) where nuclear stopping becomes dominant. Therefore, most of the lattice disorder occurs near the final ion position. This is illustrated in Fig. 21*a*.

We can estimate the damage by considering a 100 keV boron ion. Its projected range is 0.31 μm, (Fig.17*a*), and its initial nuclear energy loss is only 3 eV/Å (Fig. 16). Since the spacing between lattice planes in silicon is about 2.5 Å, this means that the boron ion will lose 7.5 eV for each lattice plane because of nuclear stopping. The energy required to displace a silicon atom from its lattice position is about 15 eV. Therefore, the incident boron ion does not release enough energy from nuclear stopping to displace a silicon atom when it first enters the silicon substrate. When the ion energy is reduced to about

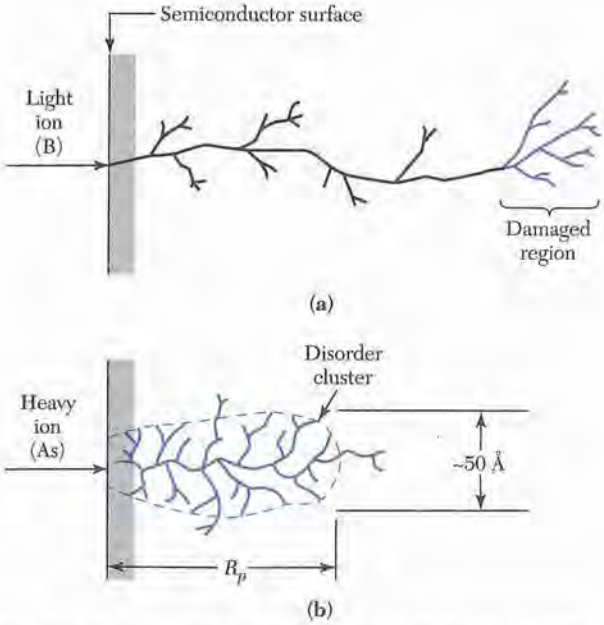


Fig. 21 Implantation disorder caused by (*a*) light ions and (*b*) heavy ions.[2,15]

50 keV (at a depth of 1500 Å), the energy loss due to nuclear stopping increases to 15 eV for each lattice plane (i.e., 6 eV/ Å), sufficient to create a lattice disorder. Assuming that one atom is displaced per lattice plane for the remaining ion range, we have 600 lattice atoms displaced (i.e., 1500 Å/2.5 Å). If each displaced atom moves roughly 25 Å from its original position, the damage volume is given by $V_D \cong \pi(25\text{Å})^2(1500\text{Å}) = 3 \times 10^{-18}\text{cm}^3$. The damage density is $600/V_D \cong 2 \times 10^{20}\ \text{cm}^{-3}$, which is only 0.4% of the atoms. Thus, very high doses of light ions are needed to create an amorphous layer.

For heavy ions, the energy loss is primarily due to nuclear collisions; therefore, we expect substantial damage. Consider a 100 keV arsenic ion with a projected range of 0.06 μm or 60 nm. The average nuclear energy loss over the entire energy range is about 1320 eV/nm (Fig. 16). This means that the arsenic ion loses about 330 eV for each lattice plane on the average. Most of the energy is given to one primary silicon atom. Each primary atom will subsequently cause 22 displaced target atoms (i.e., 330 eV/15eV). The total number of displaced atom is 5280. Assuming a range of 2.5 nm for the displaced atoms, the damage volume is $V_D \cong \pi(2.5\ \text{nm})^2\ (60\ \text{nm}) = 10^{-18}\ \text{cm}^3$. The damage density is then $5280/V_D \cong 5 \times 10^{21}\ \text{cm}^{-3}$, or about 10% of the total number of atoms in V_D. As a result of the heavy-ion implantation, the material has become essentially amorphous. Figure 21*b* illustrates the situation where the damage forms a disordered cluster over the entire projected range.

To estimate the dose required to convert a crystalline material to an amorphous form, we can use the criterion that the energy density is of the same order of magnitude as that needed for melting the material (i.e., 10^{21} keV/cm^3). For 100 keV arsenic ions, the dose required to make amorphous silicon is then

$$S = \frac{(10^{21}\ \text{keV}/\text{cm}^3)R_p}{E_0} = 6 \times 10^{13}\ \text{ions}/\text{cm}^2. \tag{32}$$

For 100 keV boron ions, the dose required is 3×10^{14} ions/cm^2 because R_p for boron is five times larger than for arsenic. However, in practice, higher doses ($>10^{16}$ ions/cm^2) are required for boron implant into a target at room temperature because of the nonuniform distribution of the damage along the ion path.

13.5.2 Annealing

Because of the damaged region and the disorder cluster that result from ion implantation, semiconductor parameters such as mobility and lifetime are severely degraded. In addition, most of the ions as implanted are not located in substitutional sites. To activate the implanted ion and to restore mobility and other material parameters, we must anneal the semiconductor at an appropriate combination of time and temperature.

In conventional annealing, we use an open-tube batch-furnace system similar to that for thermal oxidation. This process requires long time and high temperature to remove the implant damages. However, conventional annealing may cause substantial dopant diffusion and cannot meet the requirement for shallow junctions and narrow doping profiles. Rapid thermal annealing (RTA) is an annealing process that employs a variety of energy sources with a wide range of times, from 100 seconds down to nanoseconds—all short compared with the conventional annealing. RTA can activate the dopant fully with minimal redistribution.

Conventional Annealing of B and P

Annealing characteristics depend on the dopant type and the dose involved. Figure 22 shows the annealing behaviors of boron and phosphorus implantation into silicon substrates.[19] The substrate is held at room temperature (T_s) during implantation. At a given

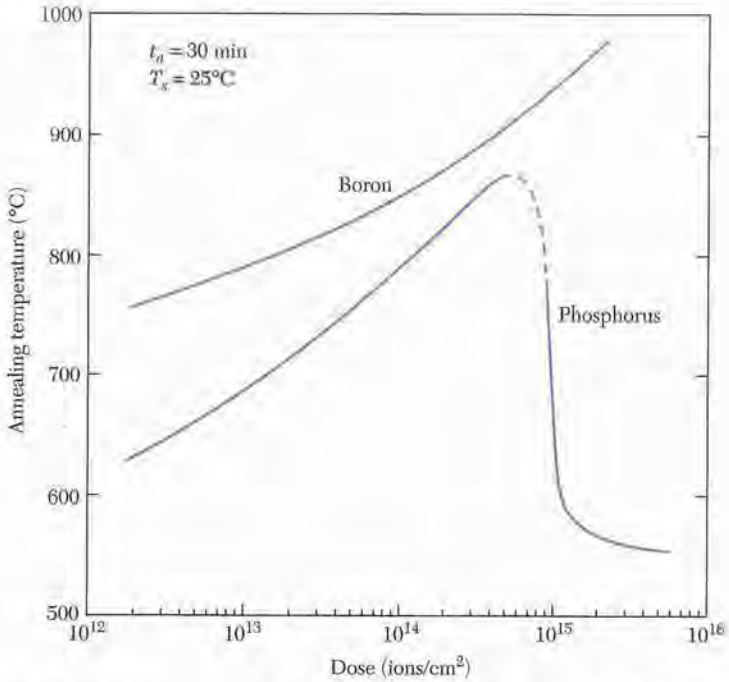


Fig. 22 Annealing temperature versus dose for 90% activation of boron and phosphorus.[1,19]

ion dose, the annealing temperature is defined as the temperature at which 90% of the implanted ions are activated by a 30 minute annealing in a conventional annealing furnace. For boron implantation, higher annealing temperatures are needed for higher doses. For phosphorus at lower doses, the annealing behavior is similar to that for boron. However, when the dose is greater than 10^{15} cm^{-2}, the annealing temperature drops to about 600°C. This phenomenon is related to the *solid-phase epitaxy* process. At phosphorus doses greater than 6×10^{14} cm^{-2}, the silicon surface layer becomes amorphous. The single-crystal semiconductor underneath the amorphous layer serves as a seeding area for recrystallization of the amorphous layer. The epitaxial-growth rate along the <100> direction is 10 nm/min at 550°C and 50 nm/min at 600°C, with an activation energy at 2.4 eV. Therefore, a 100–500 nm amorphous layer can be recrystallized in a few minutes. During the solid-phase epitaxial process, the impurity dopant atoms are incorporated into the lattice sites along with the host atoms; thus, full activation can be obtained at relatively low temperatures.

Rapid Thermal Annealing

The machine for RTA with a transient lamp heating is shown in Fig. 23. The temperature measured [22] from the heated wafer is usually from 600°–1100°C. A wafer is heated quickly under atmospheric conditions or at low pressure under isothermal conditions. Typical lamps in a RTA system are tungsten filaments or arc lamps. The processing chamber is made of either quartz, silicon carbide, stainless steel, or aluminum and has quartz windows through which the optical radiation passes to illuminate the wafer. The wafer holder is often made of quartz and contacts the wafer in a minimum number of places. A measurement system is placed in a control loop to set wafer temperature. The RTA system interfaces with a gas-handling system and a computer that controls system operation. Typically, wafer temperature in a RTA system is measured with a noncontact optical pyrometer that determines temperature from radiated infrared energy.

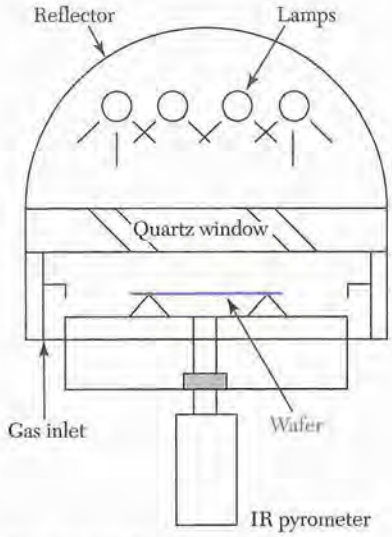


Fig. 23 Rapid thermal annealing system that is optically heated.

A comparison between conventional furnace and RTA technologies is shown in Table 2. To achieve short processing times using RTA, trade-offs must be made in temperature and process uniformity, temperature measurement and control, and wafer stress and throughput. In addition, there are concerns over the introduction of electrically active wafer defects during the very fast (100°–300°C/s) thermal transients. Rapid heating with temperature gradients in the wafers can cause wafer damage in the form of slip dislocations induced by thermal stress. On the other hand, conventional furnace processing brings with it significant problems, such as particle generation from the hot walls, limited ambient control in an open system, and a large thermal mass that restricts controlled heating times to tens of minutes. In fact, requirements on contamination, process control, and cost of manufacturing floor space have resulted in the paradigm shift to the RTA process.

TABLE 2 Technology Comparison

Determinant	Conventional furnace	Rapid thermal annealing
Process	Batch	Single-wafer
Furnace	Hot-wall	Cold-wall
Heating rate	Low	High
Cycle time	High	Low
Temperature monitor	Furnace	Wafer
Thermal budget	High	Low
Particle problem	Yes	Minimal
Uniformity and repeatability	High	Low
Throughput	High	Low

13.6 IMPLANTATION-RELATED PROCESSES

In this section we consider a few implantation-related processes, such as multiple implantation, masking, tilt-angle implantation, high-energy implantation, and high-current implantation.

13.6.1 Multiple Implantation and Masking

In many applications doping profiles other than the simple Gaussian distribution are required. One such case is the preimplantation of silicon with an inert ion to make the silicon surface region amorphous. This technique allows close control of the doping profile and permits nearly 100% dopant activation at low temperatures, as discussed previously. In such a case, a deep amorphous region may be required. To obtain this type of region, we must make a series of implants at varying ion energies and doses.

Multiple implantation can also be used to form a flat doping profile, as shown in Fig. 24. Here, four boron implants into silicon are used to provide a composite doping profile.[23] The measured carrier concentration and that predicted using range theory are shown in the figure. Other doping profiles, unavailable from diffusion techniques, can be obtained by using various combinations of impurity dose and implantation energy. Multiple implants have been used to preserve stoichiometry during the implantation and annealing of GaAs. This approach, whereby equal amount of gallium and n-type dopant (or arsenic and p-type dopant) are implanted prior to annealing, has resulted in higher carrier activation.

To form p–n junctions in selected areas of the semiconductor substrate, an appropriate mask should be used for the implantation. Because implantation is a low-temperature process, a large variety of masking materials can be used. The minimum thickness of masking material required to stop a given percentage of incident ions can be estimated from the range parameters for ions. The inset of Fig. 25 shows a profile of an implant in

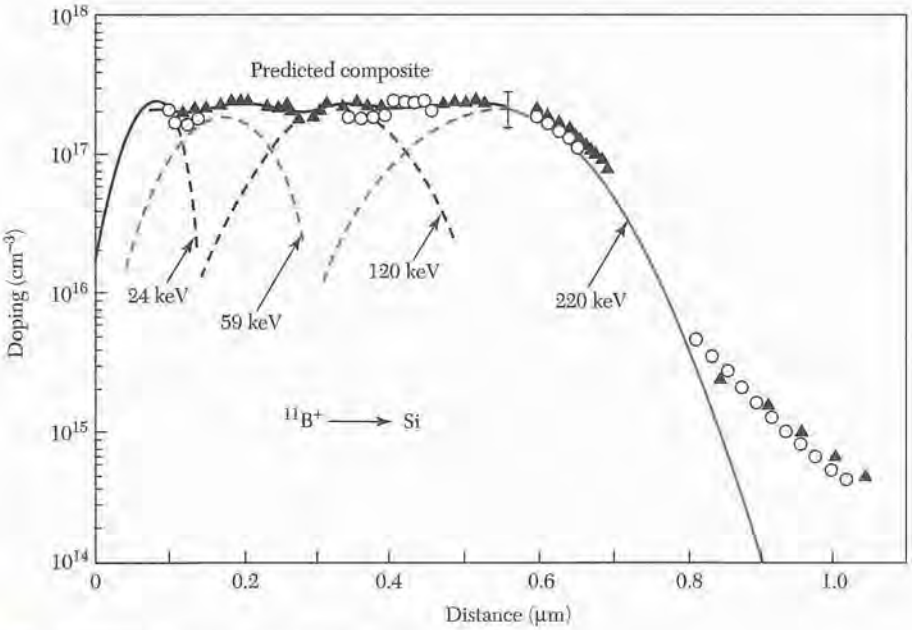


Fig. 24 Composite doping profile using multiple implants.[23]

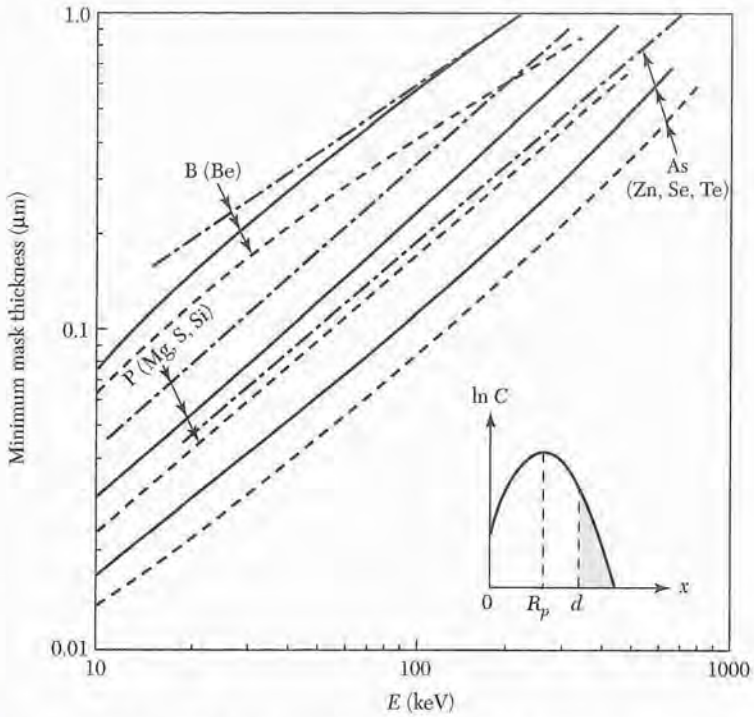


Fig. 25 Minimum thickness[24] of SiO_2 (—), Si_3N_4 (- - - -), and photoresist (– • – • –) to produce a masking effectiveness of 99.99%.

a masking material. The dose implanted in the region beyond a depth d (shown shaded) is given by integration of Eq. 25 as

$$S_d = \frac{S}{\sqrt{2\pi}\,\sigma_p}\int_d^\infty \exp\left[-\left(\frac{x-R_p}{\sqrt{2}\sigma_p}\right)^2\right]dx. \tag{33}$$

From Table 1 we can derive the expression

$$\int_x^\infty e^{-y^2}dy = \frac{\sqrt{\pi}}{2}\,\mathrm{erfc}\,(x). \tag{34}$$

Therefore, the fraction of the dose that has "transmitted" beyond a depth d is given by the transmission coefficient T:

$$\boxed{T \equiv \frac{S_d}{S} = \frac{1}{2}\,\mathrm{erfc}\left(\frac{d-R_p}{\sqrt{2}\,\sigma_p}\right).} \tag{35}$$

Once T is given, we can obtain the mask thickness d from Eq. 35 for any given R_p and σ_p.

The values of d to stop 99.99% of the incident ions ($T = 10^{-4}$) are shown in Fig. 25 for SiO_2, Si_3N_4, and a photoresist as masking materials.[19,24] Mask thicknesses given in this

figure are for boron, phosphorus, and arsenic implanted into silicon. These mask thicknesses also can be used as guidelines for impurity masking in gallium arsenide. The dopants are shown in the parentheses. Since both R_p and σ_p vary approximately linearly with energy, the minimum thickness of the masking material also increases linearly with energy. In certain applications, instead of totally stopping the beam, the masks can be used as attenuators, which can provide an amorphous surface layer to the incident ion beam to minimize the channeling effect.

EXAMPLE 4

When boron ions are implanted at 200 keV, what thickness of SiO_2 will be required to mask 99.996% of the implanted ions (R_p = 0.53 μm, σ_p = 0.093 μm)?

SOLUTION The complementary error function in Eq. 35 can be approximated if the argument is large (see Table 1):

$$T \cong \frac{1}{2\sqrt{\pi}} \frac{e^{-u^2}}{u},$$

where the parameter u is given by $(d - R_p)/\sqrt{2}\sigma$. For $T = 10^{-4}$, we can solve the above equation to give $u = 2.8$. Thus,

$$d = R_p + 3.96\sigma_p = 0.53 + 3.96 \times 0.093 = 0.898 \ \mu\text{m}$$

13.6.2 Tilt-Angle Ion Implantation

In scaling devices to submicron dimensions, it is important also to scale dopant profiles vertically. We need to produce junction depths less than 100 nm, including diffusion during dopant activation and subsequent processing steps. Modern device structures, such as the lightly doped drain (LDD), require precise control of dopant distributions vertically and laterally.

It is the ion velocity perpendicular to the surface that determines the projected range of an implanted ion distribution. If the wafer is tilted at a large angle to the ion beam, then the effective ion energy is greatly reduced. Figure 26 illustrates this for 60 keV arsenic ions as a function of the tilt angle, showing that it is possible to achieve extremely shallow distributions using a high tilt angle (86°). In tilt-angle ion implantation, we should consider the shadow effect (inset in Fig. 26) for the patterned wafer. A lower tilt angle leads to a small shadow area. For example, if the height of patterned mask is 0.5 μm, with vertical sidewall, a 7° incident ion beam will induce a 61 nm shadow. This shadow effect may introduce an unexpected series resistance in the device.

13.6.3 High-Energy and High-Current Implantation

High-energy implantors, capable of energies as high as 1.5–5 MeV, are available and have been used for a number of novel applications. The majority of these depend on the ability to dope the semiconductor to many micrometers in depth, without the need for long diffusion times at high temperatures. High-energy implantors can also be used to produce low-resistivity buried layers. For example, a buried layer 1.5–3 μm below the surface for a CMOS device can be achieved by high-energy implantation.

High-current implantors (10–20 mA), operating in the 25–30 keV range, are routinely used for the predeposition step in diffusion technology because the amount of total

dopant can be controlled precisely. After the predeposition, the dopant impurities can be driven in by a high-temperature diffusion step at the same time that implant damage at the surface region is annealed out. Another application is the threshold voltage adjustment in MOS devices. A precisely controlled amount of dopant (e.g., boron) is implanted through the gate oxide to the channel region[25] (Fig. 27*a*). Because the projected range of boron in silicon and silicon oxide are comparable, if we choose a suitable incident energy, the ions will penetrate just the thin gate oxide, not the thicker-field oxide. The threshold voltage will vary approximately linearly with the implanted dose. After boron implantation, polysilicon can be deposited and patterned to form the gate electrode of the MOSFET. The thin oxide surrounding the gate electrode is removed, and the source and drain regions are formed as shown in Fig. 27*b* by another high-dose arsenic implantation.

High-current implantors with energies in the 150–200 keV range are now available. A major use for these machines is to form high-quality silicon films, which are insulated

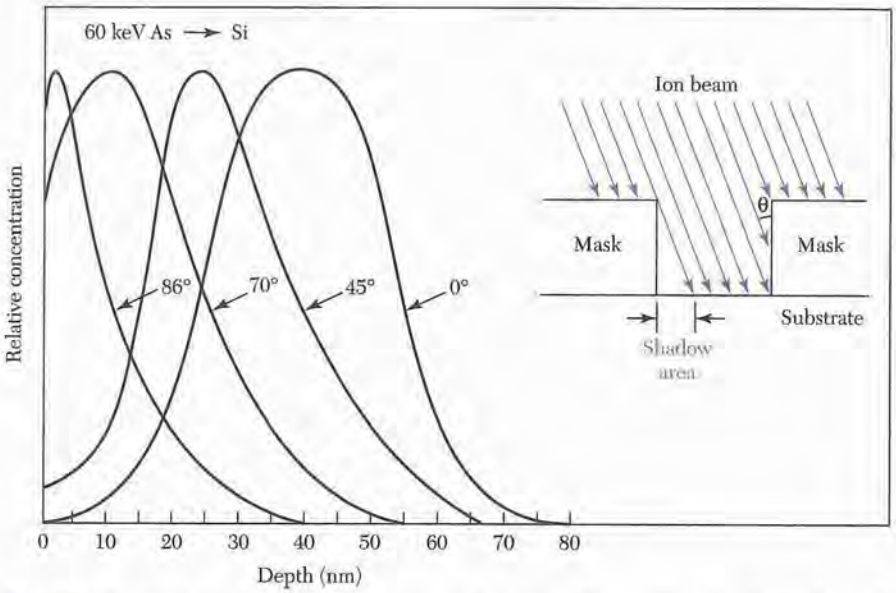


Fig. 26 60 keV arsenic implanted into silicon, as a function of beam tilt angle. Inset shows the shadow area for tilt-angle ion implantation.

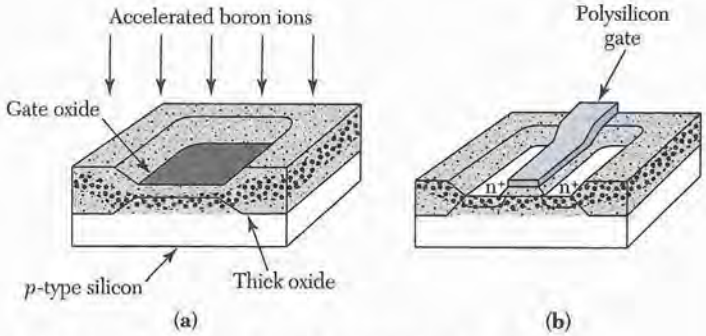


Fig. 27 Threshold voltage adjustment using boron ion implantation.[25]

from the substrate by implanting oxygen, creating an intervening layer of silicon dioxide. This separation by implantation of oxygen (SIMOX) is a key SOI (silicon-on-insulator) technology.

The SIMOX process uses a high-energy O^+ beam, typically in the 150 to 200 keV range, so that the oxygen ions have projected ranges of 100–200 nm. Additionally, a heavy dose, $1–2 \times 10^{18}$ ions/cm^2, is used to produce an insulating layer of SiO_2 that is 100–500 nm thick. The use of SIMOX material leads to a significant reduction of source/drain capacitances in MOS devices. Moreover, it reduces coupling between devices and, thus, allows tighter packing without the problem of latchup. As a result, it is widely proposed as the material of choice for advanced, high-speed CMOS circuits.

SUMMARY

Diffusion and ion implantation are the two key methods for impurity doping. We first considered the basic diffusion equation for constant diffusivity. We obtain the complementary error function (erfc) and the Gaussian function for the constant–surface-concentration case and constant–total-dopant case, respectively. The results of a diffusion process can be evaluated by measurements of the junction depth, the sheet resistance, and the dopant profile.

When the doping concentration is higher than the intrinsic carrier concentration n_i at the diffusion temperature, the diffusivity becomes concentration dependent. This dependence has profound effect on the resulting doping profile. For example, arsenic and boron diffusivity in silicon vary linearly with the impurity concentration. Their doping profiles are much more abrupt than the erfc profile. Phosphorus diffusivity in silicon varies as the square of concentration. This dependence and a dissociation effect give rise to a phosphorus diffusivity that is 100 times larger than its intrinsic diffusivity.

Lateral diffusion at the edge of a mask and impurity redistribution during oxidation are two processes in which diffusion can have an important impact on device performance. The former can substrantially reduce the breakdown voltage, and the latter will influence the threshold voltage as well as the contact resistance.

The key parameters for ion implantation are the projected range R_p and its standard deviation σ_p also called the projected straggle. The implantation profile can be approximated by a Gaussian distribution with its peak located at R_p from the surface of the semiconductor substrate. The advantages of ion implantation process are more precise control of the amount of dopant, a more reproducible doping profile, and lower processing temperature compared with the diffusion process.

We considered R_p and σ_p for various elements in silicon and gallium arsenide and discussed the channeling effect and ways to minimize this effect. However, implantation may cause severe damage to the crystal lattice. To remove the implant damage and to restore mobility and other device parameters, we must anneal the semiconductor at an appropriate combination of time and temperature. Currently, the rapid thermal annealing (RTA) is preferred to conventional furnace annealing because RTA can remove implant damage without thermal broadening of the doping profile.

Ion implantation has wide applications for advanced semiconductor devices. These include (a) multiple implantation to form novel distributions, (b) selection of masking materials and thickness to stop a given percentage of incident ions from reaching the substrate, (c) tilt-angle implantation to form ultrashallow junctions, (d) high-energy implantation to form buried layers, and (e) high-current implantation for predeposition and threshold voltage adjustment and to form an insulating layer for SOI applications.

REFERENCES

1. S. M. Sze, Ed., *VLSI Technology*, 2nd Ed., McGraw-Hill, New York, 1988, Ch. 7, 8.
2. S. K. Ghandhi, *VLSI Fabrication Principles*, 2nd Ed., Wiley, New York, 1994, Ch. 4, 6.
3. W. R. Runyan and K. E. Bean, *Semiconductor Integrated Circuit Processing Technology*, Addison-Wesley, Massachusetts, 1990, Ch. 8.
4. H. C. Casey, and G. L. Pearson, "Diffusion in Semiconductors," in J. H. Crawford, and L. M. Slifkin, Eds., *Point Defects in Solids*, Vol. 2, Plenum, New York, 1975.
5. J. P. Joly, "Metallic Contamination of Silicon Wafers," *Microelectron. Eng.*, **40**, 285 (1998).
6. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967.
7. ASTM Method F374-88, "Test Method for Sheet Resistance of Silicon Epitaxial, Diffused, and Ion-implanted Layers Using a Collinear Four-Probe Array," **V10**, 249 (1993).
8. J. C. Irvin, "Evaluation of Diffused Layers in Silicon," *Bell Syst. Tech. J.*, **41**, 2 (1962).
9. ASTM Method E1438-91, "Standard Guide for Measuring Width of Interfaces in Sputter Depth Profiling Using SIMS," **V10**, 578 (1993).
10. R. B. Fair, "Concentration Profiles of Diffused Dopants," in F. F. Y. Wang, Ed., *Impurity Doping Processes in Silicon*, North-Holland, Amsterdam, 1981.
11. L. R. Weisberg and J. Blanc, "Diffusion with Interstitial-Substitutional Equilibrium, Zinc in GaAs," *Phys. Rev.*, **131**, 1548 (1963).
12. A. F. W. Willoughby, "Double-Diffusion Processes in Silicon," in F. F. Y. Wang, Ed., *Impurity Doping Processes in Silicon*, North-Holland, Amsterdam, 1981.
13. F. A. Cunnell and C. H. Gooch, "Diffusion of Zinc in Gallium Arsenide", *J. Phys. Chem. Solid*, **15**, 127 (1960).
14. D. P. Kennedy and R. R. O'Brien, "Analysis of the Impurity Atom Distribution Near the Diffusion Mask for a Planar *p-n* Junction, " *IBM J. Res. Dev.*, **9**, 179 (1965).
15. I. Brodie and J. J. Muray, *The Physics of Microfabrication*, Plenum, New York, 1982.
16. J. F. Gibbons, "Ion Implantation," in S. P. Keller, Ed., *Handbook on Semiconductors*, Vol. 3, North-Holland , Amsterdam, 1980.
17. S. Furukawa, H. Matsumura, and H. Ishiwara, "Theoretical Consideration on Lateral Spread of Implanted Ions," *Jpn. J. Appl. Phys.*, **11**, 134 (1972).
18. B. Smith, *Ion Implantation Range Data for Silicon and Germanium Device Technologies*, Research Studies, Forest Grove, OR., 1977.
19. K. A. Pickar, "Ion Implantation in Silicon," in R. Wolfe, Ed., *Applied Solid State Science*, Vol. 5, Academic, New York, 1975.
20. L. Pauling and R. Hayward, *The Architecture of Molecules*, Freeman, San Francisco, 1964.
21. D. K. Brice, "Recoil Contribution to Ion Implantation Energy Deposition Distribution," *J. Appl. Phys.*, **46**, 3385 (1975).
22. C. Y. Chang and S. M. Sze, Eds., *ULSI Technology*, McGraw-Hill, New York, 1996, Ch. 4.
23. D. H. Lee and J. W. Mayer, "Ion-Implanted Semiconductor Devices," *Proc. IEEE*, **62**, 1241 (1974).
24. G. Dearnaley, et al., *Ion Implantation*, North-Holland, Amsterdam, 1973.
25. W. G. Oldham, "The Fabrication of Microelectronic Circuit," in *Microelectronics*, Freeman, San Francisco, 1977.

PROBLEMS (* DENOTES DIFFICULT PROBLEMS)

FOR SECTION 13.1 BASIC DIFFUSION PROCESS

1. Calculate the junction depth and the total amount of dopant introduced after boron pre-deposition performed at 950°C for 30 minutes in a neutral ambient. Assume the substrate is *n*-type silicon with $N_D = 1.8 \times 10^{16}\ \text{cm}^{-3}$ and the boron surface concentration is $C_s = 1.8 \times 10^{20}\ \text{cm}^{-3}$.

2. If the sample in Prob. 1 is subjected to a neutral drive-in at 1050°C for 60 minutes, calculate the diffusion profile and the junction depth.

3. Assume the measured phosphorus profile can be represented by a Gaussian function with a diffusivity $D = 2.3 \times 10^{-13}$ cm²/s. The measured surface concentration is 1×10^{18} atoms/cm³, and the measured junction depth is 1 μm at a substrate concentration of 1×10^{15}. Calculate the diffusion time and the total dopant in the diffused layer.

*4. To avoid wafer warp due to a sudden reduction in temperature, the temperature in a diffusion furnace is decreased linearly from 1000°C to 500°C in 20 minutes. What is the effective diffusion time at the initial diffusion temperature for a phosphorus diffusion in silicon?

*5. For a low-concentration phosphorus drive-in diffusion in silicon at 1000°C, find the percentage change of surface concentration for 1% variation in diffusion time and temperature.

6. If arsenic is diffused into a thick slice of silicon doped with 10^{15} boron atoms/cm³ at a temperature of 1100°C for 3 hours, what is the final distribution of arsenic if the surface concentration is held fixed at 4×10^{18} atoms/cm³? What are the diffusion length and junction depth?

FOR SECTION 13.2 EXTRINSIC DIFFUSION

7. If arsenic is diffused into a thick slice of silicon doped with 10^{15} boron atoms/cm³ at a temperature of 900°C for 3 hours, what is the final distribution of arsenic if the surface concentration is held fixed at 4×10^{18} atoms/cm³? What is the junction depth? Assume

$$D = D_0 e^{\frac{-E_a}{kT}} \times \frac{n}{n_i},\ D_0 = 45.8\ \text{cm}^2/\text{s},\ E_a = 4.05\ \text{eV},\ x_j = 1.6\sqrt{Dt}.$$

8. Explain the meaning of intrinsic and extrinsic diffusion.

FOR SECTION 13.3 DIFFUSION RELATED PROCESSES

9. Define the segregation coefficient.

10. Assume that the Cu concentration in SiO_2 layer is 5×10^{13} atoms/cm³ after vapor phase decomposition and is measured with atomic absorption spectometry. The Cu concentration in the Si layer is 3×10^{11} atoms/cm³ afer HF/H_2O_2 dissolution. Calculate the segregation coefficient of Cu in SiO_2/Si layers.

FOR SECTION 13.4 RANGE OF IMPLANTED IONS

11. In a 200 mm wafer boron-ion–implantation system, assume the beam current is 10 μA. For the *p*-channel transistor, calculate the implant time required to reduce threshold voltage from –1.1 V to – 0.5 V. Assume that the implanted acceptors form a sheet of negative charge just below the Si surface, and the oxide thickness is 10 nm.

12. Assume that a 100 mm diameter GaAs wafer is uniformly implanted with 100 keV zinc ions for 5 minutes with a constant ion beam current of 10 μA. What are the ion dose per unit area and the peak ion concentration?

13. A silicon *p-n* junction is formed by implanting boron ions at 80 keV through a window in an oxide. If the boron dose is 2×10^{15} cm⁻² and the *n*-type substrate concentration is 10^{15} cm⁻³, find the location of the metallurgical junction.

14. A threshold-voltage adjustment implantation is made through a 25 nm gate oxide. The substrate is a <100>-oriented *p*-type silicon with a resistivity of 10 Ω-cm. If the incremental threshold voltage due to a 40 keV boron implantation is 1 V, what is the total implanted dose per unit area? Estimate the location of the peak boron concentration.

*15. For the substrate in Prob. 14, what percentage of the total dose is in the silicon?

FOR SECTION 13.5 IMPLANT DAMAGE AND ANNEALING

16. If a 50 keV boron ion is implanted into the silicon substrate, calculate the damage density. Assume silicon atom density is 5.02×10^{22} atoms/cm^3, the silicon displacement energy is 15 eV, the range is 2.5 nm, and the spacing between silicon lattice plane is 0.25 nm.

17. Explain why high-temperature RTA is preferable to low-temperature RTA for defect-free shallow-junction formation.

18. Estimate the implant dose required to reduce a p-channel threshold voltage by 1 V if the gate oxide is 4 nm thick. Assume that the implant voltage is adjusted so that the peak of the distribution occurs at the oxide-silicon interface. Thus, half of the implant goes into the silicon. Further, assume that 90% of the implanted ions in the silicon are electrically activated by the annealing process. These assumptions allow 45% of the implanted ions to be used for threshold adjusting. Also assume that all of the charge in the silicon is effectively at the silicon-oxide interface.

FOR SECTION 13.6 IMPLANTATION-RELATED PROCESSES

19. We would like to form 0.1 μm deep, heavily doped junctions for the source and drain regions of a submicron MOSFET. Compare the options that are available to introduce and activate dopant for this application. Which option would you recommend and why?

20. When an arsenic implant at 100 keV is used and the photoresist thickness is 400 nm, find the effectiveness of the resist mask in preventing the transmission of ions (R_p = 0.6 μm, σ_p = 0.2 μm). If the resist thickness is changed to 1 μm, calculate the masking efficiency.

21. With reference to Ex. 4, what thickness of SiO_2 is required to mask 99.999% of the implanted ions?

CHAPTER 14

Integrated Devices

- 14.1 PASSIVE COMPONENTS
- 14.2 BIPOLAR TECHNOLOGY
- 14.3 MOSFET TECHNOLOGY
- 14.4 MESFET TECHNOLOGY
- 14.5 CHALLENGES FOR MICROELECTRONICS
- SUMMARY

Microwave, photonic, and power applications generally employ discrete devices. For example, an IMPATT diode is used as a microwave generator, an injection laser as an optical source, and a thyristor as a high-power switch. However, most electronic systems are built on the integrated circuit (IC), which is an ensemble of both active (e.g., transistor) and passive devices (e.g., resistor, capacitor, and inductor) formed on and within a single-crystal semiconductor substrate and interconnected by a metallization pattern.[1] ICs have enormous advantages over discrete devices connected by wire bondings. The advantages includes (a) reduction of the interconnection parasitics, because an IC with multilevel metallization can substantially reduce the overall wiring length, (b) full utilization of semiconductor wafer's "real estate," because devices can be closely packed within an IC chip, and (c) drastic reduction in processing cost, because wire bonding is a time-consuming and error-prone operation.

In this chapter we combine the basic processes described in previous chapters to fabricate active and passive components in an IC. Because the key element of an IC is the transistor, specific processing sequences are developed to optimize its performance. We consider three major IC technologies associated with the three transistor families: the bipolar transistor, the MOSFET, and the MESFET.

Specifically, we cover the following topics:

- The design and fabrication of IC resistor, capacitor, and inductor.
- The processing sequence for standard bipolar transistor and advanced bipolar devices.
- The processing sequence for MOSFET with special emphasis on CMOS and memory devices.
- The processing sequence for high-performance MESFET and monolithic microwave IC.
- The major challenges for future microelectronics, including ultrashallow junction, ultrathin oxide, new interconnection materials, low power dissipation, and isolation.

Figure 1 illustrates the interrelationship between the major process steps used for IC fabrication. Polished wafers with a specific resistivity and orientation are used as the starting material. The film formation steps include thermally grown oxide films, deposited polysilicon, dielectric, and metal films (Chapter 11). Film formation is often followed by lithography (Chapter 12) or impurity doping (Chapter 13). Lithography is generally followed by etching, which in turn is often followed by another impurity doping or film formation. The final IC is made by sequentially transferring the patterns from each mask, level by level, onto the surface of the semiconductor wafer.

After processing, each wafer contains hundreds of identical rectangular chips (or dice), typically between 1 and 20 mm on each side, as shown in Fig. 2*a*. The chips are separated by sawing or laser cutting; Figure 2*b* shows a separated chip. Schematic top views of a single MOSFET and a single bipolar transistor are shown in Fig. 2*c* to give some perspective of the relative size of a component in an IC chip. Prior to chip separation, each chip is electrically tested. Defective chips are usually marked with a dab of black ink. Good chips are selected and packaged to provide an appropriate thermal, electrical, and interconnection environment for electronic applications.[2]

IC chips may contain from a few components (transistors, diodes, resistors, capacitors, etc.) to as many as a billion or more. Since the invention of the monolithic IC in 1959, the number of components on a state-of-the-art IC chip has grown exponentially. We usually refer to the complexity of an IC as small-scale integration (SSI) for up to 100 components per chip, medium-scale integration (MSI) for up to 1000 components per chip, large-scale integration (LSI) for up to 100,000 components per chip, very-large-scale integrated (VLSI) for up to 10^7 components per chip, and ultra large-scale integration (ULSI) for larger numbers of components per chip. In Section 14.3, we show two ULSI chips, a 32-bit microprocessor chip, which contains over 42 million components, and a 1 Gbit dynamic random access memory (DRAM) chip, which contains over 2 billion components.

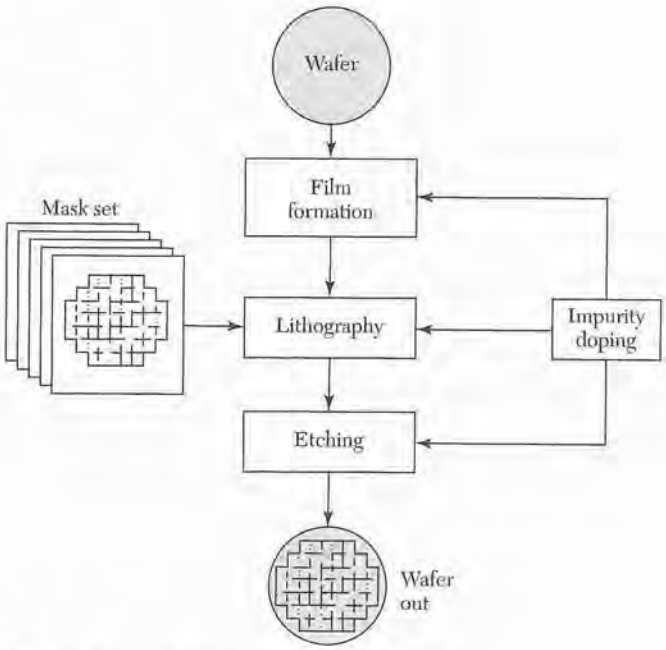


Fig. 1 Schematic flow diagram of integrated-circuit fabrication.

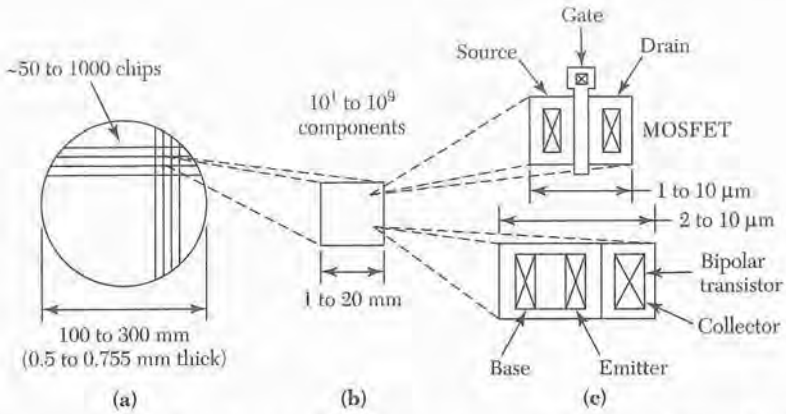


Fig. 2 Size comparison of a wafer to individual components. (*a*) Semiconductor wafer. (*b*) Chip. (*c*) MOSFET and bipolar transistor.

14.1 PASSIVE COMPONENTS

14.1.1 The Integrated-Circuit Resistor

To form an IC resistor, we can deposite a resistive layer on a silicon substrate, then pattern the layer by lithography and etching. We can also define a window in a silicon dioxide layer grown thermally on a silicon substrate and then implant (or diffuse) impurities of the opposite conductivity type into the wafer. Figure 3 shows the top and cross-sectional views of two resistors formed by the latter approach: one has a meander shape and the other has a bar shape.

Consider the bar-shaped resistor first. The differential conductance dG of a thin layer of the p-type material that is of thickness dx parallel to the surface and at a depth x (as shown by the B-B cross section) is

$$dG = q\mu_p p(x)\frac{W}{L}dx, \tag{1}$$

where W is the width of the bar, L is the length of the bar (we neglect the end contact areas for the time being), μ_p is mobility of hole, and $p(x)$ is the doping concentration. The total conductance of the entire implanted region of the bar is given by

$$\boxed{G = \int_0^{x_j} dG = q\frac{W}{L}\int_0^{x_j} \mu_p p(x)dx,} \tag{2}$$

where x_j is the junction depth. If the value of μ_p , which is a function of the hole concentration, and the distribution of $p(x)$ are known, the total conductance can be evaluated from Eq. 2. We can write

$$G \equiv g\frac{W}{L}, \tag{3}$$

where $g \equiv q\int_0^{x_j} \mu_p p(x)dx$ is the conductance of a square resistor pattern, that is, $G = g$ when $L = W$.

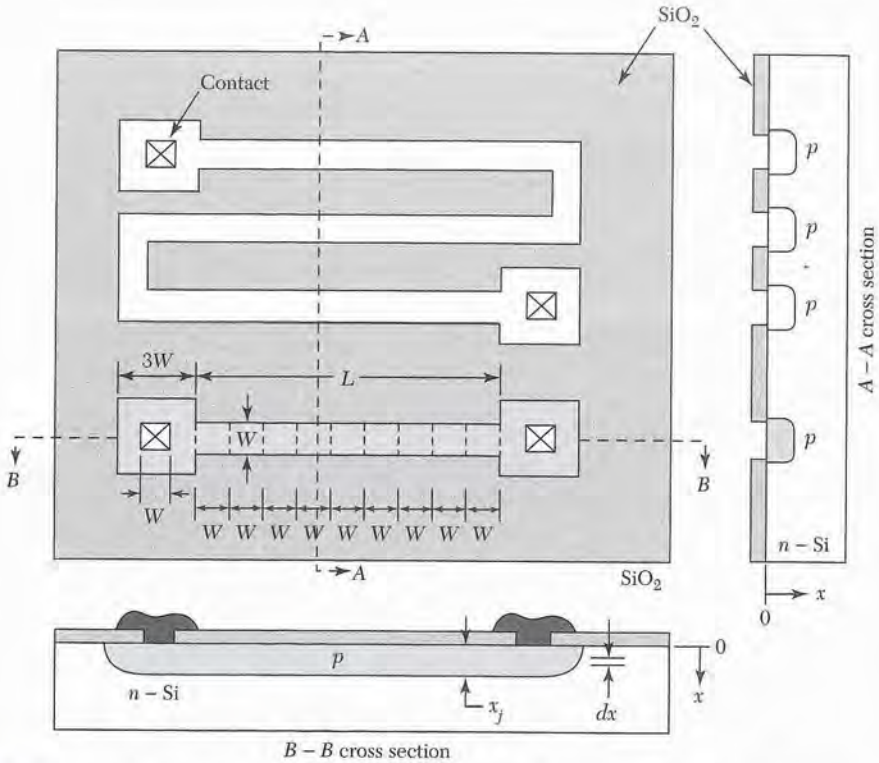


Fig. 3 Integrated-circuit resistors. All narrow lines in the large square area have the same width W, and all contacts are the same size.

The resistance is therefore given by

$$\boxed{R \equiv \frac{1}{G} = \frac{L}{W}\left(\frac{1}{g}\right),} \tag{4}$$

where $1/g$ usually is defined by the symbol $R_\square$ and is called the sheet resistance. The sheet resistance has units of ohms but is conventionally specified in units of ohms per square ($\Omega/\square$).

Many resistors in an integrated circuit are fabricated simultaneously by defining different geometric patterns in the mask such as those shown in Fig. 3. Since the same processing cycle is used for all these resistors, it is convenient to separate the resistance into two parts: the sheet resistance $R_\square$, determined by the implantation (or diffusion) process; and the ratio L/W, determined by the pattern dimensions. Once the value of $R_\square$ is known, the resistance is given by the ratio L/W, or the number of squares (each square has an area of $W \times W$) in the resistor pattern. The end contact areas will introduce additional resistance to the IC resistors. For the type shown in Fig. 3, each end contact corresponds to approximately 0.65 square. For the meander-shape resistor, the electric-field lines at the bends are not spaced uniformly across the width of the resistor but are crowded toward the inside corner. A square at the bend does not contribute exactly 1 square, but rather 0.65 square.

EXAMPLE 1

Find the value of a resistor 90 μm long and 10 μm wide, such as the bar-shaped resistor in Fig. 3. The sheet resistance is 1 kΩ /□.

SOLUTION The resistor contains 9 squares. The two end contacts correspond to 1.3 □. The value of the resistor is $(9 + 1.3) \times 1\ \text{k}\Omega/\square = 10.3\ \text{k}\Omega$.

14.1.2 The Integrated-Circuit Capacitor

There are basically two types of capacitors used in integrated circuits: MOS capacitors and p–n junctions. The MOS (metal-oxide-semiconductor) capacitor can be fabricated by using a heavily doped region (such as an emitter region) as one plate, the top metal electrode as the other plate, and the intervening oxide layer as the dielectric. The top and cross-sectional views of a MOS capacitor are shown in Fig. 4*a*. To form a MOS capacitor, a thick oxide layer is thermally grown on a silicon substrate. Next, a window is lithographically defined and then etched in the oxide. Diffusion or ion implantation is used to form a p^+-region in the window area, whereas the surrounding thick oxide serves as a mask. A thin oxide layer is then thermally grown in the window area, followed by a metallization step. The capacitance per unit area is given by

$$\boxed{C = \frac{\varepsilon_{ox}}{d}\ \text{F/cm}^2,} \tag{5}$$

where ε_{ox} is the dielectric permittivity of silicon dioxide (the dielectric constant $\varepsilon_{ox}/\varepsilon_0$ is 3.9) and d is the thin-oxide thickness. To increase the capacitance further, insulators with higher dielectric constants are being studied, such as Si_3N_4, and Ta_2O_5, with dielectric constants of 7 and 25, respectively. The MOS capacitance is essentially independent of the applied voltage, because the lower plate of the capacitor is made of heavily doped material. This also reduces the series resistance associated with it.

A p–n junction is sometimes used as a capacitor in an integrated circuit. The top and cross sectional views of an n^+-p junction capacitor are shown in Fig. 4*b*. The detailed

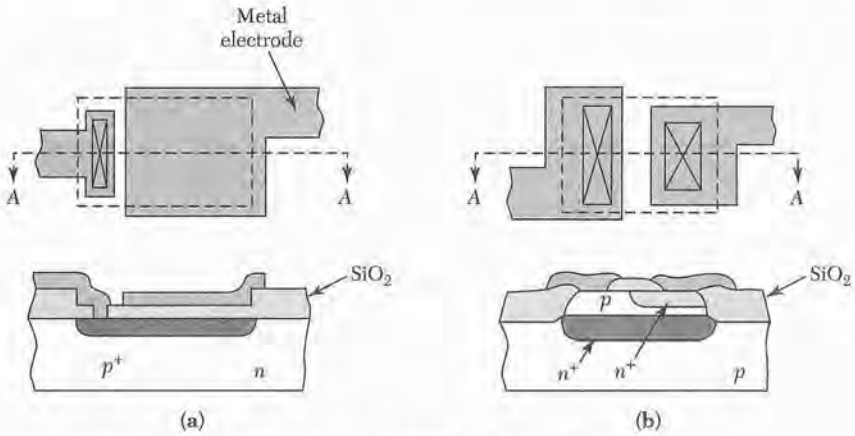


Fig. 4 (*a*) Integrated MOS capacitor. (*b*) Integrated p–n junction capacitor.

fabrication process is considered in Section 14.2, because this structure forms part of a bipolar transistor. As a capacitor, the device is usually reverse biased, that is, the p-region is reverse-biased with respect to the n^+-region. The capacitance is not a constant but varies as $(V_R + V_{bi})^{-1/2}$, where V_R is the applied reverse voltage and V_{bi} is the built-in potential. The series resistance is considerably higher than that of a MOS capacitor because the p-region has higher resistivity than does the p^+-region.

EXAMPLE 2

What is the stored charge and the number of electrons on an MOS capacitor with an area of 4 μm^2, for (a) a dielectric of 10 nm thick SiO_2 and (b) a 5 nm thick Ta_2O_5. The applied voltage is 5 V for both cases.

SOLUTION

(a) $$Q = \varepsilon_{ox} \times A \times \frac{V_s}{d} = 3.9 \times 8.85 \times 10^{-14}\ \text{F/cm} \times 4 \times 10^{-8}\,\text{cm}^2 \times \frac{5\text{V}}{1 \times 10^{-6}\,\text{cm}} = 6.9 \times 10^{-14}\text{C}$$

or

$$Q_s = 6.9 \times 10^{-14}\ \text{C}/q = 4.3 \times 10^5 \text{ electrons.}$$

(b) Changing the dielectric constant from 3.9 to 25 and the thickness from 10 nm to 5 nm, we obtain $Q_s = 8.85 \times 10^{-13}$ C, and $Q_s = 8.85 \times 10^{-13}\ \text{C}/q = 5.53 \times 10^6$ electrons. ◂

14.1.3 The Integrated-Circuit Inductor

IC inductors have been widely used in III-V based monolithic microwave integrated circuits (MMIC)[3]. With the increased speed of silicon devices and advancement in multilevel interconnection technology, IC inductors have started to receive more and more attentions in silicon-based radio frequency (rf) and high-frequency applications. Many kinds of inductors can be fabricated using IC processes. The most popular method is the thin-film spiral inductor. Figure 5*a* and *b* shows the top-view and the cross section of a silicon-based, two-level–metal spiral inductor. To form a spiral inductor, a thick oxide is thermally grown or deposited on a silicon substrate. The first metal is then deposited and defined as one end of the inductor. Next, another dielectric is deposited onto the metal 1. A via hole is defined lithographically and etched in the oxide. Metal 2 is deposited and the via hole is filled. The spiral patterned can be defined and etched on the metal 2 as the second end of the inductor.

To evaluate the inductor, an important figure of merit is the quality factor, Q. The Q is defined as $Q = L\omega/R$, where L, R, and ω are the inductance, resistance, and frequency, respectively. The higher the Q values, the lower the loss from resistance, hence the better the performance of the circuits. Figure 5*c* shows the equivalent circuit model. R_1 is the inherent resistivity of the metal, C_{p1} and C_{p2} are the coupling capacitances between the metal lines and the substrate, and R_{sub1} and R_{sub2} are the resistances of the silicon substrate associated with the metal lines, respectively. The Q increases linearly with frequency initially and then drops at higher frequencies because of parasitic resistances and capacitances.

There are some approaches to improve the Q value. The first is to use low-dielectric-constant materials (<3.9) to reduce the C_p. The other is to use a thick film metal or low-resistivity metals (e.g., Cu, Au to replace Al) to reduce the R_1. The third approach

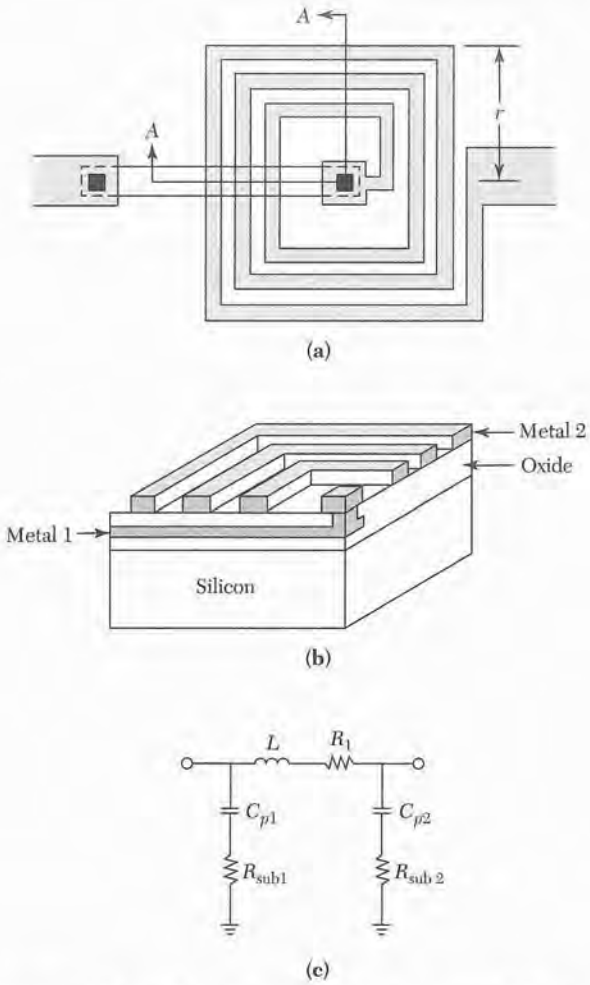


Fig. 5 (*a*) Schematic view of a spiral inductor on a silicon substrate. (*b*) Perspective view along *A*-*A*′. (*c*) An equivalent circuit model for an integrated inductor.

uses an insulating substrate (e.g., silicon-on-sapphire, silicon-on-glass, or quartz) to reduce R_{sub} *loss*.

To obtain the exact value of a thin-film inductor, complicated simulation tool, such as computer aided design, must be employed for both circuit simulation and inductor optimization. The model for thin-film inductor must take into account the resistance of the metal, the capacitance of the oxide, line-to-line capacitance, the resistance of the substrate, the capacitance to the substrate, and the inductance and mutual inductance of the metal lines. Hence, it is more difficult to calculate the integrated inductance compared with the integrated capacitors or resistors. However, a simple equation to estimate the square planar spiral inductor is given as[3]

$$\boxed{L \approx \mu_0 n^2 r \approx 1.2 \times 10^{-6} n^2 r,} \tag{6}$$

where μ_0 is the permeability in vacuum ($4\pi \times 10^{-7}$ H/m), L is in henries, n is the number of turns, and r is the radius of the spiral in meters.

EXAMPLE 3

For an integrated inductor with an inductance of 10 nH, what is the required radius if the number of turns is 20?

SOLUTION According to the Eq. 6,

$$r = \frac{10 \times 10^{-9}}{1.2 \times 10^{-6} \times 20^2} = 2.08 \times 10^{-5}\,(\text{m}) = 20.8\ \mu\text{m}.$$

14.2 BIPOLAR TECHNOLOGY

For IC applications, especially for VLSI and ULSI, the size of bipolar transistors must be reduced to meet the high-density requirement. Figure 6 illustrates the reduction in the size of the bipolar transistor in recent years.[4] The main differences in a bipolar transistor in an IC compared with a discrete transistor are that all electrode contacts are located on the *top surface* of the IC wafer, and each transistor must be electrically *isolated* to prevent interactions between devices. Prior to 1970, both the lateral and vertical isolations were provided by p–n junctions (Fig. 6*a*) and the lateral p-isolation region was always reverse biased with respect to the n-type collector. In 1971, thermal oxide was used for lateral isolation, resulting in a substantial reduction in device size (Fig. 6*b*), because the base and collector contacts abut the isolation region. In the mid-1970s, the emitter extended to the walls of the oxide, resulting in an additional reduction in area (Fig. 6*c*). At the present time, all the lateral and vertical dimensions have been scaled down and emitter stripe widths have dimensions in the submicron region (Fig. 6*d*).

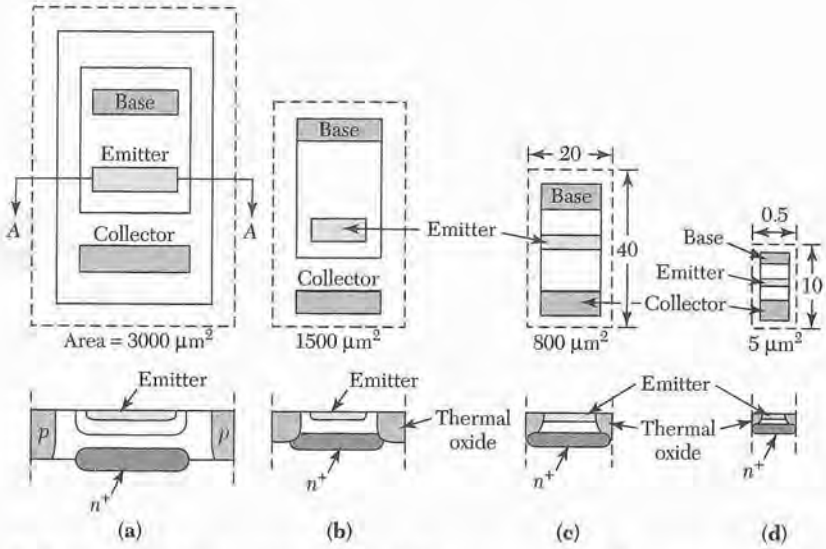


Fig. 6 Reduction of the horizontal and vertical dimensions of a bipolar transistor. (a) Junction isolation. (*b*) Oxide isolation. (*c* and *d*) Scaled oxide isolation.[4]

14.2.1 The Basic Fabrication Process

The majority of bipolar transistor used in ICs are of the *n-p-n* type because the higher mobility of minority carriers (electrons) in the base region results in higher-speed performance than can be obtained with *p-n-p* types. Figure 7 shows a perspective view of an *n-p-n* bipolar transistor, in which lateral isolation is provided by oxide walls and vertical isolation is provided by the n^+-p junction. The lateral oxide isolation approach reduces not only the device size but also the parasitic capacitance because of the smaller dielectric constant of silicon dioxide (3.9, compared with 11.9 for silicon). We consider the major process steps that are used to fabricate the device shown in Fig. 7.

For an *n-p-n* bipolar transistor, the starting material is a *p*-type lightly doped ($\sim 10^{15}$ cm^{-3}), ⟨111⟩- or ⟨100⟩-oriented, polished silicon wafer. Because the junctions are formed inside the semiconductor, the choice of crystal orientation is not as critical as for MOS devices. The first step is to form a buried layer. The main purpose of this layer is to minimize the series resistance of the collector. A thick oxide (0.5–1 μm) is thermally grown on the wafer, and a window is then opened in the oxide. A precisely controlled amount of low-energy arsenic ions (~30 keV, $\sim 10^{15}$ cm^{-2}) is implanted into the window region to serve as a predeposit (Fig. 8*a*). Next, a high temperature (~1100°C) drive-in step forms the n^+-buried layer, which has a typical sheet resistance of 20 Ω/□.

The second step is to deposit an *n*-type epitaxial layer. The oxide is removed and the wafer is placed in an epitaxial reactor for epitaxial growth. The thickness and the doping concentration of the epitaxial layer are determined by the ultimate use of the device. Analog circuits (with their higher voltages for amplification) require thicker layer (~10 μm) and lower dopings ($\sim 5 \times 10^{15}$ cm^{-3}), whereas digital circuits (with their lower voltages for switching) require thinner layers (~3 μm) and higher dopings ($\sim 2 \times 10^{16}$ cm^{-3}). Figure 8*b* shows a cross-sectional view of the device after the epitaxial process. Note that there is some outdiffusion from the buried layer into the epitaxial layer. To minimize the outdiffusion, a low-temperature epitaxial process should be employed, and low-diffusivity impurities should be used in the buried layer (e.g., As).

The third step is to form the lateral oxide isolation region. A thin-oxide pad (~50 nm) is thermally grown on the epitaxial layer, followed by a silicon-nitride deposition (~100 nm). If nitride is deposited directly onto the silicon without the thin-oxide pad, the nitride may cause damages to the silicon surface during the subsequent high-temperature steps. Next, the nitride-oxide layers and about half of the epitaxial layer are etched using a photoresist

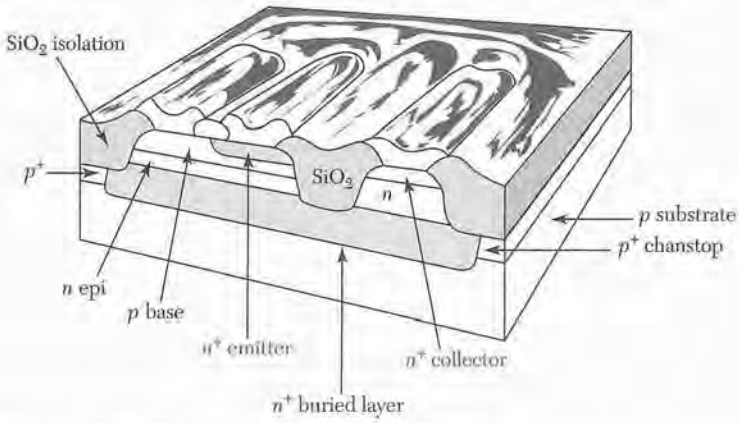


Fig. 7 Perspective view of an oxide-isolated bipolar transistor.

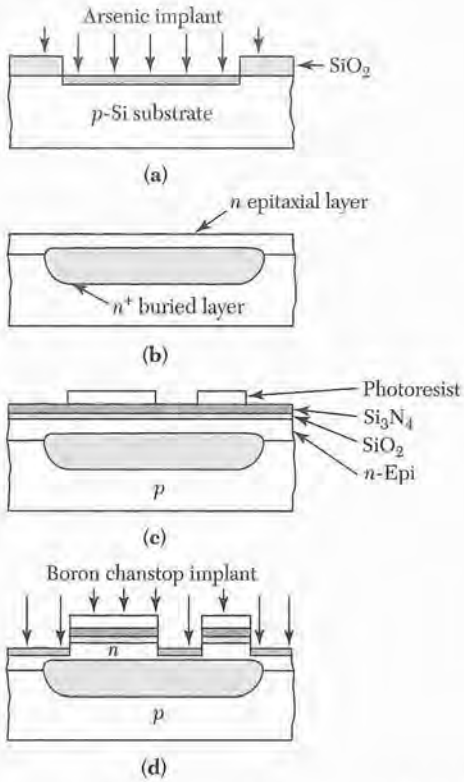


Fig. 8 Cross-sectional views of bipolar transistor fabrication. (*a*) Buried-layer implantation. (*b*) Epitaxial layer. (*c*) Photoresist mask. (*d*) Chanstop implant.

as mask (Fig. 8*c* and 8*d*). Boron ions are then implanted into the exposed silicon areas (Fig. 8*d*).

The photoresist is removed and the wafer is placed in an oxidation furnace. Since the nitride layer has a very low oxidation rate, thick oxides will be grown only in the areas not protected by the nitride layer. The isolation oxide is usually grown to a thickness such that the top of the oxide becomes coplanar with the original silicon surface to minimize the surface topography. This oxide isolation process is called local oxidation of silicon (LOCOS). Figure 9*a* shows the cross section of the isolation oxide after the removal of the nitride layer. Becaue of segregation effects, most of the implanted boron ions are pushed underneath the isolation oxide to form a p^+-layer. This is called p^+ channel stop (or chanstop), because the high concentration of *p*-type semiconductor will prevent surface inversion and eliminate possible high-conductivity paths (or channels) among neighboring buried layers.

The fourth step is to form the base region. A photoresist is used as a mask to protect the right half of the device; then, boron ions ($\sim 10^{12}$ cm^{-2}) are implanted to form the base regions, as shown in Fig. 9*b*. Another lithographic process removes all the thin-pad oxide except a small area near the center of the base region (Fig. 9*c*).

The fifth step is to form the emitter region. As shown in Fig. 9*d*, the base contact area is protected by a photoresist mask; then, a low-energy, high-arsenic–dose ($\sim 10^{16}$ cm^{-2}) implantation forms the n^+-emitter and the n^+-collector contact regions. The photoresist

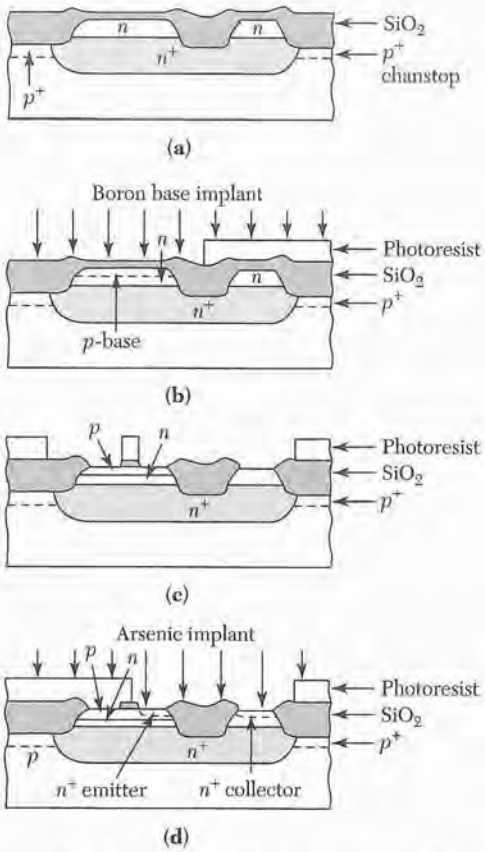


Fig. 9 Cross-section views of bipolar transistor fabrication. (*a*) Oxide isolation. (*b*) Base implant. (*c*) Removal of thin oxide. (*d*) Emitter and collector implant.

is removed; and a final metallization step forms the contacts to the base, emitter, and collector as shown in Fig. 7.

In this basic bipolar process, there are six film formation operations, six lithographic operations, four ion implantations, and four etching operations. Each operation must be precisely controlled and monitored. Failure of any one of the operations generally will render the wafer useless.

The doping profiles of the completed transistor along a coordinate perpendicular to the surface and passing through the emitter, base, and collector are shown in Fig. 10. The emitter profile is abrupt because of the concentration-dependent diffusivity of arsenic. The base doping profile beneath the emitter can be approximated by a Gaussian distribution for a limited-source diffusion. The collector doping is given by the epitaxial doping level ($\sim 2 \times 10^{16}$ cm^{-3}) for a representative switching transistor; however, at larger depths, the collector doping concentration increases because of outdiffusion from the buried layer.

14.2.2 Dielectric Isolation

In the isolation scheme described previously for the bipolar transistor, the device is isolated from other devices by the oxide layer around its periphery and is isolated from its common substrate by a n^+–p junction (buried layer). In high-voltage applications, a different

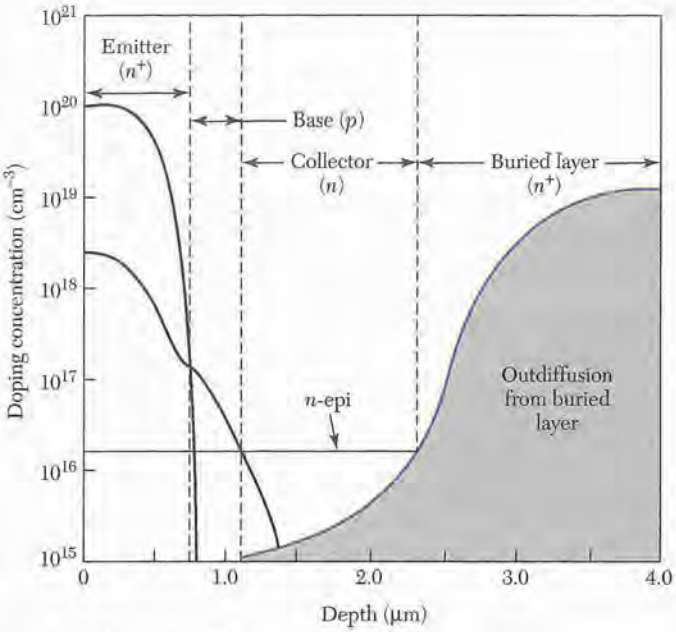


Fig. 10 n-p-n transistor doping profiles.

approach, called dielectric isolation, is used to form insulating tubs to isolate a number of pockets of single-crystal semiconductors. In this approach the device is isolated from both its common substrate and its surrounding neighbors by a dielectric layer.

A process sequence for the dielectric isolation is shown in Fig. 11. An oxide layer is formed inside a <100>-oriented n-type silicon substrate using high-energy oxygen ion implantation (Fig. 11*a*). Next, the wafer undergoes a high-temperature annealing process so that the implanted oxygen will react with silicon to form the oxide layer. The damage resulting from implantation is also annealed out in this process (Fig. 11*b*). After this, we can obtain an n-silicon layer that is fully isolated on an oxide [namely, silicon-on-insulator, (SOI)]. This process is called SIMOX (separation by implanted oxygen). Since the top silicon is so thin, the isolation region is easily formed by the LOCOS process illustrated in Fig. 8*c* or by etching a trench (Fig. 11*c*) and refilling it with oxide (Fig. 11*d*). The other processes are almost the same as those from Fig. 8*c* through Fig. 9 to form the p-type base, n^+-emitter, and collector.

The main advantage of this technique is its high breakdown voltage between the emitter and the collector, which can be in excess of several hundreds volts. This technique is also compatible with modern CMOS integration. This CMOS-compatible process is very useful for mixed high-voltage and high-density IC.

14.2.3 Self-Aligned Double-Polysilicon Bipolar Structure

The process shown in Fig. 9*c* needs another lithographic process to define an oxide region to separate the base and emitter contact regions. This gives rise to a large inactive device area within the isolated boundary, which increases not only the parasitic capacitances but also the resistance that degrades the transistor performance. The most effective way to reduce these effects is by using the self-aligned structure.

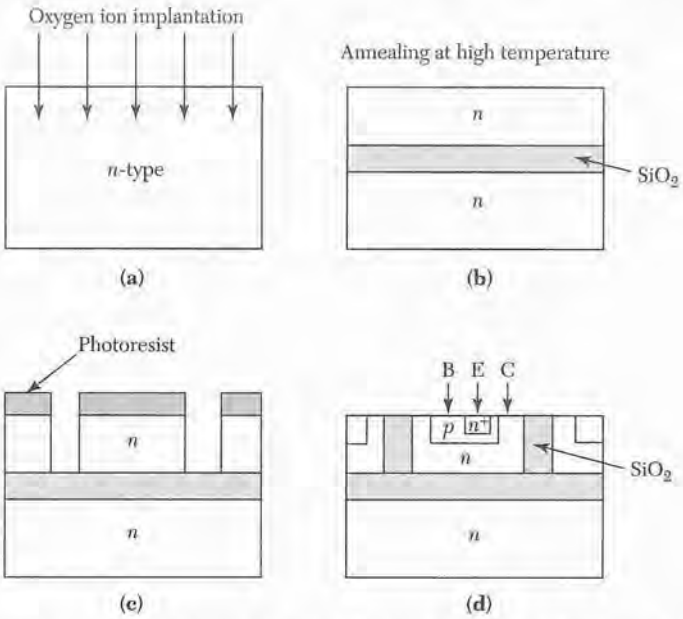


Fig. 11 Process sequence for dielectric isolation bipolar device using silicon-on-insulator for high-voltage application. (*a*) Oxygen ion implantation. (*b*) Annealing at high temperature to form the isolation dielectric. (*c*) Trench isolation formed by a dry-etching process. (*d*) Base, emitter, and collector formation.

The most widely used self-aligned structure is the double-polysilicon structure with the advanced isolation provided by a trench refilled with polysilicon,[5] shown in Fig. 12. Figure 13 shows the detail sequence of the steps for the self-aligned double-polysilicon (n-p-n) bipolar structure.[6] The transistor is built on an n-type epitaxial layer. A trench of 5.0 μm in depth is etched by reactive ion etching through the n^+-subcollector region into the p^-substrate region. A thin layer of thermal oxide is then grown and serves as the screen oxide for the channel stop implant of boron at the bottom of the trench. The trench is then filled with undoped polysilicon and capped by a thick planar field oxide.

The first polysilicon layer is deposited and heavily doped with boron. The p^+-polysilicon (called poly 1) will be used as a solid-phase diffusion source to form the extrinsic base region and the base electrode. This layer is covered with a chemical-vapor deposition

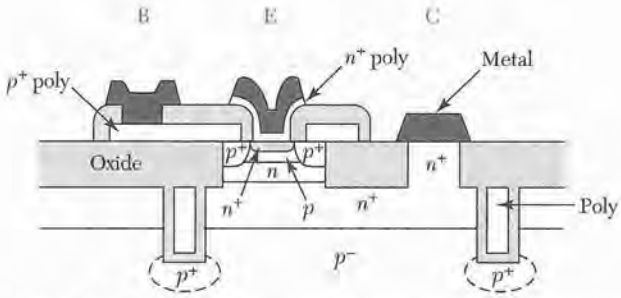


Fig. 12 Cross-section of a self-aligned, double-polysilicon bipolar transistor with advanced trench isolation.[5]

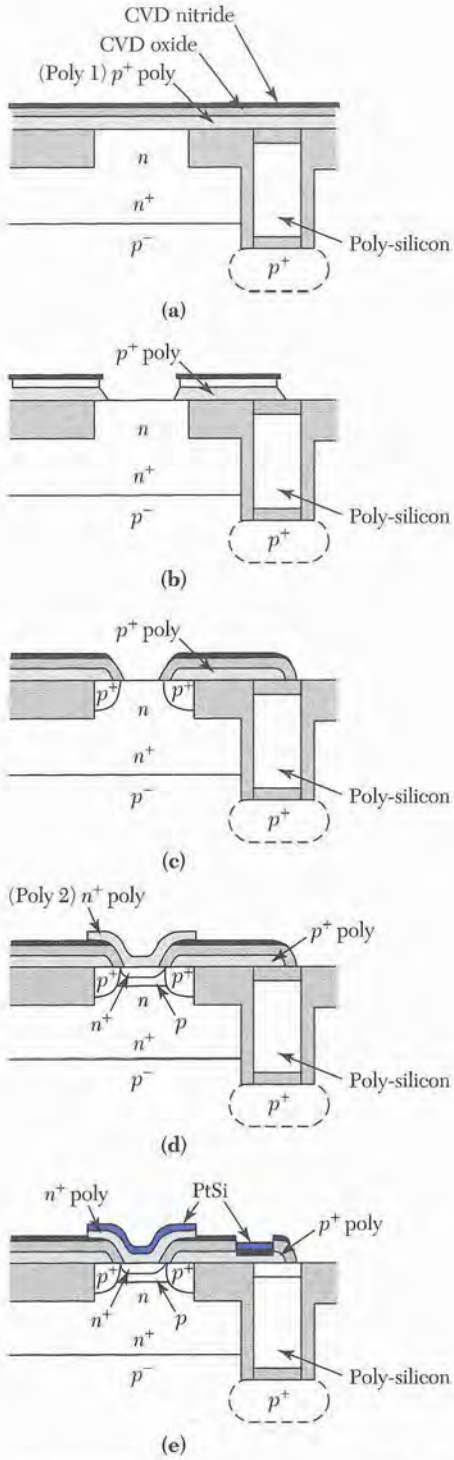


Fig. 13 Process sequence for fabricating double-polysilicon self-aligned n-p-n transistors.[6]

(CVD) oxide and nitride (Fig. 13*a*). The emitter mask is used to pattern the emitter-area regions, and a dry-etch process is used to produce an opening in the CVD oxide and poly 1 (Fig. 13*b*). A thermal oxide is then grown over the etched structure, and a relatively thick oxide (approximately 0.1–0.4 μm) is grown on the vertical sidewalls of the heavily doped poly. The thickness of this oxide determines the spacing between the edges of the base and emitter contacts. The extrinsic p^+ base regions are also formed during the thermal-oxide growth step as a result of the outdiffusion of boron from the poly 1 into the substrate (Fig. 13*c*). Because boron diffuses laterally as well as vertically, the extrinsic base region will be able to make contact with the intrinsic base region that is formed next, under the emitter contact.

Following the oxide-grown step, the intrinsic base region is formed using ion implantation of boron (Fig. 13*d*). This serves to self-align the intrinsic and extrinsic base regions. After the contact is cleaned to remove any oxide layer, the second polysilicon layer is deposited and implanted with As or P. The n^+–polysilicon (called poly 2) is used as a solid-phase diffusion source to form the emitter region and the emitter electrode. A shallow emitter region is then formed through dopant outdiffusion from poly 2. A rapid thermal anneal for the base and emitter outdiffusion steps facilitates the formation of shallow emitter-base and collector-base junctions. Finally, Pt film is deposited and sintered to form PtSi over the n^+-polysilicon emitter and the p^+-polysilicon base contact (Fig. 13*e*).

This self-aligned structure allows the fabrication of emitter regions smaller than the minimum lithographic dimension. When the sidewall-spacer oxide is grown, it fills the contact hole to some degree because the thermal oxide occupies a larger volume than the original volume of polysilicon. Thus, an opening 0.8 μm wide will shrink to about 0.4 μm if sidewall oxide a 0.2 μm thick is grown on each side.

14.3 MOSFET TECHNOLOGY

At present, the MOSFET is the dominant device used in ULSI circuits because it can be scaled to smaller dimensions than other types of devices. The dominant technology for MOSFET is the CMOS (complementary MOSFET) technology, in which both *n*-channel and *p*-channel MOSFETs (called NMOS and PMOS, respectively) are provided on the same chip. CMOS technology is particular attractive for ULSI circuits because it has the lowest power consumption of all IC technology.

Figure 14 shows the reduction in the size of the MOSFET in recent years. In the early 1970s, the gate length was 7.5 μm and the corresponding device area was about 6000 μm^2. As the device is scaled down, there is a drastic reduction in the device area. For a MOSFET with a gate length of 0.5 μm, the device area shrinks to less than 1% of the early MOSFET. We expect that device miniaturization will continue. The gate length will be less than 0.10 μm in the early twenty-first century. We consider the future trends of the devices in Section 14.5.

14.3.1 The Basic Fabrication Process

Figure 15 shows a perspective view of an *n*-channel MOSFET prior to its final metallization.[7] The top layer is a phosphorus-doped silicon dioxide (P-glass) that is used as an insulator between the polysilicon gate and the gate metallization and also as a gettering layer for mobile ions. Compare Fig. 15 with Fig. 7 for the bipolar transistor and note that a MOSFET is considerably simpler in its basic structure. Although both devices use

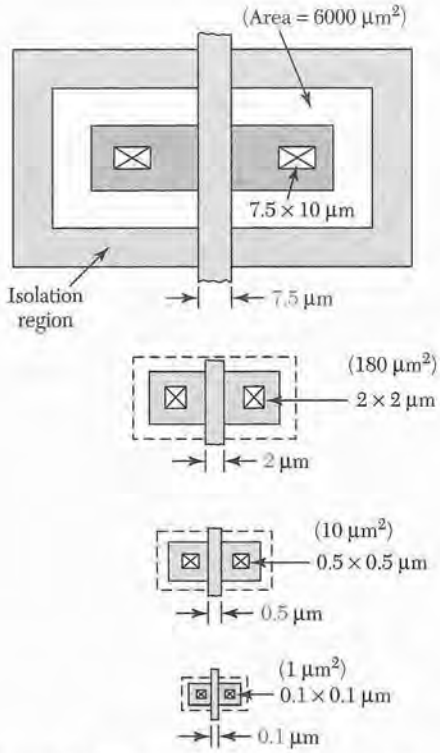


Fig. 14 Reduction in the area of the MOSFET as the gate length (minimum feature length) is reduced.

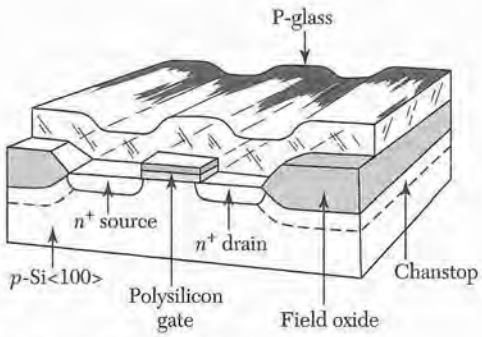


Fig. 15 Perspective view of an n-channel MOSFET.[7]

lateral oxide isolation, there is no need for vertical isolation in the MOSFET, whereas a buried-layer n^+-p junction is required in the bipolar transistor. The doping profile in a MOSFET is not as complicated as that in a bipolar transistor and the control of the dopant distribution is also less critical. We consider the major process steps that are used to fabricate the device shown in Fig. 15.

To process an n-channel MOSFET (NMOS), the starting material is a p-type, lightly doped ($\sim 10^{15}$ cm^{-3}), ⟨100⟩-oriented, polished silicon wafer. The ⟨100⟩-orientation is preferred over ⟨111⟩ because it has an interface-trap density that is about one-tenth that of ⟨111⟩. The first step is to form the oxide isolation region using LOCOS technology. The

process sequence for this step is similar to that for the bipolar transistor. A thin-pad oxide (~35 nm) is thermally grown, followed by a silicon nitride (~150 nm) deposition (Fig. 16*a*).[7] The active device area is defined by a photoresist mask and a boron chanstop layer is then implanted through the composite nitride-oxide layer (Fig. 16*b*). The nitride layer not covered by the photoresist mask is subsequently removed by etching. After stripping the photoresist, the wafer is placed in an oxidation furnace to grow an oxide (called the field oxide), where the nitride layer is removed, and to drive in the boron implant. The thickness of the field oxide is typically 0.5–1 μm.

The second step is to grow the gate oxide and to adjust the threshold voltage (see Section 6.2.3). The composite nitride-oxide layer over the active device area is removed, and a thin-gate oxide layer (less than 10 nm) is grown. For an enhancement-mode n-channel device, boron ions are implanted in the channel region, as shown in Fig. 16*c*, to increase the threshold voltage to a predetermined value (e.g., + 0.5V). For a depletion-mode *n*-channel device, arsenic ions are implanted in the channel region to decrease the threshold voltage (e.g., −0.5V).

The third step is to form the gate. A polysilicon is deposited and is heavily doped by diffusion or implantation of phosphorus to a typical sheet resistance of 20–30 Ω/□. This resistance is adequate for MOSFETs with gate lengths larger than 3 μm. For smaller

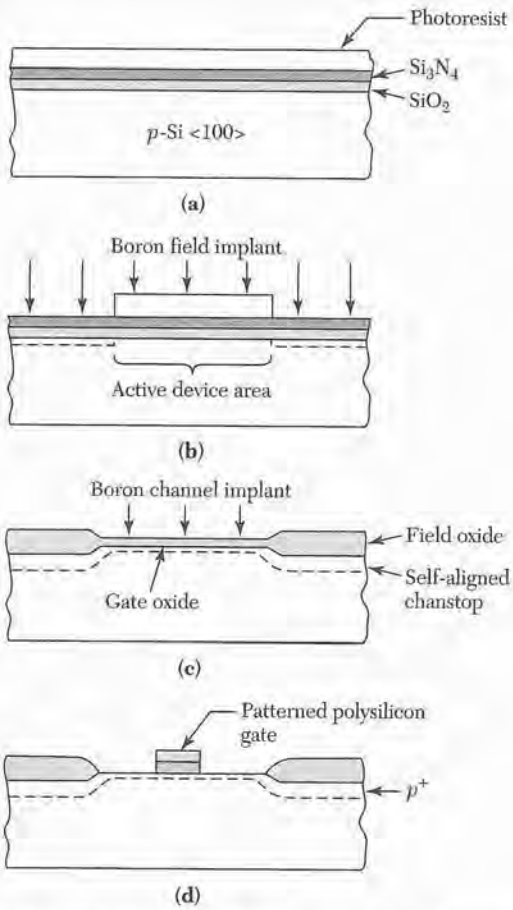


Fig. 16 Cross-sectional view of NMOS fabrication sequence.[7] (*a*) Formation of SiO_2, Si_3N_4, and photoresist layer. (*b*) Boron implant. (*c*) Field oxide. (*d*) Gate.

devices, polycide, a composite layer of metal silicide and polysilicon such as W-polycide, can be used as the gate materials to reduce the sheet resistance to about 1 Ω/□.

The fourth step is to form the source and drain. After the gate is patterned (Fig. 16*d*), it serves as a mask for the arsenic implantation (~30 keV, ~5×10^{15} cm^{-2}) to form the source and drain (Fig. 17*a*), which are self-aligned with respect to the gate[7]. At this stage, the only overlapping of the gate is due to lateral straggling of the implanted ions (for 30 keV As, $\sigma_{\perp}$ is only 5 nm). If low-temperature processes are used for subsequent steps to minimize lateral diffusion, the parasitic gate-drain and gate-source coupling capacitances can be much smaller than the gate-channel capacitance.

The last step is the metallization. A phosphorus-doped oxide (P-glass) is deposited over the entire wafer and is flowed by heating the wafer to give a smooth surface topography (Fig. 17*b*). Contact windows are defined and etched in the P-glass. A metal layer, such as aluminum, is then deposited and patterned. A cross-section view of the completed MOSFET is shown in Fig. 17*c*, and the corresponding top view is shown in Fig. 17*d*. The gate contact is usually made outside the active device area to avoid possible damage to the thin-gate oxide.

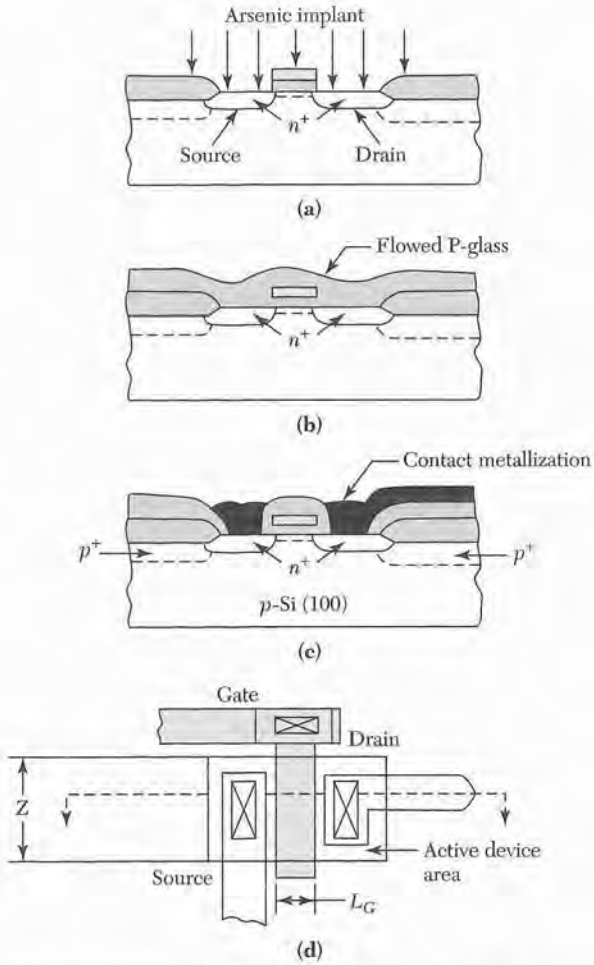


Fig. 17 NMOS fabrication sequence.[7] (*a*) Source and drain. (*b*) P-glass deposition. (*c*) Cross section of the MOSFET. (*d*) Top view of the MOSFET.

EXAMPLE 4

What is the maximum gate-to-source voltage that a MOSFET with a 5 nm gate oxide can withstand. Assume that the oxide breaks down at 8 MV/ cm and the substrate voltage is zero.

SOLUTION

$V = \mathscr{E} \times d = 8 \times 10^6 \times 5 \times 10^{-7} = 4$ V.

14.3.2 Memory Devices

Memories are devices that can store digital information (or data) in terms of *bits* (*bi*nary digi*ts*). Various memory chips have been designed and fabricated using NMOS technology. For most large memories, the random access memory (RAM) organization is preferred. In a RAM, memory cells are organized in a matrix structure and data can be accessed (i.e., stored, retrieved, or erased) in random order, independent of their physical locations. A static random access memory (SRAM) can retain stored data indefinitely as long as the power supply is on. The SRAM is basically a flip-flop circuit that can store one bit of information. A SRAM cell has four enhancement-mode MOSFETs and two depletion-mode MOSFETs. The depletion-mode MOSFETs can be replaced by resistors formed in undoped polysilicon to minimize power consumption.[8]

To reduce the cell area and power consumption, the dynamic random access memory (DRAM) has been developed. Figure 18*a* shows the circuit diagram of the one-transistor DRAM cell in which the transistor serves as a switch and one bit of information can be stored in the storage capacitor. The voltage level on the capacitor determines the state of the cell. For example, +1.5 V may be defined as logic 1 and 0 V defined as logic 0. The stored charge will be removed typically in a few milliseconds mainly because of the leakage current of the capacitors; thus, dynamic memories require periodic "refreshing" of the stored charge.

Figure 18*b* shows the layout of a DRAM cell, and Fig. 18*c* shows the corresponding cross section through AA′. The storage capacitor uses the channel region as one plate, the polysilicon gate as the other plate, and the gate oxide as the dielectric. The row line is a metal track to minimize the delay due to parasitic resistance (R) and parasitic capacitance (C), the RC delay. The column line is formed by n^+-diffusion. The internal drain region of the MOSFET serves as a conductive link between the inversion layers under the storage gate and the transfer gate. The drain region can be eliminated by using the double-level polysilicon approach shown in Fig. 18*d*. The second polysilicon electrode is separated from the first polysilicon capacitor plate by an oxide layer that is thermally grown on the first-level polysilicon before the second electrode has been defined. The charge from the column line can therefore be transmitted directly to the area under the storage gate by the continuity of inversion layers under the transfer and storage gates.

To meet the requirements of high-density DRAM, the DRAM structure has been extended to the third dimension with stacked or trench capacitors. Figure 19*a* shows a simple trench cell structure.[9] The advantage of the trench type is that the capacitance of the cell could be increased by increasing the depth of the trench without increasing the surface area of silicon occupied by the cell. The main difficulties of making trench-type cells are the etching of the deep trench, which needs a rounded bottom corner and the growth of a uniform thin dielectric film on trench walls. Figure 19*b* shows a stacked cell structure. The storage capacitance increases as a result of stacking the storage capacitor on top of the access transistor. The dielectric is formed using the thermal oxidation

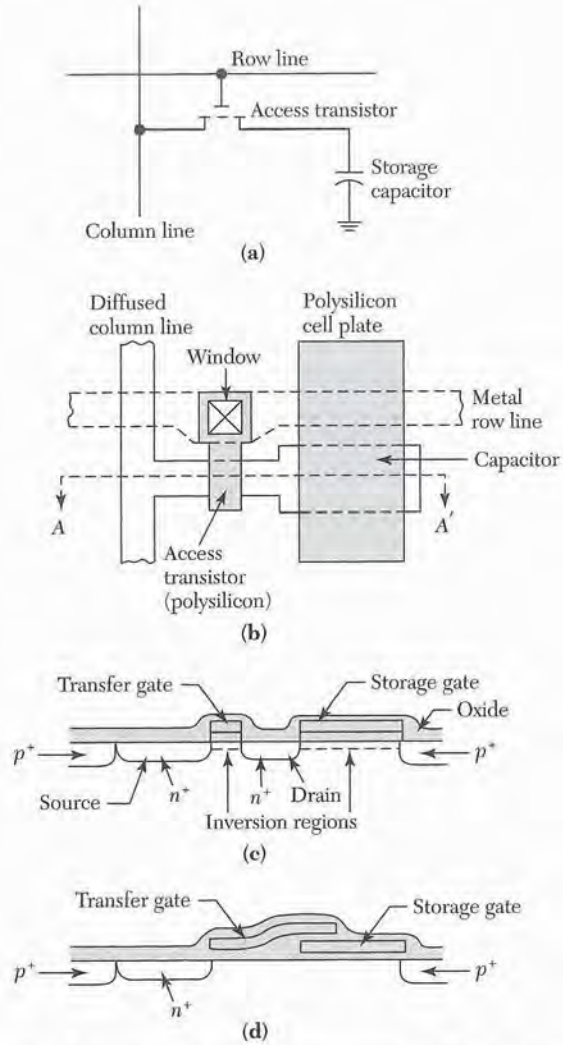


Fig. 18 Single-transistor dynamic random access memory (DRAM) cell with a storage capacitor.[8] (*a*) Circuit diagram. (*b*) Cell layout. (*c*) Cross section through *A-A′*. (*d*) Double-level polysilicon.

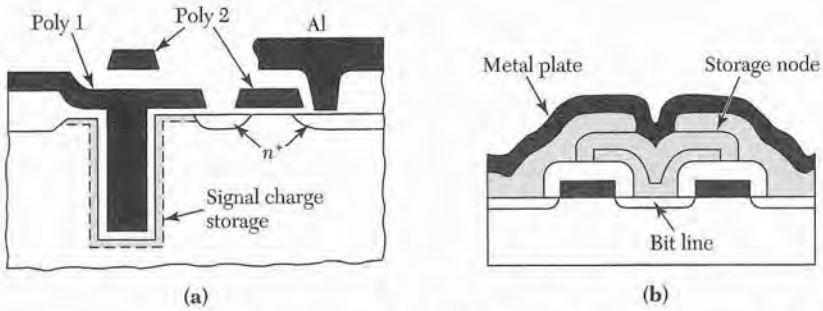


Fig. 19 (*a*) DRAM with a trench cell structure.[9] (*b*) DRAM with a single-layer stacked-capacitor cell.

or CVD nitride methods between the two-polysilicon plates. Hence, the stacked cell process is easier than the trench type process.

Figure 20 shows a 1 Gb DRAM chip. This memory chip uses 0.18 μm design rules. Trench capacitors and its peripheral circuits are in CMOS, which are considered in Section 14.3.3. The memory chip has an area of 390 mm^2 (14.3 mm × 27.3 mm) that contains over 2 billion components and operates at 2.5 V. This 1 Gb DRAM is mounted in an 88-pin ceramic package, which can provide adequate heat dissipation.

Both SRAM and DRAM are volatile memories, that is, they lose their stored data when power is switched off. Nonvolatile memories, on the other hand, can retain their data. Figure 21a shows a floating-gate nonvolatile memory, which is basically a conventional MOSFET that has a modified gate electrode. The composite gate has a regular (control gate) and a floating gate which is surrounded by insulators. When a large positive voltage is applied to the control gate, charge will be injected from the channel region through the gate oxide into the floating gate. When the applied voltage is removed, the injected charge can be stored in the floating gate for a long time. To remove this charge, a large negative voltage must be applied to the control gate, so that the charge will be injected back into the channel region.

Another version of the nonvolatile memory is the metal-nitride-oxide-semiconductor (MNOS) type shown inn Fig. 21*b*. When a positive gate voltage is applied, electrons can tunnel through the thin oxide layer (~2 nm) and be captured by the traps at the oxide-nitride interface, and thus become stored charges there. The equivalent circuit for both types of nonvolatile memories can be represented by two capacitors in series for the gate structure, as illustrated in Fig. 21*c*. The charge stored in the capacitor C_1 causes a shift in the threshold voltage, and the device remains at the higher threshold voltage-state (logic 1). For a well-designed memory device, the charge retention time can be over 100 years.

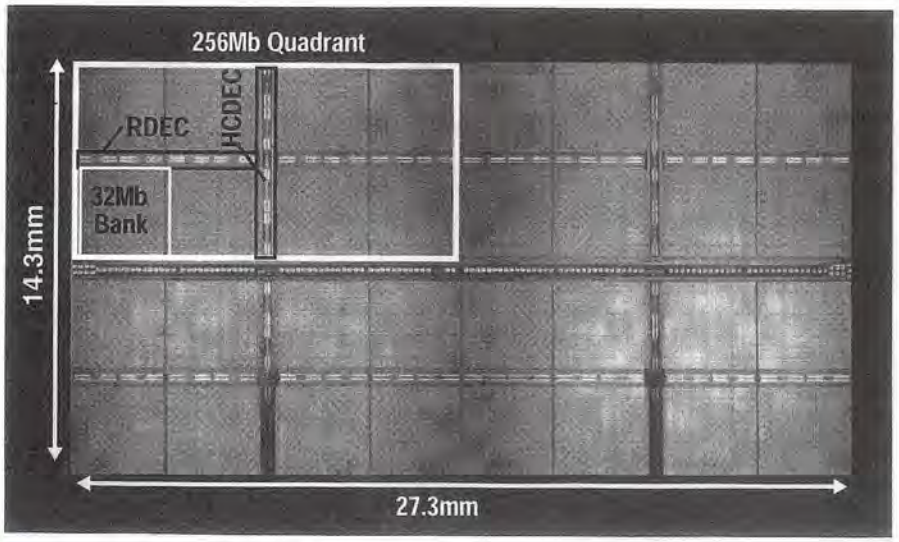


Fig. 20 A 1 Gb DRAM that contains over 2 billion components. (Photography courtesy of IBM/Siemens, 1999 IEEE Int. Solid State Circuit Conference.)

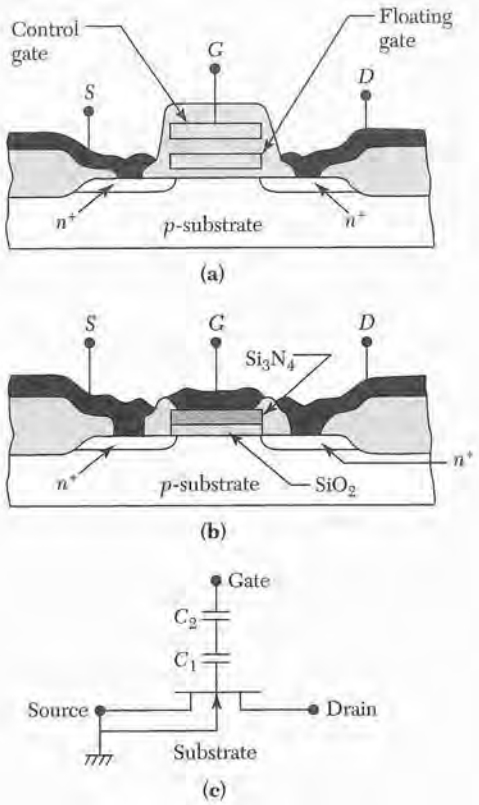


Fig. 21 Nonvolatile memory devices. (*a*) Floating-gate, nonvolatile memory. (*b*) MNOS nonvolatile memory. (*c*) Equivalent circuit of either type of nonvolatile memory.

To erase the memory (e.g., the store charge) and return the device to a lower threshold voltage state (logic 0), a gate voltage or other means (such as ultraviolet light) can be used.

The nonvolatile semiconductor memory (NVSM) has been extensively used in portable electronics systems, such as cellular phones and the digital cameras. Another interesting application is the chip card, also called IC card.

The top photo in Fig. 22 shows an IC card. The diagram at the bottom of Fig. 22 illustrates the nonvolatile memory device that stores the data that can be read and written through the bus to a central processing unit (CPU). In contrast to the limited volume (1 kbytes) inside a conventional magnetic tape card, the size of the nonvolatile memory can be increased to 16 kbytes, 64 kbytes, or even larger depending on the applications (e.g., you can store personal photos or finger prints). Through the IC card read/write machines, the data can be used in numerous applications, such as telecommunications (card telephone, mobile radio), payment transactions (electronic purse, credit card), pay television, transport (electronic ticket, public transport), health care (patient-data card), and access control. The IC card will play a central role in the global information and service society of the future.[10]

14.3.3 CMOS Technology

Figure 23*a* shows a CMOS inverter. The gate of the upper PMOS device is connected to the gate of the lower NMOS device. Both devices are enhancement-mode MOSFETs

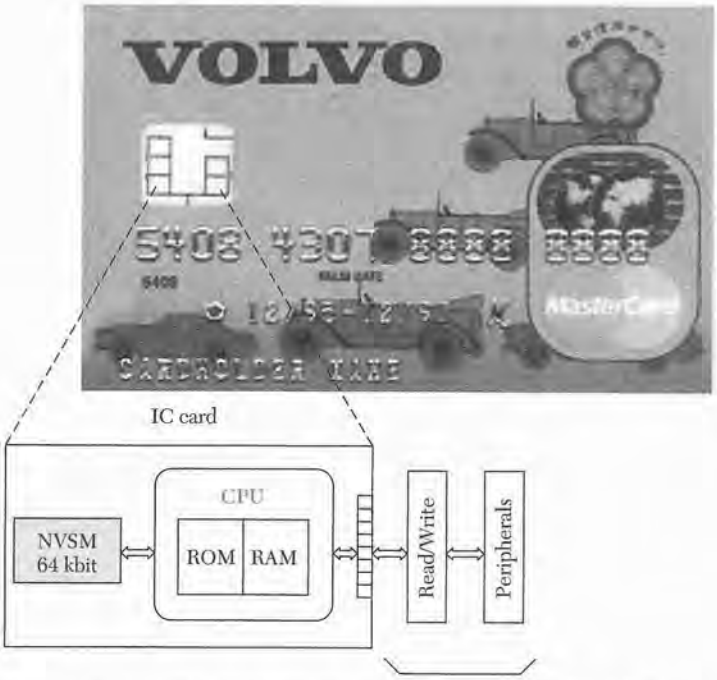


Fig. 22 An integrated-circuit (IC) card. The data stored in the NVSM can be accessed through the bus of the central processing unit (CPU). There are several metal pads connecting to the read/write machine. (Photography courtesy of Retone Information System Co., LTD.)

with the threshold voltage V_{Tp} less than zero for the PMOS device and V_{Tn} greater than zero for the NMOS device (typically the threshold voltage is about 1/4 V_{DD}). When the input voltage V_i is at ground or at small positive values, the PMOS device is turned on (the gate-to-ground potential of PMOS is $-V_{DD}$, which is more negative than V_{Tp}), and the NMOS device is off. Hence, the output voltage V_o is very close to V_{DD} (logic 1). When the input is at V_{DD}, the PMOS (with $V_{GS} = 0$) is turned off, and the NMOS is turned on ($V_i = V_{DD} > V_{Tn}$). Therefore, the output voltage V_o equals zero (logic 0). The CMOS inverter has a unique feature: in either logic state, one device in the series path from V_{DD} to ground is nonconductive. The current that flows in either steady state is a small leakage current, and only when both devices are on during switching does a significant current flow through the CMOS inverter. Thus, the average power dissipation is small, on the order of nanowatts. As the number of components per chip increases, the power dissipation becomes a major limiting factor. The low power consumption is the most attractive feature of the CMOS circuit.

Figure 23*b* shows a layout of the CMOS inverter, and Fig. 23*c* shows the device cross section along the *A*-*A′* line. In the processing, a *p*-tub (also called a *p*-well) is first implanted and subsequently driven into the *n*-substrate. The *p*-type dopant concentration must be high enough to overcompensate the background doping of the *n*-substrate. The subsequent processes for the *n*-channel MOSFET in the *p*-tub are identical to those described previously. For the *p*-channel MOSFET, $^{11}B^+$ or $^{49}(BF_2)^+$ ions are implanted into the *n*-substrate to form the source and drain regions. A channel implant of $^{75}As^+$ ions may be used to adjust the threshold voltage and a n^+-chanstop is formed underneath the field

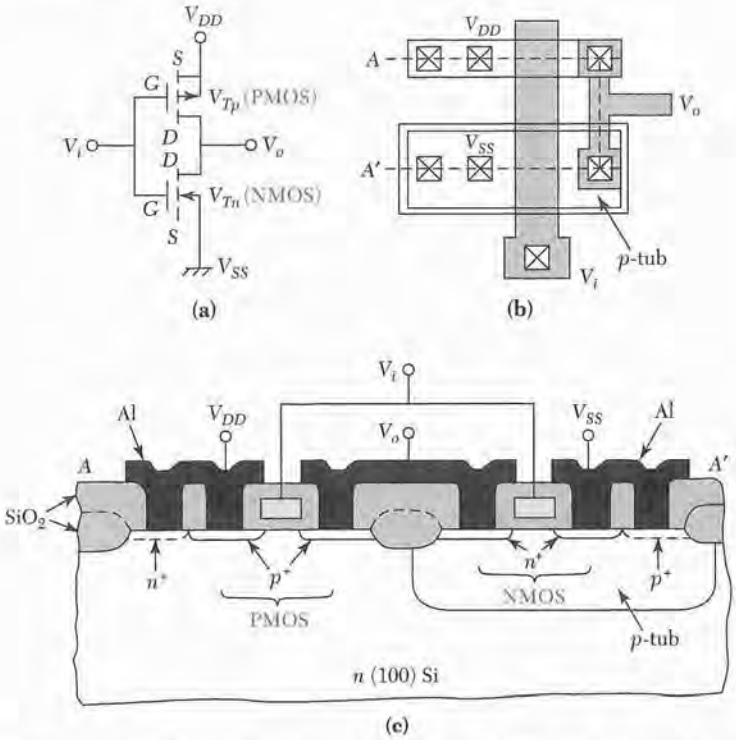


Fig. 23 Complementary MOS (CMOS) inverter. (*a*) Circuit diagram. (*b*) Circuit layout. (*c*) Cross section along dotted A-A′ line of (*b*).

oxide around the *p*-channel device. Because of the *p*-tub and the additional steps needed to make the *p*-channel MOSFET, the number of steps to make a CMOS circuit is essentially double that to make an NMOS circuit. Thus, we have a trade-off between the complexity of processing and a reduction in power consumption.

Instead of the *p*-tub described above, an alternate approach is to use an *n*-tub formed in *p*-type substrate, as shown in Fig. 24*a*. In this case, the *n*-type dopant concentration must be high enough to overcompensate for the background doping of the p-substrate (i.e., $N_D > N_A$). In both the *p*-tub and the *n*-tub approach, the channel mobility will be degraded because mobility is determined by the total dopant concentration ($N_A + N_D$). A recent approach using two separated tubs implanted into a lightly doped substrate is shown in Fig. 24*b*. This structure is called a *twin* tub.[1] Because no overcompensation is needed in either of the twin tubs, higher channel mobility can be obtained.

All CMOS circuits have the potential for a troublesome problem called latchup that is associated with parasitic bipolar transistors (to see how this problem can occur, see Chapter 5). An effective processing technique to eliminate latchup problem is to use the deep-trench isolation, as shown[11] in Fig. 24*c*. In this technique, a trench with a depth deeper than the well is formed in the silicon by anisotropic reactive sputter etching. An oxide layer is thermally grown on the bottom and walls of the trench, which is then refilled by deposited polysilicon or silicon dioxide. This technique can eliminate latchup because the *n*-channel and *p*-channel devices are physically isolated by the refilled trench. The detailed steps for trench isolation and some related CMOS processes are now considered.

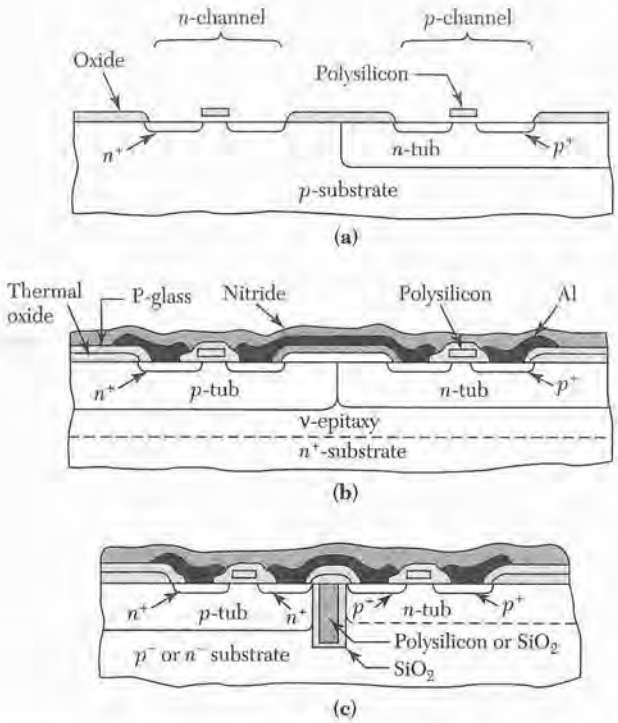


Fig. 24 Various CMOS structures. (*a*) *n*-tub. (*b*) Twin tub[1]. (*c*) Refilled trench.[11]

Well-Formation Technology

The well of a CMOS can be a single well, a twin well, or a retrograde well. The twin-well process exhibits some disadvantages, e.g., it needs high temperature processing (above 1050°C) and a long diffusion time (longer than 8 hours) to achieve the required depth of 2–3 μm. In this process, the doping concentration is highest at the surface and decreases monotonically with depth. To reduce the process temperature and time, high-energy implantation is used, i.e., implanting the ion to the desired depth instead of diffusion from the surface. Since the depth is determined by the implantion energy, we can design the well depth with different implantation energy. The profile of the well in this case can have a peak at a certain depth in the silicon substrate. This is called a retrograde well. Figure 25 shows a comparison of the impurity profiles in the retrograde well and the conventional thermal diffused well.[12] The energy for the *n*- and *p*-type retrograde wells is around 700 keV and 400 keV, respectively. As mentioned above, the advantage of the high-energy implantation is that it can form the well under low-temperature and short-time conditions, hence, it can reduce the lateral diffusion and increase the device density. The retrograde well can offer some additional advantages over the conventional well: (a) because of high doping near the bottom, the well resistivity is lower than that of the conventional well and the latchup problem can be minimized, (b) the chanstop can be formed at the same time as the retrograde well implantation, reducing processing steps and time, (c) higher well doping in the bottom can reduce the chance of punchthrough from the drain to the source.

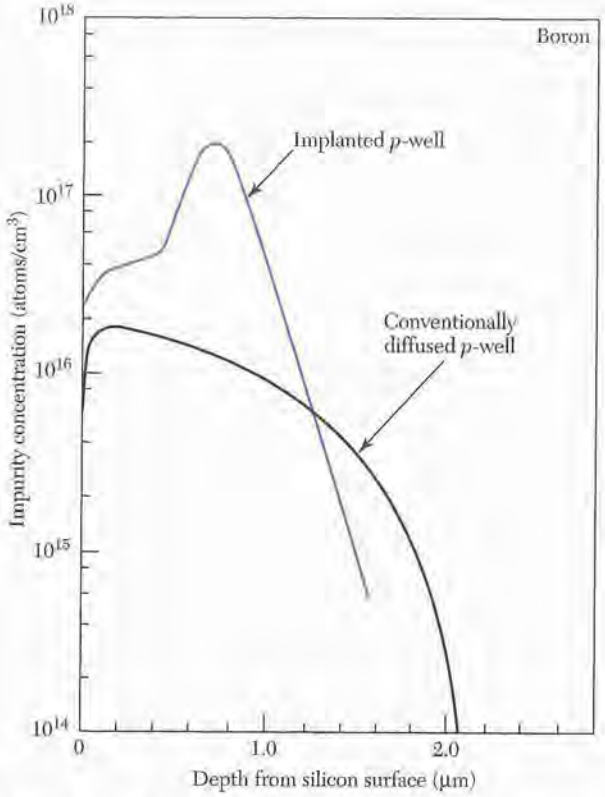


Fig. 25 Retrograded p-well implanted impurity concentration profile. Also shown is a conventionally diffused well.[12]

Advanced Isolation Technology

The conventional isolation process (Section 14.3.1) has some disadvantages that make it unsuitable for deep-submicron (0.25 μm and smaller) fabrications. The high-temperature oxidation of silicon and long oxidation time result in the encroachment of the chanstop implantation (usually boron for n-MOSFET) to the active region and cause V_T shift. The area of the active region is reduced because of the lateral oxidation. In addition, the field-oxide thickness in submicron-isolation spacings is significantly less than the thickness of field oxide grown in wider spacings. The trench isolation technology can avoid these problems and has become the mainstream technology for isolation. Figure 26 shows the process sequence for forming a deep (larger than 3 μm) but narrow (less than 2 μm) trench-isolation structure. There are four steps: patterning the area, trench etching and oxide growth, refilling with dielectric materials such as oxide or undoped polysilicon, and planarization. This deep trench isolation can be used in both advanced CMOS and bipolar devices and for the trench-type DRAM. Since the isolation material is deposited by CVD, it does not need a long-time or a high-temperature process, and it eliminates the lateral oxidation and boron encroachment problems.

Another example is the shallow trench (depth is less than 1 μm) isolation for CMOS, shown in Fig. 27. After patterning (Fig. 27*a*), the trench area is etched (Fig. 27*b*) and then re-filled with oxide (Fig. 27*c*). Before refilling, a chanstop implantation can be performed. Since the oxide has over filled the trench, the oxide on the nitride should be removed. Chemical-mechanical polishing (CMP) is used to remove the oxide on the nitride

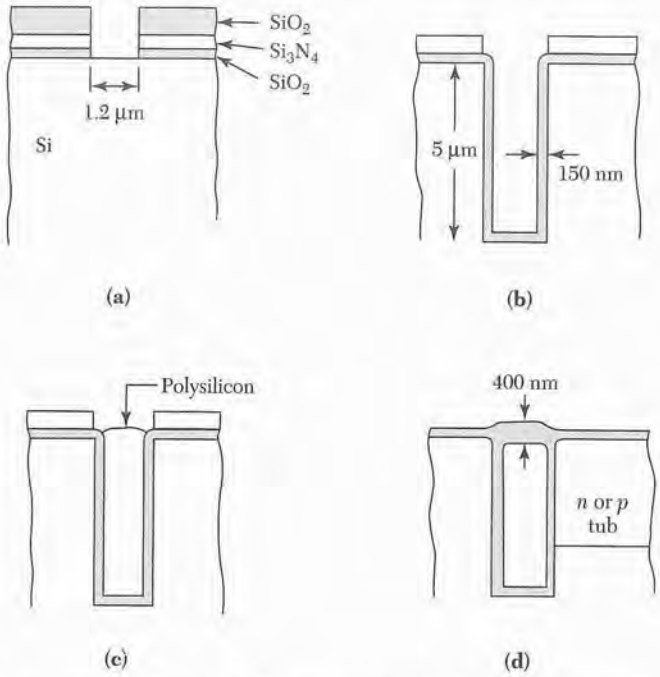


Fig. 26 Process sequence for forming a deep, narrow-trench, isolation structure. (*a*) Trench mask patterning. (*b*) Trench etching and oxide growth. (*c*) Polysilicon deposition to fill the trench. (*d*) planarization.

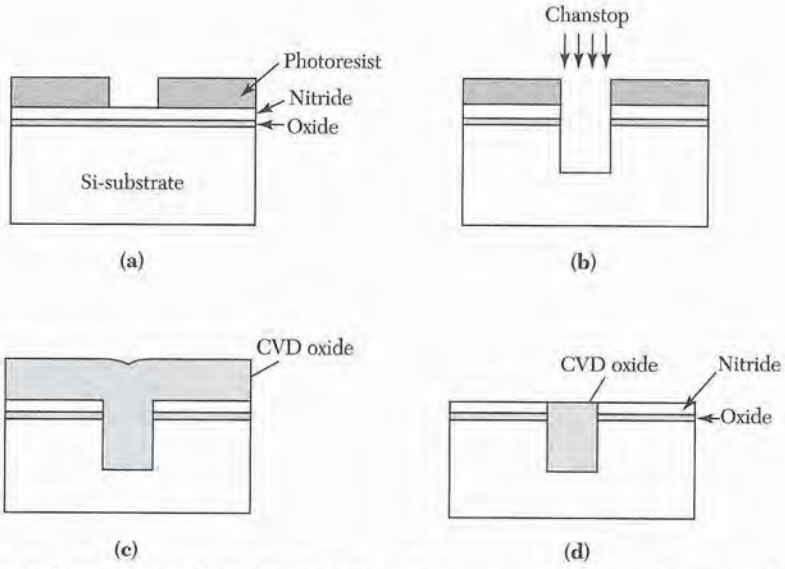


Fig. 27 A shallow trench isolation for CMOS. (*a*) Patterning with photoresist on nitride/oxide films. (*b*) Dry etching and chanstop implantation. (*c*) Chemical-vapor deposition (CVD) oxide to refill. (*d*) Planar surface after chemical-mechanical polishing (CMP).

and to get a flat surface (Fig. 27*d*). Due to its high resistance to polishing, the nitride acts as a stop-layer for the CMP process. After the polishing, the nitride layer and the oxide layer can be removed by H_3PO_4 and HF, respectively. This initial planarization step at the beginning is helpful for the subsequent polysilicon patterning and planarizations of the multilevel interconnection processes.

Gate-Engineering Technology

If we use n^+-polysilicon for both PMOS and NMOS gates, the threshold voltage for PMOS ($V_{TP} \cong -0.5$ to -1.0 V) has to be adjusted by boron implantation. This makes the channel of the PMOS a buried type, shown in Fig. 28*a*. The buried-type PMOS suffers serious short-channel effects as the device size shrinks to 0.25 μm and less. The most noticeable phenomena for short-channel effects are the V_T roll-off, drain-induced barrier lowering (DIBL), and the large leakage current at the off state so that even with the gate voltage at zero, leakage current flows through source and drain. To alleviate this problem, one can change n^+-polysilicon to p^+-polysilicon for PMOS. Due to the work function difference (there is a 1.0 eV difference from n^+- to p^+-polysilicon), one can obtain a surface p-type channel device without the boron V_T adjustment implantation. Hence, as the technology shrinks to 0.25 μm and less, dual-gate structures are required, i.e., p^+-polysilicon gate for PMOS, and n^+-polysilicon for NMOS (Fig. 28*b*). A comparison of V_T for the surface channel and the buried channel is shown in Fig. 29. We note that the V_T of surface channel rolls off slowly in the deep-submicron regime compared with the buried-channel device. This makes the surface-channel device with the p^+-polysilicon suitable for deep-submicron device operation.

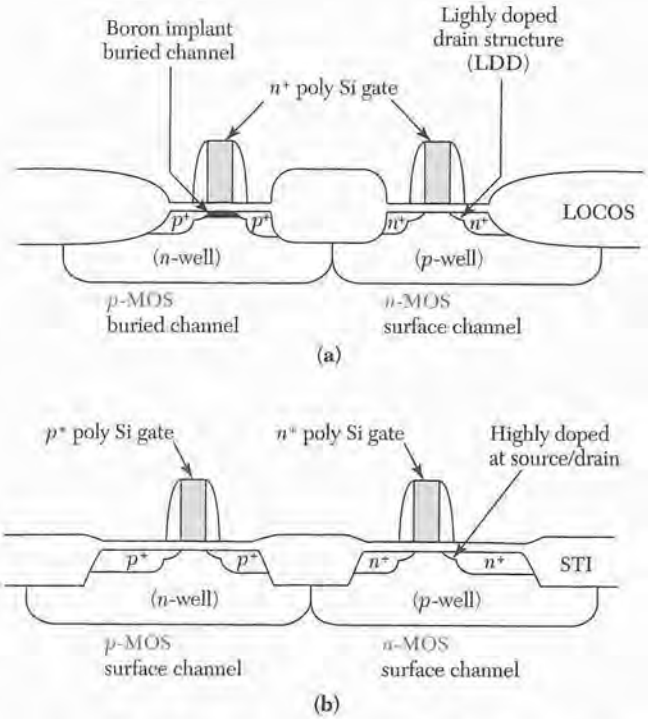


Fig. 28 (*a*) A conventional long-channel CMOS structure with a single-polysilicon gate (n^+). (*b*) Advanced CMOS structures with dual-polysilicon gates.

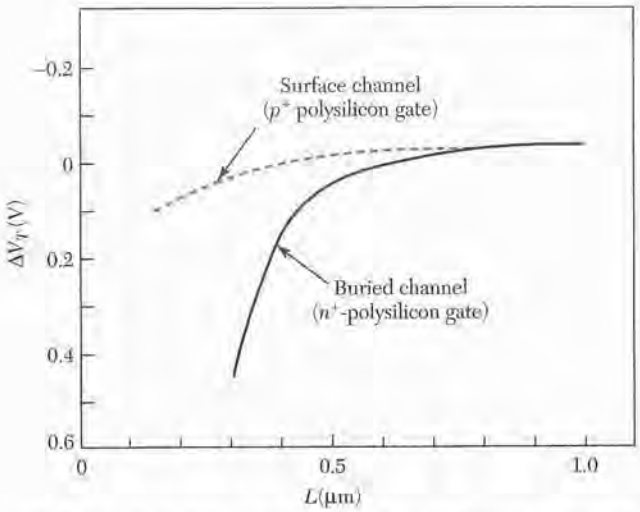


Fig. 29 The V_T roll-off for a buried type channel and for a surface type channel. The V_T drops very quickly as the channel length becomes less than 0.5 μm.

To form the p^+-polysilicon gate, ion implantation of BF_2^+ is commonly used. However, boron penetrates easily from the polysilicon through the oxide into the silicon substrate at high temperatures, resulting in a V_T shift. This penetration is enhanced in the presence of a F-atom. There are methods to reduce this effect: use of rapid thermal annealing to reduce the time at high temperatures and, consequently, the diffusion of boron; use of nitrided oxide to suppress the boron penetration, since boron can easily combine with nitrogen and becomes less mobile; and the making of a multilayer of polysilicon to trap the boron atoms at the interface of the two layers.

Figure 30 shows a microprocessor chip (Pentium 4) that has an area of about 200 mm^2 and contains 42 million components. This ULSI chip is fabricated using 0.18 μm CMOS technology with a six-level aluminum metallization.

14.3.4 BiCMOS Technology

BiCMOS is a technology that combines both CMOS and bipolar device structures in a single IC. The reason to combine these two different technologies is to create an IC chip that has the advantages of both CMOS and bipolar devices. As we know that CMOS exhibits advantages in power dissipation, noise margin, and packing density, whereas bipolar shows advantages in switching speed, current drive capability, and analog capability. As a result, for a given design rule, BiCMOS can have a higher speed than CMOS, better performance in analog circuits than CMOS, a lower power-dissipation than bipolar, and a higher component density than bipolar. Figure 31 shows the comparison of a BiCMOS and a CMOS logic gates. For a CMOS inverter, the current to drive (or to charging) the next loading, C_L, is the drain current I_{DS}. For a BiCMOS inverter, the current is $h_{fe}I_{DS}$, where h_{fe} is the current-gain of the bipolar transistor and I_{DS} is the base current of the bipolar transistor and is equal to the drain current of M_2 in the CMOS. Since h_{fe} is much larger than 1, the speed can be substantially enhanced.

BiCMOS has been widely used in many applications. In the early days, it was used in SRAM. At the present time, BiCMOS technology has been successfully developed for transceiver, amplifier, and oscillator applications in wireless-communication equipment.

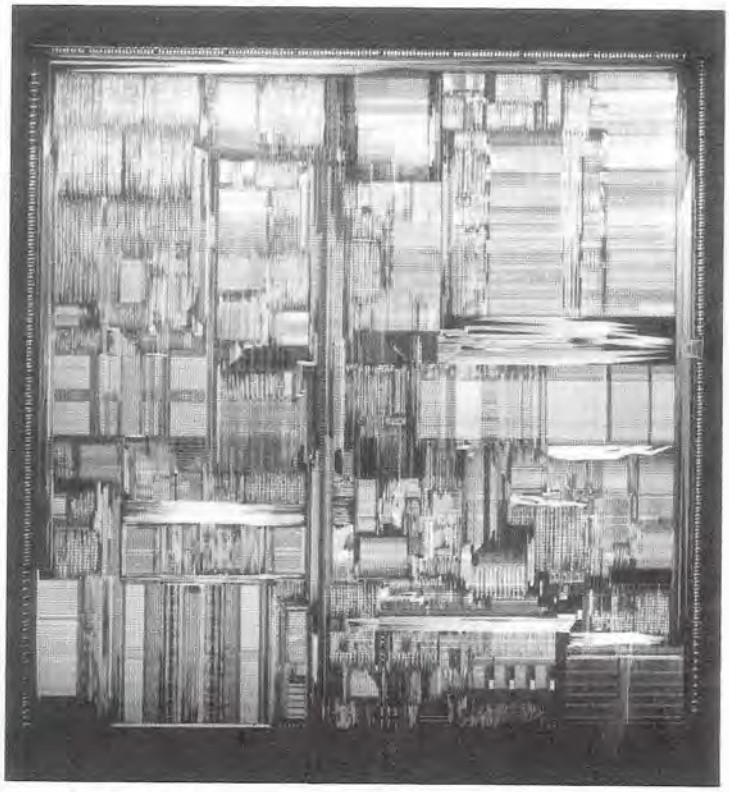


Fig. 30 Micrograph of a 32-bit microprocessor chip, Pentium 4. (Photography courtesy of Intel Corporation.)

Most of the BiCMOS processes are based on the CMOS process, with some modifications, such as adding masks for bipolar transistor fabrication. The following example is for a high-performance BiCMOS process based on the twin-well CMOS process, shown[13] in Fig. 32.

The initial material is a p-type silicon substrate, and then an n^+-buried layer is formed to reduce the collector's resistance. The buried p-layer is formed through ion implantation to increase the doping level to prevent punchthrough. A lightly doped n-epi layer is grown on the wafer and a twin-well process for the CMOS is performed. To achieve high performance of the bipolar transistor, four additional masks are needed. They are the

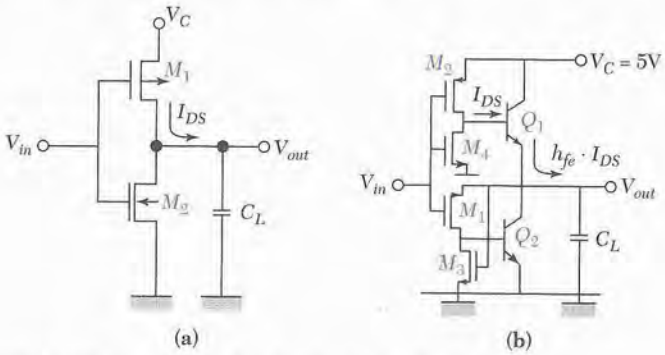


Fig. 31 (*a*) CMOS logic gate. (*b*) Bipolar CMOS (BiCMOS) logic gate.

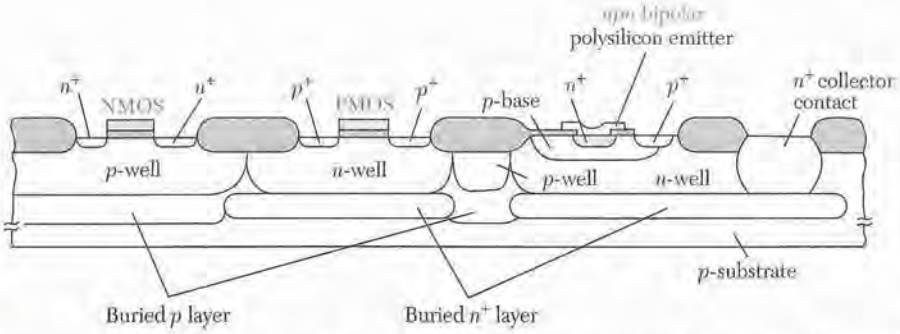


Fig. 32 Optimized BiCMOS device structure. Key features include self-aligned p and n^+ buried layers for improved packing density, separately optimized n- and p-well (twin-well CMOS) formed in an epitaxial layer with intrinsic background doping, and a polysilicon emitter for improved bipolar performance.[13]

buried n^+-mask, the collector deep-n^+-mask, the base p-mask, and the poly-emitter mask. In other processing steps, the p^+-region for base contact can be formed with the p^+-implant in the source/drain implantation of the PMOS, and the n^+-emitter can be formed with the source/drain implantation of the NMOS. The additional masks and longer processing time compared with a standard CMOS are the main drawbacks of BiCMOS. The additional cost should be justified by the enhanced performances of BiCMOS.

14.4 MESFET TECHNOLOGY

Recent advances in gallium arsenide processing techniques in conjunction with new fabrication and circuit approaches have made possible the development of "silicon-like" gallium arsenide IC technology. There are three inherent advantages of gallium arsenide compared with silicon: higher electron mobility, which results in lower series resistance for a given device geometry; higher drift velocity at a given electric field, which improves device speed; and the ability to be made semiinsulating, which can provide a lattice-matched dielectric-insulated substrate. However, gallium arsenide also has three disadvantages: a very short minority-carrier lifetime; lack of a stable, passivating native oxide; and crystal defects that are many orders of magnitude higher than in silicon. The short minority-carrier lifetime and the lack of high-quality insulating films have prevented the development of bipolar devices and delayed MOS technology using gallium arsenide. Thus, the emphasis of gallium arsenide IC technology is in the MESFET area, in which our main concerns are the majority carriers transport and the metal-semiconductor contact.

A typical fabrication sequence[14] for a high-performance MESFET is shown in Fig. 33. A layer of GaAs is epitaxially grown on a semiinsulating GaAs substrate, followed by an n^+-contact layer (Fig. 33*a*). A mesa etch step is performed for isolation (Fig. 33*b*), and a metal layer is evaporated for the source and drain ohmic contacts (Fig. 33*c*). A channel recess etch is followed by a gate recess etch and gate evaporation (Fig. 33*d* and *e*). After a liftoff process that removes the photoresist, shown in Fig. 33*e*, the MESFET is completed (Fig. 33*f*).

The n^+-contact layer reduces the source and drain ohmic contact resistances. Note that the gate is offset toward the source to minimize the source resistance. The epitaxial layer is thick enough to minimize the effect of surface depletion on the source and drain resistance. The gate electrode has maximal cross-sectional area with a minimal foot

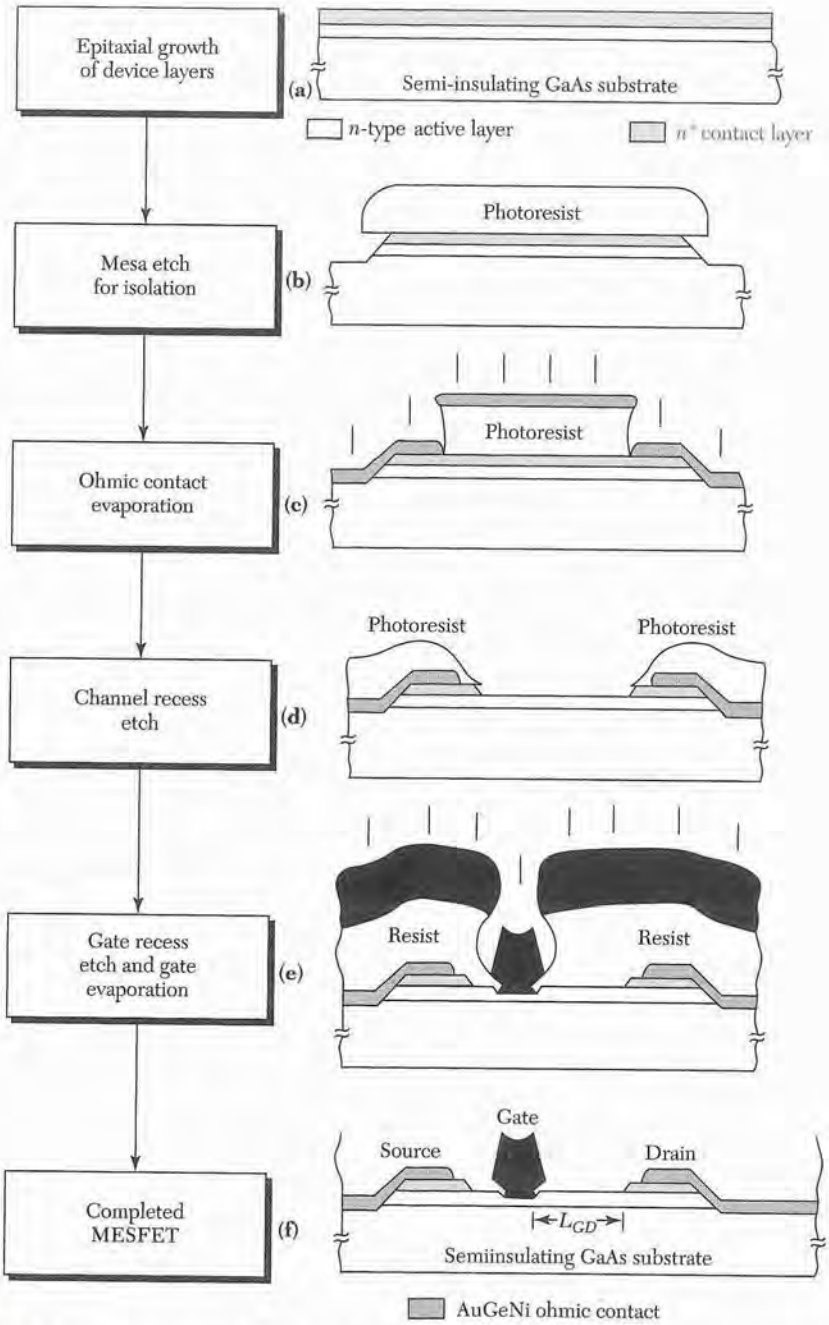


Fig. 33 Fabrication sequence of a GaAs MESFET.[14]

print, which provides low gate resistance and minimal gate length. In addition, the length L_{GD} is designed to be greater than the depletion width at gate-drain breakdown.

A representative fabrication sequence for a MESFET integrated circuit is shown [15] in Fig. 34. In this process, n^+-source and drain regions are self-aligned to the gate of each MESFET. A relatively light channel implant is used for the enhancement-mode switching

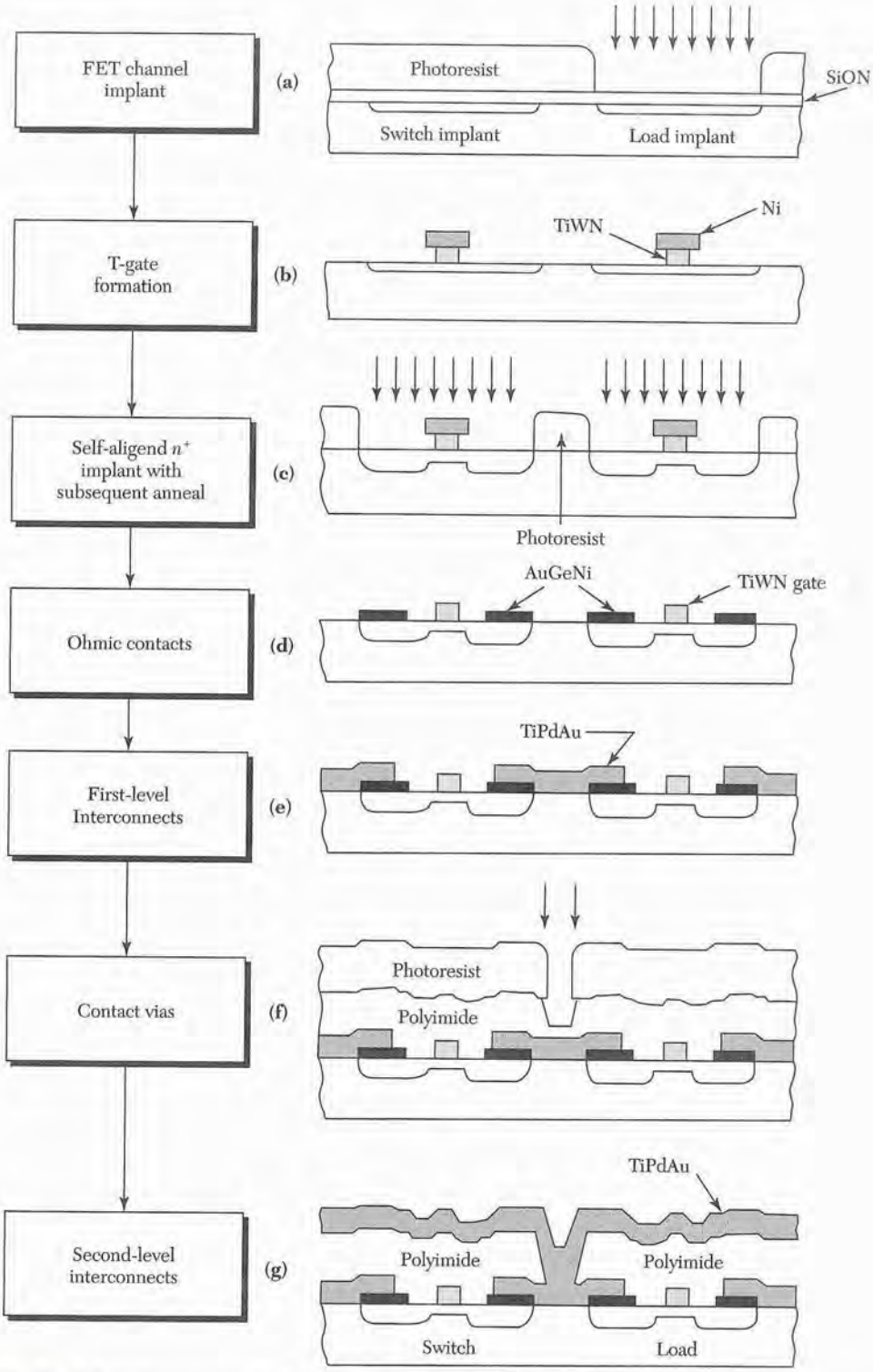


Fig. 34 Fabrication process for MESFET direct-coupled FET logic (DCFL) with active loads. Note that the n^+-source and drain regions are self-aligned to the gate.[15]

device and a heavier implant is used for the depletion-mode load device. A gate recess is usually not used for such digital IC fabrication because the uniformity of each depth has been difficult to control, leading to an unacceptable variation of the threshold voltage. This process sequence can also be used for a monolithic microwave integrated circuit (MMIC). Note that the gallium arsenide MESFET processing technology is similar to the silicon-based MOSFET processing technology.

Gallium arsenide ICs with complexities up to the large-scale integration level (~10,000 components per chip) have been fabricated. Because of the higher drift velocity (~20% higher than silicon), gallium arsenide ICs will have a 20% higher speed than silicon ICs that use the same design rules. However, substantial improvements in crystal quality and processing technology are needed before gallium arsenide can seriously challenge the preeminent position of silicon in ULSI applications.

14.5 CHALLENGES FOR MICROELECTRONICS

Since the beginning of the integrated-circuit era in 1959, the minimum device dimension, also called the minimum feature length, has been reduced at an annual rate of about 13% (i.e., a reduction of 30% every 3 years). According to the prediction by the International Technology Roadmap for Semiconductors[16], the minimum feature length will shrink from 130 nm (0.13 μm) in the year 2002 to 35 nm (0.035 μm) around 2014, as shown in Table 1. Also shown in Table 1 is the DRAM size. The DRAM has increased its memory cell capacity four times every 3 years and 64 Gbit DRAM is expected to be available in year 2011 using 50 nm design rules. The table also shows that the wafer size will increase to 450 mm (18 in. diameter) in 2014. In addition to the feature size reduction, challenges come from the device level, material level, and system level, discussed in the following subsections.

14.5.1 Challenges for Integration

Figure 35 shows the trends of power supply voltage V_{DD}, threshold voltage V_T, and gate oxide thickness d versus channel length for CMOS logic technology.[17] From the figure, one can find that the gate oxide thickness will soon approach the tunneling-current limit of 2 nm. V_{DD} scaling will slow down because of nonscalable V_T (i.e., to a minimum V_T of about 0.3 V due to subthreshold leakage and circuit noise immunity). Some challenges of the 180 nm technology and beyond are shown[18] in Fig. 36. The most stringent requirements are as follows.

TABLE 1 The Technology Generation[16] from 1997 to 2014

Year of the first product shipment	1997	1999	2002	2005	2008	2011	2014
Feature size (nm)	250	180	130	100	70	50	35
DRAM[a] size (bit)	256M	1G	—	8G	—	64G	—
Wafer size (mm)	200	300	300	300	300	300	450
Gate oxide (nm)	3–4	1.9–2.5	1.3–1.7	0.9–1.1	<1.0	—	—
Junction depth (nm)	50–100	42–70	25–43	20–33	15–30	—	—

[a]DRAM, dynamic random access memory.

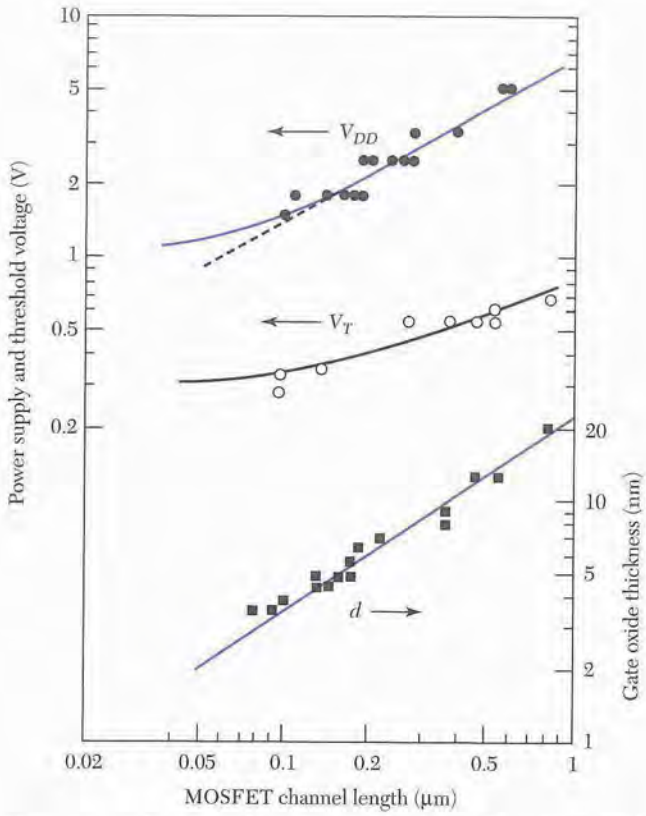


Fig. 35 Trends of power supply voltage V_{DD}, threshold voltage V_T, and gate oxide thickness d versus channel length for CMOS logic technologies. Points are collected from data published over recent years.[17]

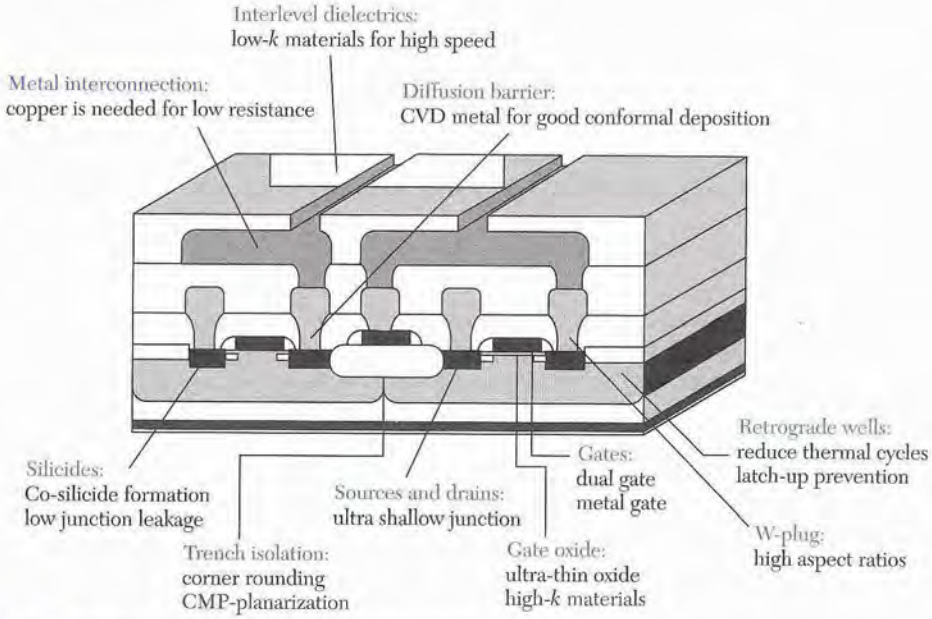


Fig. 36 Challenges for 180 nm and smaller MOSFET.[18]

Ultrashallow Junction Formation

As mentioned in Chapter 6, the short-channel effect happens as the channel length is reduced. This problem becomes critical as the device dimension is scaled down to 100 nm. To achieve an ultrashallow junction with low sheet resistance, low-energy (less than 1 keV) implantation technology with high dosage must be employed to reduce the short-channel effect. Table 1 shows the required junction depth versus the technology generation. The requirements of the junction for 100 nm are depths around 20–33 nm with a doping concentration of 1×10^{20} /cm^3.

Ultrathin Oxide

As the gate length shrinks below 130 nm, the oxide equivalent thickness of gate dielectric must be reduced around 2 nm to maintain the performance. However, if only SiO_2 (with a dielectric constant of 3.9) is used, the leakage through the gate becomes very high because of direct tunneling. For this reason, thicker high-k dielectric materials that have lower leakage current are suggested to replace oxide. Candidates for the short term are silicon nitride (with a dielectric constant of 7), Ta_2O_5 (25), and TiO_2 (60–100).

Silicide Formation

Silicide-related technology has become an integral part of submicron devices for reducing the parasitic resistance to improve device and circuit performance. The conventional Ti-silicide process has been widely use in 350–250 nm technology. However, the sheet resistance of a $TiSi_2$ line increases with decreasing line width, which limits the use of $TiSi_2$ in 100 nm CMOS applications and beyond. $CoSi_2$ or NiSi processes will replace $TiSi_2$ in the technology beyond 100 nm.

New Materials for Interconnection

To achieve high-speed operation, the RC time delay of the interconnection must be reduced. In Fig. 14 of Chapter 11 we have shown the delay as a function of feature size.[19] It is obvious that the gate delay decreases as the channel length decreases, meanwhile the delay resulting from interconnect increases significantly as the size decreases. This causes the total delay time to increase as the dimension of the device size scales down to 250 nm. Consequently, both high-conductivity metals, such as Cu, and low-dielectric-constant (low-k) insulators, such as organic (polyimide) or inorganic (F-doped oxide) materials offer major performance gains. Cu exhibits superior performance because of its high conductivity (1.7 $\mu\Omega$-cm compared with 2.7 $\mu\Omega$-cm of Al) and is 10–100 times more resistant to electromigration. The delay using the Cu and low-k material shows a significant decrease compared with that of the conventional Al and oxide. Hence, Cu with the low-k material is essential in multilevel interconnection for future deep-submicron technology.

Power Limitations

The power required merely to charge and discharge circuit nodes in an IC is proportional to the number of gates and the frequency at which they are switched (clock frequency). The power can be expressed as $P \cong 1/2CV^2 nf$, when C is the capacitance per device, V is the applied voltage, n is the number of devices per chip, and f is the clock frequency. The temperature rise caused by this power dissipation in an IC package is limited by the thermal conductivity of the package material, unless auxiliary liquid or gas cooling is used. The maximum allowable temperature rise is limited by the bandgap of the semiconductor (~100°C for Si with a bandgap of 1.1 eV). For such a temperature rise, the maximum power dissipation of a typical high-performance package is about 10

W. As a result, we must limit either the maximum clock rate or the number of gates on a chip. As an example, in an IC containing 100 nm MOS devices with $C = 5 \times 10^{-2}$ fF, running at a 20 GHz clock rate, the maximum number of gates we can have is about 10^7 if we assume a 10% duty cycle. This is a design constraint fixed by basic material parameters.

SOI Integration

Mentioned in Section 14.2.2 was the isolation of the SOI wafer. Recently SOI technology has received more attention. The advantages of the SOI integration become significant as the minimum feature length approaches 100 nm. From the process point of view, SOI does not need the complex well structure and isolation processes. In addition, shallow junctions are directly obtained through the SOI film thickness. There is no risk of nonuniform interdiffusion of silicon and Al in the contact regions because of oxide isolation at the bottom of the junction. Hence, the contact barrier is not necessary. From the device point of view, the modern bulk silicon device needs high doping at the drain and substrate to eliminate short-channel effects and punch-through. This high doping results in high capacitance when the junction is reversed bias. On the other hand, in SOI, the maximum capacitance between the junction and substrate is the capacitance of the buried insulator whose dielectric constant is three times smaller than that of silicon (3.9 versus 11.9). Based on the ring oscillator performance, the 130 nm SOI CMOS technology can achieve 25% faster speed or require 50% less power compared to a similar bulk technology.[20] SRAM, DRAM, CPU, and rf CMOS have all been successfully fabricated using SOI technology. Therefore SOI is a key candidate for the future system-on-a-chip technology, considered in the following section.

EXAMPLE 5

For an equivalent oxide thickness of 1.5 nm, what will be the physical thickness when high-k materials nitride ($\varepsilon_i/\varepsilon_0 = 7$), Ta_2O_5 (25), or TiO_2 (80) are used?

SOLUTION For nitride,

$$\left(\frac{\varepsilon_{ox}}{1.5}\right) = \left(\frac{\varepsilon_{\text{nitride}}}{d_{\text{nitride}}}\right),$$

$$d_{\text{nitride}} = 1.5\left(\frac{7}{3.9}\right) = 2.69 \text{ nm}.$$

Using the same calculation, we obtain 9.62 nm for Ta_2O_5 and 10.77 nm for TiO_2.

14.5.2 System-On-A-Chip

The increased component density and improved fabrication technology have helped the realization of the system-on-a-chip (SOC), that is, an IC chip that contain a complete electronic system. The designers can build all the circuitry needed for a complete electronic system, such as a camera, radio, television, or personal computer (PC), on a single chip. Figure 37 shows the SOC application in the PC's mother-board. Components (11 chips in this case) once found on boards are becoming virtual components on the chip at the right.[21]

There are two obstacles in the realization of the SOC. The first is the huge complexity of the design. Since the component board is presently designed by different companies and different design tools, it is difficult to integrate them into one chip. The other

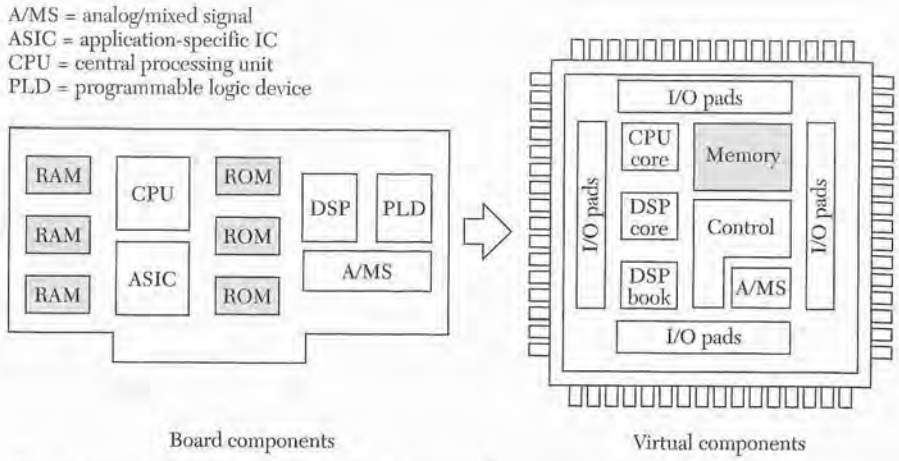


Fig. 37 System-on-a-chip of a conventional personal computer mother-board.[21]

is the difficulty of fabrication. In general, the fabricating processes of the DRAM are significantly different from those of logic IC (e.g., CPU). Speed is the first priority for the logic, whereas leakage of the stored charge is the priority for memory. Therefore, multilevel interconnection using five to six levels of metals is essential for logic IC to improve the speed. However, DRAM needs only two to three levels. In addition, to increase the speed, a silicide process must be used to reduce the series resistance, and ultrathin gate oxide is needed to increase the drive current. These requirements are not critical for the memory.

To achieve the SOC goal, an embedded DRAM technology is introduced, i.e., to merge logic and DRAM into a single chip with compatible processes. Figure 38 shows the schematic cross section of the embedded DRAM, including the DRAM cells and the logic CMOS devices.[22] Some processing steps are modified as a compromise. The trench-type capaci-

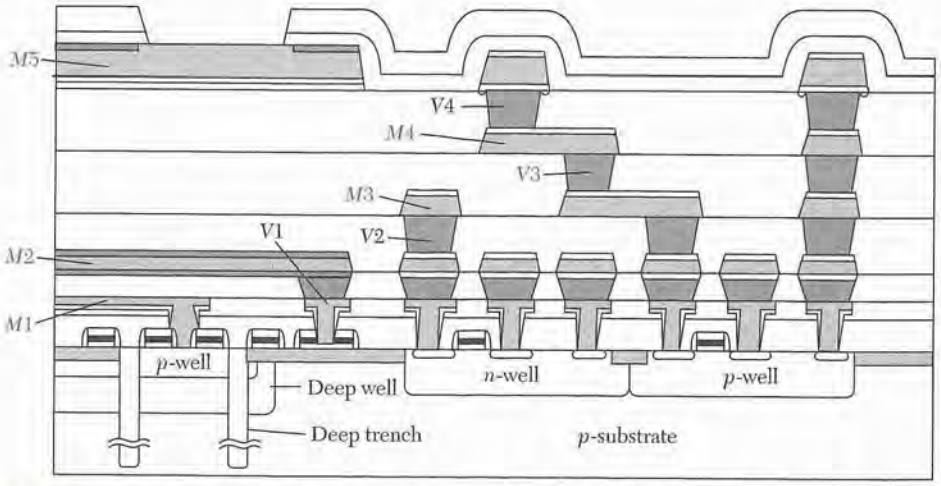


Fig. 38 Schematic cross section of the embedded DRAM including DRAM cells and logic MOSFETs. There is no height difference in the trench capacitor cell because of the DRAM cell structure. $M1$ to $M5$ are metal interconnections, and $V1$ to $V4$ are via holes.[22]

tor, instead of the stacked type, is used so that there is no height difference in the DRAM cell structure. In addition, multiple gate oxide thicknesses exist on the same wafer to accommodate multiple supply voltages and/or combine memory and logic circuits on one chip.

SUMMARY

In this chapter we considered processing technologies for passive components, active devices, and IC. Three major IC technologies based on the bipolar transistor, the MOSFET, and the MESFET were discussed in detail. It appears that the MOSFET will be the dominant technology at least until 2014 because of its superior performance compared with the bipolar transistor. For 100 nm CMOS technology, a good candidate is the combination of an SOI-substrate with interconnections using Cu and low-k materials.

Because the rapid reduction in feature length, the technology will soon reach its practical limit as the channel length is reduced to about 20 nm. What will be the device beyond the CMOS is the question being asked by research scientists. Major candidates include many innovative devices based on quantum mechanical effects. This is because when the lateral dimension is reduced to below 100 nm, depending on the materials and the temperature of operation, electronic structures will exhibit nonclassical behaviors. The operation of such devices will be on the scale of single-electron transport. This approach has been demonstrated by the single-electron memory cell. The realization of such systems with trillions of components will be a major challenge beyond CMOS.[23]

REFERENCES

1. For a detailed discussion on IC process integration, see C. Y. Liu and W. Y. Lee, "Process Integration," in C. Y. Chang and S. M. Sze, Eds., *ULSI Technology*, McGraw-Hill, New York, 1996.
2. T. Tachikawa, "Assembly and Packaging," in C. Y. Chang and S. M. Sze, Eds., *ULSI Technology*, McGraw-Hill, New York, 1996.
3. T. H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*, Cambridge Univ. Press, Cambridge, U.K., 1998, Ch. 2.
4. D. Rise, "Isoplanar-S Scales Down for New Heights in Performance," *Electronics*, **53**, 137 (1979).
5. T. C. Chen, et. al., "A submicrometer High-Performance Bipolar Technology," *IEEE Electron. Device Lett.*, **10**(8), 364, (1989).
6. G. P. Li et. al., "An Advanced High-Performance Trench-Isolated Self-Aligned Bipolar Technology," *IEEE Trans. Electron Devices*, **34**(10), 2246 (1987).
7. W. E. Beasle, J. C. C. Tsai, and R. D. Plummer, Eds., *Quick Reference Manual for Semiconductor Engineering*, Wiley, New York, 1985.
8. R. W. Hunt, "Memory Design and Technology," in M. J. Howes and D. V. Morgan, Eds., *Large Scale Integration*, Wiley, New York, 1981.
9. A. K. Sharma, *Semiconductor Memories—Technology, Testing, and Reliability*, IEEE, New York, 1997.
10. U. Hamann, "Chip Cards—The Application Revolution," *IEEE Tech. Dig. Int. Electron Devices Meet.*, p. 15, 1997.
11. R. D. Rung, H. Momose, and Y. Nagakubo, "Deep Trench Isolation CMOS Devices," *IEEE Tech. Dig. Int. Electron. Devices Meet.*, p. 237, 1982.
12. D. M. Bron, M. Ghezzo, and J. M. Primbley, "Trends in Advanced CMOS Process Technology," *Proc. IEEE*, p. 1646, (1986).
13. H. Higuchi, et al., "Performance and Structure of Scaled-Down Bipolar Devices Merge with CMOSFETs," *IEEE Tech.Dig. Int. Electron. Devices Meet.*, 694, 1984.
14. M. A. Hollis and R. A. Murphy, "Homogeneous Field-Effect Transistors," in S. M. Sze, Ed., *High-Speed Semiconductor Devices*, Wiley, New York, 1990.

15. H. P. Singh, et al., "GaAs Low Power Integrated Circuits for a High Speed Digital Signal Processor," *IEEE Trans. Electron Devices*, **36**, 240 (1989).

16. *International Technology Roadmap for Semiconductor (ITRS)*, Semiconductor Ind. Assoc., San Jose, 1999.

17. Y. Taur and E. J. Nowak, "CMOS Devices below 0.1 μm: How High Will Performance Go?" *IEEE Tech. Dig. Int. Electron Devices Meet.*, 215, 1997.

18. L. Peters, "Is the 0.18 μm Node Just a Roadside Attraction," *Semicond. Int.*, **22**, 46 (1999).

19. M. T. Bohr, "Interconnect Scaling—The Real Limiter to High Performance ULSI," *IEEE Tech. Dig. Int. Electron Devices Meet.*, p. 241, 1995.

20. E. Leobandung, et al.,"Scalability of SOI Technology into 0.13 μm 1.2 V CMOS Generation," *IEEE Int. Electron Devices Meet.*, p. 403, 1998.

21. B. Martin, "Electronic Design Automation," *IEEE Spectr.*, **36**, 61 (1999).

22. H. Ishiuchi, et al., " Embedded DRAM Technologies," *IEEE Tech. Dig. Int. Electron Devices Meet.*, p. 33, 1997.

23. S. Luryi, J. Xu, and A. Zaslavsky, Eds, *Future Trends in Microelectronics*, Wiley, New York, 1999.

PROBLEMS (* DENOTES DIFFICULT PROBLEMS)

FOR SECTION 14.1 PASSIVE COMPONENTS

1. For a sheet resistance of 1 kΩ/□, find the maximum resistance that can be fabricated on a 2.5 × 2.5-mm chip using 2 μm lines with a 4 μm pitch (i.e., distance between the centers of the parallel lines).

2. Design a mask set for a 5 pF MOS capacitor. The oxide thickness is 30 nm. Assume that the minimum window size is 2 × 10 μm and the maximum registration errors are 2 μm.

3. Draw a complete step-by-step set of masks for the spiral inductor with three turns on a substrate.

4. Design a 10 nH square spiral inductor in which the total length of the interconnect is 350 μm; the spacing between turns is 2 μm.

FOR SECTION 14.2 BIPOLAR TECHNOLOGY

5. Draw the circuit diagram and device cross section of a clamped transistor.

6. Identify the purpose of the following steps in self-aligned double-polysilicon bipolar structure: (a) undoped polysilicon in trench in Fig. 13*a*, (b) the poly 1 in Fig. 13*b*, and (c) the poly 2 in Fig. 13*d*.

FOR SECTION 14.3 MOSFET TECHNOLOGY

***7.** In NMOS processing, the starting material is a *p*-type 10 Ω-cm <100>-oriented silicon wafer. The source and drain are formed by arsenic implantation of 10^{16} ions/cm^2 at 30 keV through a gate oxide of 25 nm. (a) Estimate the threshold voltage change of the device. (b) Draw the doping profile along a coordinate perpendicular to the surface and passing through the channel region or the source region.

8. (a) Why is <100>-orientation preferred in NMOS fabrication? (b) What are the disadvantages if too thin a field oxide is used in NMOS devices? (c) What problems occur if a polysilicon gate is used for gate lengths less than 3 μm? Can another material be substituted for polysilicon? (d) How is a self-aligned gate obtained and what are its advantages? (e) What purpose does P-glass serve?

***9.** For a floating-gate nonvolatile memory, the lower insulator has a dielectric constant of 4 and is 10 nm thick. The insulator above the floating gate has a dielectric constant of 10 and is 100 nm thick. If the current density J in the lower insulators is given by $J = \sigma\mathscr{E}$,

where $\sigma = 10^{-7}$ S/cm, and the current in the other insulator is negligibly small, find the threshold voltage shift of the device caused by a voltage of 10 V applied to the control gate for (a) 0.25 μs, and (b) a sufficiently long time that J in the lower insulator becomes negligibly small.

10. Draw a complete step-by step set of masks for CMOS inverter shown in Fig. 23. Pay particular attention to the cross section shown in Fig. 23*c* for your scale.

***11.** A 0.5 μm digital CMOS technology has 5 μm wide transistors. The minimum wire width is 1 μm and the metallization layer consists of 1 μm thick aluminum. Assume that μ_n is 400 cm²/V-s, d is 10-nm, V_{DD} is 3.3 V, and the threshold voltage is 0.6 V. Finally, assume that the maximum voltage drop that can be tolerated is 0.1 V when a 1 μm² cross section aluminum wire is carrying the maximum current that can be supplied by the NMOS transistor. How long a wire can be allowed? Use a simple square-law, long-channel model to predict the MOS current drive (resistivity of aluminum is 2.7×10^{-8} Ω-cm).

12. Plot the cross-sectional views of a twin-tub CMOS structure of the following stages of processing: (a) n-tub implant, (b) p-tub implant, (c) twin-tub drive-in, (d) nonselective p^+-source/drain implant, (e) selective n^+-source/drain implant using photoresist as mask, and (f) P-glass deposition.

13. Why do we use a p^+-polysilicon gate for PMOS?

14. What is the boron penetration problem in p^+-polysilicon PMOS? How would you eliminate it?

15. To obtain a good interfacial property, a buffered layer is usually deposited between the high-k material and substrate. Calculate the effective oxide thickness if the stacked gate dielectric structure is (a) a buffered nitride of 0.5 nm and (b) a Ta_2O_5 of 10 nm.

16. Describe the disadvantages of LOCOS technology and the advantages of shallow-trench isolation technology.

FOR SECTION 14.4 MESFET TECHNOLOGY

17. What is the purpose for the polyimide used in Fig. 34*f*?

18. What is the reason that it is difficult to make bipolar transistor and MOSFET in GaAs?

FOR SECTION 14.5 CHALLENGES FOR MICROELECTRONICS

19. (a) Calculate the RC time constant of a aluminum runner 0.5 μm thick formed on a thermally grown SiO_2 0.5 μm thick. The length and width of the runner are 1 cm and 1 μm, respectively. The resistivity of the runner is 10^{-5} Ω-cm. (b) What will be the RC time constant for a polysilicon runner ($R_\square = 30$ Ω/□) of identical dimension?

20. Why do we need multiple oxide thicknesses for a system-on-a-chip (SOC)?

21. Normally we need a buffered layer placed between a high-k Ta_2O_5 and the silicon substrate. Calculate the effective oxide thickness (EOT) when the stacked gate dielectric is Ta_2O_5 ($k = 25$) with a thickness of 75Å on a buffered nitride layer ($k = 7$ and a thickness of 10 Å). Also calculate EOT for a buffered oxide layer ($k = 3.9$, and a thickness of 5 Å).

Appendix A

List of Symbols

Symbol	Description	Unit
a	Lattice constant	Å
$\mathscr{B}$	Magnetic induction	Wb/m^2
c	Speed of light in vacuum	cm/s
C	Capacitance	F
$\mathscr{D}$	Electric displacement	C/cm^2
D	Diffusion coefficient	cm^2/s
E	Energy	eV
E_C	Bottom of conduction band	eV
E_F	Fermi energy level	eV
E_g	Energy bandgap	eV
E_V	Top of valence band	eV
$\mathscr{E}$	Electric field	V/cm
$\mathscr{E}_c$	Critical field	V/cm
$\mathscr{E}_m$	Maximum field	V/cm
f	Frequency	Hz(cps)
$F(E)$	Fermi–Dirac distribution function	
h	Planck constant	J·s
$h\nu$	Photon energy	eV
I	Current	A
I_C	Collector current	A
J	Current density	A/cm^2
J_{th}	Threshold current density	A/cm^2
k	Boltzmann constant	J/K
kT	Thermal energy	eV
L	Length	cm or μm
m_0	Electron rest mass	kg
m_n	Electron effective mass	kg
m_p	Hole effective mass	kg
$\overline{n}$	Refractive index	
n	Density of free electrons	cm^{-3}
n_i	Intrinsic carrier concentration	cm^{-3}

(continued)

Symbol	Description	Unit
N	Doping concentration	cm^{-3}
N_A	Acceptor-impurity density	cm^{-3}
N_C	Effective density of states in conduction band	cm^{-3}
N_D	Donor-impurity density	cm^{-3}
N_V	Effective density of states in valence band	cm^{-3}
p	Density of free holes	cm^{-3}
P	Pressure	Pa
q	Magnitude of electronic charge	C
Q_{it}	Interface-trapped charge	charges/cm^2
R	Resistance	Ω
t	Time	s
T	Absolute Temperature	K
v	Carrier velocity	cm/s
v_s	Saturation velocity	cm/s
v_{th}	Thermal velocity	cm/s
V	Voltage	V
V_{bi}	Built-in potential	V
V_{EB}	Emitter-base voltage	V
V_B	Breakdown voltage	V
W	Thickness	cm or μm
W_B	Base thickness	cm or μm
ε_0	Permittivity in vacuum	F/cm
ε_s	Semiconductor permittivity	F/cm
ε_{ox}	Insulator permittivity	F/cm
$\varepsilon_s/\varepsilon_0$ or $\varepsilon_{ox}/\varepsilon_0$	Dielectric constant	
τ	Lifetime or decay time	s
θ	Angle	rad
λ	Wavelength	μm or nm
ν	Frequency of light	Hz
μ_0	Permeability in vacuum	H/cm
μ_n	Electron mobility	cm^2/V·s
μ_p	Hole mobility	cm^2/V·s
ρ	Resistivity	Ω-cm
ϕ_{Bn}	Schottkey barrier height on n-type semiconductor	V
ϕ_{Bp}	Schottkey barrier height on p-type semiconductor	V
$q\phi_m$	Metal work function	eV
ω	Angular frequency ($2\pi f$ or $2\pi\nu$)	Hz
Ω	Ohm	Ω

Appendix B

International System of Units (SI Units)

Quantity	Unit	Symbol	Dimensions
Length[§]	meter	m	
Mass	kilogram	kg	
Time	second	s	
Temperature	kelvin	K	
Current	ampere	A	
Light intensity	candela	Cd	
Angle	radian	rad	
Frequency	hertz	Hz	1/s
Force	newton	N	kg-m/s^2
Pressure	pascal	Pa	N/m^2
Energy[§]	joule	J	N-m = kg·m^2/s^2
Power	watt	W	J/s
Electric charge	coulomb	C	A·s
Potential	volt	V	J/C
Conductance	siemens	S	A/V
Resistance	ohm	Ω	V/A
Capacitance	farad	F	C/V
Magnetic flux	weber	Wb	V·s
Magnetic induction	tesla	T	Wb/m^2
Inductance	henry	H	Wb/A
Light flux	lumen	Lm	Cd-rad

[§] It is more common in the semiconductor field to use cm for length and eV for energy (1 cm = 10^{-2} m, 1 eV = 1.6×10^{-19} J).

Appendix C

Unit Prefixes*

Multiple	Prefix	Symbol
10^{18}	exa	E
10^{15}	peta	P
10^{12}	tera	T
10^{9}	giga	G
10^{6}	mega	M
10^{3}	kilo	k
10^{2}	hecto	h
10	deka	da
10^{-1}	deci	d
10^{-2}	centi	c
10^{-3}	milli	m
10^{-6}	micro	μ
10^{-9}	nano	n
10^{-12}	pico	p
10^{-15}	femto	f
10^{-18}	atto	a

* Adopted by International Committee on Weights and Measures. (Compound prefixes should not be used, e.g., not μμ but p.)

Appendix D

Greek Alphabet

Letter	Lowercase	Uppercase
Alpha	α	Α
Beta	β	Β
Gamma	γ	Γ
Delta	δ	Δ
Epsilon	ε	Ε
Zeta	ζ	Ζ
Eta	η	Η
Theta	θ	Θ
Iota	ι	Ι
Kappa	κ	Κ
Lambda	λ	Λ
Mu	μ	Μ
Nu	ν	Ν
Xi	ξ	Ξ
Omicron	o	Ο
Pi	π	Π
Rho	ρ	Ρ
Sigma	σ	Σ
Tau	τ	Τ
Upsilon	υ	Υ
Phi	ϕ	Φ
Chi	χ	Χ
Psi	ψ	Ψ
Omega	ω	Ω

Appendix E

Physical Constants

Quantity	Symbol	Value
Angstrom unit	Å	10 Å = 1 nm = 10^{-3} μm = 10^{-7} cm = 10^{-9} m
Avogadro constant	N_{av}	6.02214×10^{23}
Bohr radius	a_B	0.52917 Å
Boltzmann constant	k	1.38066×10^{-23} J/K (R/N_{av})
Elementary charge	q	1.60218×10^{-19} C
Electron rest mass	m_0	0.91094×10^{-30} kg
Electron volt	eV	1 eV = 1.60218×10^{-19} J = 23.053 kcal/mol
Gas constant	R	1.98719 cal/mol-K
Permeability in vacuum	μ_0	1.25664×10^{-8} H/cm ($4\pi \times 10^{-9}$)
Permittivity in vacuum	ε_0	8.85418×10^{-14} F/cm ($1/\mu_0 c^2$)
Planck constant	h	6.62607×10^{-34} J·s
Reduced Planck constant	$\hbar$	1.05457×10^{-34} J·s ($h/2\pi$)
Proton rest mass	M_p	1.67262×10^{-27} kg
Speed of light in vacuum	c	2.99792×10^{10} cm/s
Standard atmophere		1.01325×10^5 Pa
Thermal voltage at 300 K	kT/q	0.025852 V
Wavelength of 1 eV quantum	λ	1.23984 μm

$1.3806503 \times 10^{-23}\ \underbrace{m^2\ kg\ s^{-2}\ K^{-1}}_{J \cdot K^{-1}}$

$8.61732 \times 10^{-5}\ eV \cdot K^{-1}$

Appendix F

Properties of Important Element and Binary Compound Semiconductors at 300 K

Semiconductor		Lattice constant (Å)	Bandgap (eV)	Band[a]	Mobility[b] (cm^2/V-s) μ_n	μ_p	Dielectric constant
Element	Ge	5.65	0.66	I	3900	1800	16.2
	Si	5.43	1.12	I	1450	505	11.9
IV-IV	SiC	3.08	2.86	I	300	40	9.66
III-V	AlSb	6.13	1.61	I	200	400	12.0
	GaAs	5.65	1.42	D	9200	320	12.4
	GaP	5.45	2.27	I	160	135	11.1
	GaSb	6.09	0.75	D	3750	680	15.7
	InAs	6.05	0.35	D	33000	450	15.1
	InP	5.86	1.34	D	5900	150	12.6
	InSb	6.47	0.17	D	77000	850	16.8
II-VI	CdS	5.83	2.42	D	340	50	5.4
	CdTe	6.48	1.56	D	1050	100	10.2
	ZnO	4.58	3.35	D	200	180	9.0
	ZnS	5.42	3.68	D	180	10	8.9
IV-VI	PbS	5.93	0.41	I	800	1000	17.0
	PbTe	6.46	0.31	I	6000	4000	30.0

[a]I, indirect, D, direct.

[b]The values are for drift mobilities obtained in the purest and most perfect materials available to date.

Appendix G

Properties of Si and GaAs at 300 K

Properties	Si	GaAs
Atoms/cm^3	5.02×10^{22}	4.42×10^{22}
Atomic weight	28.09	144.63
Breakdown field (V/cm)	$\sim 3 \times 10^5$	$\sim 4 \times 10^5$
Crystal structure	Diamond	Zincblende
Density (g/cm^3)	2.329	5.317
Dielectric constant	11.9	12.4
Effective density of states in conduction band, N_C(cm^{-3})	2.86×10^{19}	4.7×10^{17}
Effective density of states in valence band, N_V(cm^{-3})	2.66×10^{19}	7.0×10^{18}
Effective mass (conductivity)		
Electrons (m_n/m_0)	0.26	0.063
Holes (m_p/m_0)	0.69	0.57
Electron affinity, χ(V)	4.05	4.07
Energy gap (eV)	1.12	1.42
Index of refraction	3.42	3.3
Intrinsic carrier concentration(cm^{-3})	9.65×10^9	2.25×10^6
Intrinsic resistivity (Ω-cm)	3.3×10^5	2.9×10^8
Lattice constant (Å)	5.43102	5.65325
Linear coefficient of thermal expansion, $\Delta L/L \times T$ (°C^{-1})	2.59×10^{-6}	5.75×10^{-6}
Melting point (°C)	1412	1240
Minority-carrier lifetime (s)	3×10^{-2}	$\sim 10^{-8}$
Mobility (cm^2/V·s)		
μ_n (electrons)	1450	9200
μ_p (holes)	505	320
Specific heat (J/g -°C)	0.7	0.35
Thermal conductivity(W/cm-K)	1.31	0.46
Vapor pressure (Pa)	1 at 1650°C 10^{-6} at 900°C	100 at 1050°C 1 at 900°C

Appendix H

Derivation of the Density of States in a Semiconductor

To calculate the electron and hole concentrations in the conduction and valence bands, respectively, we need to know the density of states, that is, the number of allowed energy states per unit energy per unit volume (i.,e., in the units of number of states /eV/cm^3).

When electrons move back and forth along the x-direction in a semiconductor material, the movements can be described by standing-wave oscillations. The wavelength λ of a standing wave is related to the length of the semiconductor L by

$$\frac{L}{\lambda} = n_x, \tag{1}$$

where n_x is an integer. The wavelength can be expressed by de Broglie hypothesis:

$$\lambda = \frac{h}{p_x}, \tag{2}$$

where h is the Planck constant and p_x is the momentum in the x-direction. Substituting Eq. 2 into Eq. 1 gives

$$Lp_x = hn_x. \tag{3}$$

The incremental momentum dp_x required for a unity increase in n_x is

$$L\,dp_x = h. \tag{4}$$

For a three-dimensional cube of side L, we have

$$L^3\,dp_x\,dp_y\,dp_z = h^3. \tag{5}$$

The volume $dp_x\,dp_y\,dp_z$ in the momentum space for a unit cube ($L = 1$) is thus equal to h^3. Each incremental change in n corresponds to a unique set of integers (n_x, n_y, n_z), which in turn corresponds to an allowed energy state. Thus, the volume in momentum space for an energy state is h^3. The figure shows the momentum space in spherical coordinates. The volume between two concentric spheres (from p to $p + dp$) is $4\pi p^2 dp$. The number of energy states contained in the volume is then $2(4\pi p^2 dp)/h^3$, where the factor 2 accounts for the electron spins.

The energy E of the electron (here we only consider the kinetic energy) is given by

$$E = \frac{p^2}{2m_n} \tag{6}$$

or

$$p = \sqrt{2m_n E}, \tag{7}$$

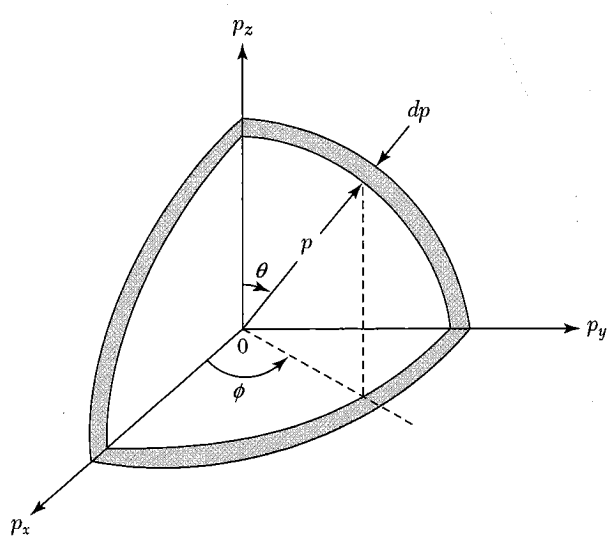
where p is the total momentum (with components p_x, p_y, and p_z in Cartesian Coordinates) and m_n is the effective mass. From Eq. 7, we can substitute E for p and obtain

$$N(E)dE = \frac{8\pi p^2 dp}{h^3} = 4\pi\left(\frac{2m_n}{h^2}\right)^{\frac{3}{2}} E^{\frac{1}{2}}\, dE \tag{8}$$

and

$$N(E) = 4\pi\left(\frac{2m_n}{h^2}\right)^{\frac{3}{2}} E^{\frac{1}{2}}, \tag{9}$$

where $N(E)$ is called the density of states.



Appendix I

Derivation of Recombination Rate for Indirect Recombination

A schematic diagram showing the various transitions that occur in the recombination through recombination centers is shown in Fig. 12 of Chapter 3. If the concentration of centers in the semiconductor is N_t, the concentration of unoccupied centers is given by $N_t(1\text{-}F)$, where F is the Fermi distribution function for the probability that a center is occupied by an electron. In equilibrium,

$$F = \frac{1}{1+e^{(E_t - E_F)/kT}} \tag{1}$$

where E_t is the energy level of the center and E_F is the Fermi level.

Therefore, the capture rate of an electron by a recombination center, Fig. 12*a* of Chapter 3, is given by

$$R_a \approx nN_t(1-F). \tag{2}$$

We designate the proportionality constant by the product $v_{th}\sigma_n$, so that

$$R_a = v_{th}\sigma_n nN_t(1-F). \tag{3}$$

The product $v_{th}\sigma_n$ may be visualized as the volume swept out per unit time by an electron with cross section σ_n. If the center lies within this volume, the electron will be captured by it.

The rate of emission of electrons from the center, Fig. 12*b*, is the inverse of the electron capture process. The rate is proportional to the concentration of centers occupied by electrons, that is, N_tF. We have

$$R_b = e_n N_t F. \tag{4}$$

The proportionality constant e_n is called the emission probability. At thermal equilibrium the rates of capture and emission of electrons must be equal ($R_a = R_b$). Thus, the emission probability can be expressed in terms of the quantities already defined in Eq. 3:

$$e_n = \frac{v_{th}\sigma_n n(1-F)}{F}. \tag{5}$$

Since the electron concentration in thermal equilibrium is given by

$$n = n_i e^{(E_F - E_i)/kT}, \tag{6}$$

we obtain

$$e_n = v_{th}\sigma_n n_i e^{(E_t - E_i)/kT}. \tag{7}$$

The transitions between the recombination center and valence band are analogous to those described above. The capture rate of a hole by an occupied recombination center, Fig. 12*c*, is given by

$$R_c = v_{th}\sigma_p p N_t F. \tag{8}$$

By arguments similar to those for electron emission, the rate of hole emission, Fig. 12*d*, is

$$R_d = e_p N_t (1 - F). \tag{9}$$

The emission probability e_p of a hole may be expressed in terms of v_{th} and σ_p by considering the thermal equilibrium condition for which $R_c = R_d$:

$$e_p = v_{th}\sigma_p n_i e^{(E_i - E_t)/kT}. \tag{10}$$

Let us now consider the nonequilibrium case in which an n-type semiconductor is illuminated uniformly to give a generation rate G_L. Thus in addition to the process shown in Fig. 12, electron-hole pairs are generated as a result of light. In steady state the electrons entering and leaving the conduction band must be equal. This is called the *principle of detailed balance*, and it yields

$$\frac{dn_n}{dt} = G_L - (R_a - R_b) = 0. \tag{11}$$

Similarly, in steady state the detailed balance of holes in valence band leads to

$$\frac{dp_n}{dt} = G_L - (R_c - R_d) = 0. \tag{12}$$

Under equilibrium conditions, that is, $G_L = 0$, $R_a = R_b$ and $R_c = R_d$. However, under state-state nonequilibrium conditions, $R_a \neq R_b$ and $R_c \neq R_d$. From Eqs. 11 and 12 we obtain

$$G_L = R_a - R_b = R_c - R_d \equiv U. \tag{13}$$

We can get the net recombination rate U from Eqs. 3, 4, 8, and 9:

$$U \equiv R_a - R_b = \frac{v_{th}\sigma_n\sigma_p N_t\left(p_n n_n - n_i^2\right)}{\sigma_p[p_n + n_i e^{(E_i - E_t)/kT}] + \sigma_n[n_n + n_i e^{(E_t - E_i)/kT}]}. \tag{14}$$

Appendix J

Calculation of the Transmission Coefficient for a Symmetric Resonant-Tunneling Diode

To calculate the transmission coefficient, we consider Fig. 13*a* of Chapter 8 where the five regions (I, II, III, IV, V) are specified by the coordinates (x_1, x_2, x_3, x_4). The Schrödinger equation for the electron in any region can be written as

$$-\frac{\hbar^2}{2m_i^*}\left(\frac{d^2\psi_i}{dx^2}\right)+V_i\psi_i = E\psi_i \qquad i=1,2,3,4,5, \tag{1}$$

where $\hbar$ is the reduced Planck constant, m_i^* the effective mass in the ith region, E the incident energy, and V_i and ψ_i the potential energy and the wave function in the ith region, respectively. The wavefunction ψ_i can be expressed as

$$\psi_i(x) = A_i \exp(jk_i x) + B_i \exp(-jk_i x), \tag{2}$$

where A_i and B_i are constants to be determined from the boundary conditions, and $k_i = \sqrt{2m_i^*(E-V_i)}/\hbar$ Since the wavefunctions and their first derivatives (i.e., $\psi_i / m_i^* = \psi_{i+1} / m_{i+1}^*$) at each potential discontinuity must be continuous, we obtain the transmission coefficient (for identical effective mass across the five regions),

$$T_t = \frac{1}{1+E_0^2(\sinh^2\beta L_B)H^2/[4E^2(E_0-E)^2]}, \tag{3}$$

where

$$H \equiv 2\left[E(E_0-E)\right]^{1/2}\cosh\beta L_B \cos kL_W - (2E-E_0)\sinh\beta L_B \sin kL_W$$

and

$$\beta \equiv \frac{\sqrt{2m^*(E_0-E)}}{\hbar}, \qquad k = \frac{\sqrt{2m^*E}}{\hbar}.$$

The resonant condition occurs when $H = 0$, and thus $T_t = 1$. The resonant-tunneling energy levels E_n can be calculated by solving the transcendental equation:

$$\frac{2\left[E(E_0 - E)\right]^{1/2}}{(2E - E_0)} = \tan kL_W \tanh \beta L_B. \tag{4}$$

As a first-order estimate of the energy levels, one can use the results of a quantum well with infinite barrier height:

$$E_n \approx \left(\frac{\pi^2 \hbar^2}{2m^* L_W^2}\right) n^2. \tag{5}$$

For a double-barrier structure with finite barrier height and width, the energy level (for a given n) will be lower; however, it will have a similar dependence on the effective mass and well width, that is, E_n increases with decreasing m^* or L_W.

Appendix K

Basic Kinetic Theory of Gases

The ideal gas law states that

$$PV = RT = N_{av}kT, \tag{1}$$

where P is the pressure, V is the volume of one mole of gas, R is the gas constant (1.98 cal/mol-K, or 82 atm-cm^3/mol-K), T is the absolute temperature in K, N_{av} is the Avogadro constant (6.02×10^{23} molecules/mole), and k is the Boltzmann constant (1.38×10^{-23} J/K, or 1.37×10^{-22} atm-cm^2/K). Since real gases behave more and more like the ideal gas as the pressure is lowered, Eq.1 is valid for most vacuum processes. We can use Eq.1 to calculate the molecular concentration n (the number of molecules per unit volume):

$$n = \frac{N_{av}}{V} = \frac{P}{kT} \tag{2}$$

$$= 7.25 \times 10^{16} \frac{P}{T} \text{ molecules/cm}^3, \tag{2a}$$

where P is in Pa. The density ρ_d of a gas is given by the product of its molecular weight and its concentration:

$$\rho_d = \text{molecular weight} \times \left(\frac{P}{kT}\right). \tag{3}$$

The gas molecules are in constant motion and their velocities are temperature dependent. The distribution of velocities is described by the Maxwell-Boltzmann distribution law, which states that for a given speed v,

$$\frac{1}{n}\frac{dn}{dv} \equiv f_v = \frac{4}{\sqrt{\pi}}\left(\frac{m}{2kT}\right)^{3/2} v^2 \exp\left(-\frac{mv^2}{2kT}\right), \tag{4}$$

where m is the mass of a molecule. This equation states that if there are n molecules in the volume, there will be dn molecules having a speed between v and $v + dv$. The average speed is obtainable from Eq. 4:

$$v_{av} = \frac{\int_0^\infty v f_v dv}{\int_0^\infty f_v dv} = \frac{2}{\sqrt{\pi}}\sqrt{\frac{2kT}{m}}. \tag{5}$$

An important parameter for vacuum technology is the molecular *impingement rate*, that is, how many molecules impinge on a unit area per unit time. To obtain this parameter, first consider the distribution function f_{vx} for the velocities of molecules in the x-direction. This function can be expressed by an equation similar to Eq. 4:

$$\frac{1}{n}\frac{dn_x}{dv_x} \equiv f_{v_x} = \left(\frac{m}{2\pi kT}\right)^{1/2} v_x^2 \exp\left(\frac{-mv_x^2}{2kT}\right). \tag{6}$$

The molecular impingement rate ϕ is given by

$$\phi = \int_0^\infty v_x dn_x. \tag{7}$$

Substituting dn_x from Eq. 6 and integrating gives

$$\phi = n\sqrt{\frac{kT}{2\pi m}}. \tag{8}$$

The relationship between the impingement rate and the gas pressure is obtained by using Eq. 2:

$$\phi = P(2\pi mkT)^{-1/2} \tag{9}$$

$$= 2.64 \times 10^{20}\left(\frac{P}{\sqrt{MT}}\right), \tag{9a}$$

where P is the pressure in Pa and M is the molecular weight.

Appendix L

Answers to Selected Problems

Answers are provided for odd-numbered problems, which have numerical solutions.

CHAPTER 2

1. (a) 2.35 Å; (b) 6.78×10^{14} (100), 9.6×10^{15} (110), 7.83×10^{14} (111) atoms/cm^2.
3. 52% (simple cubic), 74% (fcc), 34% (diamond).
5. (643) plane
11. 0.583 eV (77 K), 0.569 eV (300 K), 0.557 eV (373 K).
13. 72.7 Å, 1154 Å.
15. 2.26×10^{16} cm^{-3}.
17. (a) 9.3×10^4 cm^{-3}, 0.26 eV; (b) 10^{15} cm^{-3}, 0.26 eV
19. $n = 10^{15}$ cm^{-3} ($N_D = 10^{15}$ cm^{-3}), 0.93×10^{17} (10^{17}), 0.27×10^{19}(10^{19}).

CHAPTER 3

1. 3.31×10^5 Ω-cm (Si), 2.92×10^8 Ω-cm (GaAs).
3. 167 cm^2/V·s.
5. 3.5×10^7 cm^{-3}, 400 cm^2/V·s.
7. 0.226 Ω-cm
9. $N_A = 50\ N_D$.
11. (b) 259 V/cm.
13. (a) $\Delta n = 10^{11}$ cm^{-3}, $n = 10^{15}$ cm^{-3}, $p = 10^{11}$ cm^{-3}
19. $p(x) = 10^{14}(1 - 0.9e^{-x/Lp})$, $L_p = 31.6$ μm.
23. 0.403, 7.8×10^{-9}
25. 1.35×10^5 cm/s < 9.5×10^6 cm/s (100 V/cm) 1.35×10^7 cm/s ≈ 9.5×10^6 cm/s (10^4 V/cm).

CHAPTER 4

1. 1.867 μm, $V_{bi} = 0.52$ V, $\mathscr{E}_m = 4.86 \times 10^3$ V/cm.
3. At 300 K, $V_{bi} = 0.714$ V, W = 0.97 μm, $\mathscr{E}_m = 1.47 \times 10^4$ V/cm.
5. For $N_D = 10^{15}$ cm^{-3}, $1/C_j^2 = 1.187 \times 10^{16}$ (0.834 – V).
7. $N_D = 3.43 \times 10^{15}$ cm^{-3}.
9. 2.5×10^{17} cm^{-3}.
11. $N_A = 2.2 \times 10^{15}$ cm^{-3}, $N_D = 5.4 \times 10^{15}$ cm^{-3}.
13. 0.79 V.
15. 8.78×10^{-3} C/cm^2.
17. cross-sectional area is 8.6×10^{-5} cm^2.
19. (a) 587 V, (b) 42.8 V.
21. For V= +0.5 V, $V_{bi} = 1.1$ and 3.4×10^{-4} V; depletion widths = 3.82×10^{-5} and 1.27×10^{-8} cm.

CHAPTER 5

1. (a) 0.995, 199, (b) 2×10^{-6} A.
3. (a) 0.904 μm, (b) 2.54×10^{11} cm^{-3}.
5. (a)$I_E = 1.606 \times 10^{-5}$A, $I_C = 1.596 \times 10^{-5}$A, $I_B = 1.041 \times 10^{-7}$A; (b) $\beta_0 = 160$.
13. $\mu_E = 87.6$ cm^2/V-s, $\mu_B = 1186$, $\mu_c = 453$.
15. $\beta_0 = 50{,}000$.
17. 131.6.
21. $I_E = 1.715 \times 10^{-4}$A, $I_C = 1.715 \times 10^{-4}$A.
23. $f_T = 1.27$ GHz, $f_\alpha = 1.275$ GHz, $f_\beta = 2.55$ MHz.

25. 0.29.

27. 3.96×10^{-3} cm, 44.4 cm^2.

CHAPTER 6

5. 0.15 μm.

7. 0.59 V, 1.11×10^5 V/cm.

9. 2.32×10^{-2} V.

11. 7.74×10^{-2} V.

15. 3.42 V, 2.55×10^{-2} A.

17. 3.45×10^{-4} S.

19. 8×10^{11} cm^{-2}.

21. 1.7×10^{12} cm^{-2}.

23. 0.457 μm.

25. 0.83 V.

29. 49 nm.

31. 0.134 to 0.226 V.

33. 3.1×10^{-11} A.

35. 4.34 V.

CHAPTER 7

1. 0.54 eV, V_{bi} = 0.352 V.

3. ϕ_{Bn} = 0.64 V, V_{bi} = 0.463 V, W = 0.142 μm, $\mathscr{E}_m = 6.54 \times 10^4$ V/cm.

5. V_{bi} = 0.605 V, ϕ_m = 4.81 V.

7. 0.108 μm.

9. (b) –2.06V.

11. 0.152 μm, 0.0496 μm.

13. 5.8 nm.

15. 44.5 nm, –0.93 V.

CHAPTER 8

1. 11.25 nH.

3. 18.9 nm, 6.13×10^{-7} F/cm^2.

5. (a) 137 Ω; (b) 318.5 V.

7. (a) 74.8 V; (b) 2.2×10^5 V/cm; (c) 19 GHz.

9. (a) 10^{16} cm^{-3}; (b) 10 ps; (c) 2.02 W.

11. 3 meV, 11 meV.

CHAPTER 9

1. 1.57 mW.

3. (a) 40.58 cm^{-1}; (b) 36.6%.

7. 3.83 nA.

9. 0.0091 $(°C)^{-1}$.

11. 9.

13. 138 V, 90 ps.

15. V_m= 0.64 V, P_m = 52 mW.

17. 35.6 mW (R_s =0), 9 mW (R_s = 5Ω).

CHAPTER 10

1. At x = 0, $C_s = 3 \times 10^{16}$ cm^{-3}; at x = 0.9, $C_s = 1.5 \times 10^{17}$ cm^{-3}.

3. 0.75 g.

5. 6.56 m.

9. 24 cm.

11. ±30%, ±1%.

15. 2.14×10^{14} cm^{-3} at 900°C.

17. 4.68×10^4 cm/s.

19. 5.27×10^{14} atoms/cm^2.

CHAPTER 11

1. 44 min.

5. (a) x = 0.83, y = 0.46; (b) 2×10^{11} Ω-cm.

7. 0.0093.

9. 757°C.

13. 2.1×10^{11} cm^{-2}.

15. 71.1 nm for $TiSi_2$.

17. (a) 0.93 ns; (b) 0.42 ns; (c) 0.45.

19. 72 Ω, 0.18 V.

CHAPTER 12

1. (a) 2765; (b) 578; (c) 157.

3. 7 wafers/hr for positive resist, 120 for negative resist.

9. (a) W_b = 1.22 μm; (b) 0.93 μm; (c) 0.65 μm.

11. Etch from the top, wafer area lost = 127 cm^2; etch from the bottom, wafer area lost is small.

13. 224.7 nm/min.

17. 433.3 nm.

CHAPTER 13

1. 0.15 μm, 5.54×10^{14} atoms/cm^2.

3. 25 min, 3.4×10^{13} atoms/cm^2.

5. 16.9%.

7. x_j = 32.3 nm.

11. 6.7 s.

13. 0.53 μm.

15. 99.6%.

21. 0.927 μm.

CHAPTER 14

1. 781 MΩ.

3. 13 turns.

7. (a) 0.91 V; (b) peak concentration = 2.2×10^{21} cm^{-3}.

9. (a) 0.565 V; (b) 9.98 V.

11. 740 μm.

15. 1.84 nm.

19. 1.38 ns, 207 ns.

21. 17.3 Å, 16.7 Å.

Index

(100) plane, 24, 26, 376, 428, 450
(110) plane, 24, 301, 428
(111) plane, 24, 376, 428, 431, 450
⟨100⟩ direction, 181, 182, 347, 479, 497, 500, 504, 528
⟨110⟩ direction, 475
⟨111⟩ direction, 25, 79, 334, 347,
⟨111⟩ surface, 181, 182
⟨111⟩-axis, 475
3–D integration, 182

Abrupt junction, 91–93, 95, 96, 101, 103, 104, 114, 118, 119, 120–124, 128, 129
Absorption, 282–287, 306, 314–316, 319, 328, 330, 331
  coefficient, 286, 287, 314, 315, 330
Acceptor, 17, 37, 38, 41, 45, 46
Acceptor vacancy, 462
Accumulation, 171, 174, 177, 220
Acetic acid, 428, 430
Activation energy, 114, 384, 389, 395, 455, 479
Active mode, 131–135, 138, 139, 142, 144, 146, 151, 153, 167
Air mass 0, 319
Air mass 1.5, 319, 320
Air molecules, 360
Air traffic control beacons, 255
AlCu etch chamber, 439
AlF_3, 441
AlGaAs, 246, 247, 248, 250, 251, 253, 307, 309, 310, 330
AlGaAs/GaAs heterojunction, 248, 253
AlGaInN, 292
AlInAs/GaInAs HBT, 275
AlN, 292
AlSb, 537
Aluminum (Al) 18, 20, 21, 171, 179, 370, 378, 379, 384, 388, 390, 392, 393, 395, 396, 401, 402, 428–430, 440, 442, 448–450, 452
  alloy, 430
  alloying with copper, 395
  arsenide (AlAs), 3
  corrosion, 442
  silicon diodes, 236
  spiking, 394, 395
$Al_xGa_{1-x}As$, 152, 299, 333, 358, 361, 362
$Al_xGa_{1-x}As_ySb_{1-y}$, 298
$Al_xGa_{1-x}N$, 292
Ammonia, 384, 401
Ammonium fluoride(NH_4F), 428
Amorphous Si (a-Si) solar cell, 325, 326, 328
Amorphous silicon, 210, 286, 325
Amorphous surface layer, 476, 483
Amplifier, 517
Amplifying circuit, 146
Analog circuits, 497, 517
Angular frequency, 297
Anisotropic etching, 325, 431, 440, 441, 449
  profiles, 434
  reactive sputter etching, 512
Annealing, 452, 466, 470, 477, 479–481, 485, 488
Anode, 156, 157, 161, 164
Antimony (Sb), 389
Antireflection coating, 314, 315, 317, 318, 320, 328
Ar, 25, 334, 376
Arc lamps, 479
Area defect, 349, 365
Argon, 25
Aromatic diamine, 294
Arsenic, 66, 333, 343, 344, 345, 346, 356, 357, 359, 360, 365, 366, 379, 453, 478, 483, 497, 498, 499, 505, 506, 511, 528
Arsine (AsH_3), 356, 453, 469
As (see Arsenic)
As_2O_3, 453
$AsCl_3$, 453, 469
AsH_3 (see Arsine)
a-Si: H TFT, 210, 211, 212
Aspect ratio, 439, 440, 441, 448
Asymmetric thyristor, 163, 164, 166
Atmospheric "windows," 256
Atmospheric pollution monitoring, 298
Atmospheric pressure, 354, 358
  CVD, 400
Atmospheric-pollution monitoring, 6
Atomic number, 421, 424
Attenuation, 296
Attenuator, 483
Auger recombination, 78, 79, 81
Automobile industry, 1
Avalanche breakdown voltage, 120, 128, 467
Avalanche multiplication, 85, 117, 118, 127
Avalanche noise, 317
Avalanche photodiode, 317, 318, 328, 331

B_2H_6, 453
Backscattering, 422, 423, 425
Ballistic collector transistor (BCT), 155, 156
Ballistic devices, 279
Ballistic electron, 275
Bandgap, 17, 19, 31–33, 35–37, 42–46, 124, 246, 247, 288, 292, 295, 298, 311, 324, 325, 537
Bandgap narrowing effect, 44
Band-to-band recombination, 60, 80
Barium strontium titanate (BST), 387
Barrier height, 226, 227, 230–234, 236, 242, 246, 248, 251–253

Barrier metal layer, 392
Barrier-lowering effect, 202
Base, 156, 157, 158, 162, 163, 164, 165, 166, 167
  region, 134, 137, 138, 140, 141, 144, 146, 147, 151, 153, 155, 167, 168
  transport factor, 136, 137, 151, 156, 167
  width modulation, 146, 148
  width modulation effect, 148
Basic fabrication steps, 84, 85
BBr_3, 453
BCl_3, 440, 441, 450
Be, 346, 361
Beam-blanking plates, 429
Beryllium, 424
BF_3, 469
BiCMOS, 205, 209, 223
  inverter, 517
Bidirectional thyristor, 161, 162, 167
Binary compound, 19, 20, 22
Biotechnology, 444
Bipolar junction transistor (BJT), 130, 152, 154, 168
Bipolar technology, 489, 496, 527, 528
Bipolar transistor, 3, 4, 7, 8, 11, 13, 130, 131, 132, 135, 137, 142, 146, 147, 151–157, 165–168, 489–491, 494, 496–499, 501, 503–505, 517, 518, 528, 529
Bits, 507
BN, 453
Body-centered cubic (BCC), 21
Bphr model, 28
Boron (B), 195, 196, 222, 335–338, 346, 353, 358, 367, 379, 380, 453, 455, 462, 464, 469, 472, 477–479, 481–488, 498, 499, 501, 503, 505, 514, 516, 517, 529
Boron implantation, 195
Boron trioxide (B_2O_3), 346
Bow, 347, 349
Bragg reflector, 424
  wavelength, 308, 330
Breakdown region, 242
Breakdown voltages, 118–120, 122, 128, 155, 160, 262–264, 281, 405, 500
Bridgman technique, 7, 345, 347
Buffered HF solution (BHF), 381, 384, 428–430
Buffered-oxide-etch (BOE), 428
Built-in electric field, 149, 155
Built-in potential, 91, 92, 94, 95, 99, 105, 125–128, 227, 234, 243, 252, 494
Bulk micromachining, 443, 448
Bulk punch-through effect, 202
Buried layer, 469, 483, 485, 497–500, 504, 518, 519
Buried-type PMOS, 516

C_2F_6, 441
C_4F_8, 441
C_5F_8, 441
Cadmium, 346, 357
Camera, 510, 525
Cantilever-beam piezoresistive acceleration sensor, 443
Capacitance, 493–497, 500, 506, 507, 524, 525
  voltage technique , 103, 462
  voltage characteristics, 101, 102, 178
Capture cross section, 64
Carbon, 333, 351, 353
Card telephone, 510
Carrier drift, 47, 48, 77, 80, 81
Carrier generation, 60, 311, 312
Carrier injection, 60, 68
Carrier lifetime, 297, 312, 313, 331
Carrier mobility, 461
Carrier transport, 311
Cathode, 156, 157, 161, 164
CdS, 537
CdTe, 537
Cell projection, 448
Cellular phones, 6, 214, 255, 510
Central processing unit, 10, 13, 510, 526
CF_4, 440–442
Chain scission, 421
Channel conductance, 187, 190, 192, 193, 198, 222
Channel length, 187, 199–202
Channel resistance, 190, 204
Channel width, 187, 218
Characteristic impedance, 256, 257, 280
Charge density, 174, 183, 184, 204, 222, 223
Charge distribution, 228, 229
Charge neutrality, 90, 94, 100
Charge sheet, 248, 249
Charge-coupled device (CCD), 7, 170, 185, 186
Charge-sharing model, 199, 223
Chemical etching, 426–428, 431, 432, 434, 443, 449
Chemical vapor deposition (CVD), 87, 354, 358, 365, 376–381, 384, 386, 387, 391, 392, 396, 400, 401, 429
Chemical-amplified resist, 414
Chemical-mechanical polishing (CMP), 11, 396–398, 442, 514
Chemical-vapor deposition (CVD) SiO_2, 376, 391
CHF_3, 440, 441
Chip card, 510, 527
Chromium, 111, 411
Circuit symbol, 131, 132
Clean room, 405–407, 414, 438, 447, 449
Clear air turbulence detection, 255
Cleaved ends, 301
Clustered plasma processing, 439
Clustered tools, 439
CMOS(complementary MOSFET), 8, 9, 11, 14, 15, 169, 200, 205–209, 212–215, 220– 223, 465, 483, 485, 489, 500, 503, 509, 512–516, 518, 519, 524, 526–529
  inverter, 205–208, 215, 220, 510, 511, 517, 529
  logic gates, 517
  technology, 200, 205, 220, 221, 503, 510, 517, 525, 527, 529
Cobalt silicide, 398, 400
Cold wall reactor, 379
Collector, 155–158, 166
  -emitter leakage current, 145
Collimating tubes, 391
Color bands, 282
Columnar structure, 389
Common-base configuration, 133, 136, 143

Common-base current gain, 145, 148, 157, 167, 168
Common-base cutoff frequency, 148
Common-emitter configuration, 143–148, 152, 155, 167, 168
Complementary error function, 457, 483
Complementary MOSFET (see CMOS)
Compound semiconductor, 17, 19, 22
Computer aided design (CAD), 411, 495
Concentration-dependent diffusivity, 452, 462, 499
Conduction band, 29–46
Conduction band discontinuity, 154, 247
Conductivity, 17, 18, 32, 41, 44, 76, 81, 82, 190, 311, 312
Conductor, 17, 32–34
Confinement factor, 300, 301, 306, 330
Conformal step coverage, 381, 382, 391, 400
Constant-surface-concentration diffusion, 457, 458, 460, 463, 467, 485
Constant-total-dopant diffusion, 456, 459, 460, 467, 471, 485
Constant-field scaling, 202
Contact holes, 391
Contact imaging sensors (CIS), 195, 210, 221
Contact plug, 391
Contact printing, 407
Contact resistance, 469, 485
Continuity equation, 47, 66–68, 80, 82, 137, 168, 338
Contrast ratio, 413
Conversion efficiency, 264
Coplanar waveguide stripline, 256
Copper (Cu), 18, 111, 431, 441, 442, 456, 478
  halides, 442
  interconnect, 8, 11
  metallization, 396, 400
  polyimide interconnect, 396
Correction factor, 54
$CoSi_2$, 398, 399, 402, 524
Coulomb force, 49
Coulombic interaction, 471
Coupled impurity vacancy pair, 465
Covalent boning, 27
CRAY 1, 11
Critical angle, 292, 295, 300, 330
Critical dimension, 408, 439
Critical dimension (CD) control, 437
Critical field, 119, 120, 122, 262
Critical layer thickness, 362, 363, 365, 366
Crystal defect, 349
Crystal growth techniques, 17, 24, 45, 332, 343, 345, 367
Crystal lattice, 452, 454, 485
Crystal plane, 17, 22
Crystal puller, 334
Crystal purification, 341
Crystal structure, 17, 19, 22, 25, 44, 45
Cubic lattice, 349
Current density, 231–234, 252
  equations, 59, 60
Current gain, 131, 135–137, 141, 142, 144, 145, 148, 151, 153–155, 157–159, 165–168, 206, 209
Current transport, 231, 234, 251
Current voltage characteristics, 84, 88, 104, 106, 108, 112, 113, 127, 128, 137, 138, 143, 145, 156, 158, 160–162, 165
Cutoff frequency, 131, 148, 149, 153–155, 165, 166, 168, 237, 244, 245, 250, 251, 273, 277, 279
Cutoff mode, 142, 149
CVD (see chemical vapor deposition)
  for GaAs, 356
  for silicon, 355
Czochralski crystal, 26, 334, 352
Czochralski technique, 7, 25, 26, 333, 334, 339, 341, 345, 346, 365, 367

Damage layer, 426
Damascene process, 442
Dangling bonds, 65, 210
Data communication, 255
De Broglie wavelength, 309, 539
Deep UV lithography, 414
Defect, 332, 333, 335, 349, 351, 361, 363– 368, 431, 449
  density, 210, 397, 400, 411, 449
  free zone, 352
Degenerate energy levels, 28
Degenerate semiconductors, 39, 44, 299
Degree of anisotropy, 431
Degrees of freedom, 48
Density, 371–373, 375, 380, 384, 387, 390, 394, 397, 400, 401, 402
Density of states, 34–36, 39, 40, 44, 46, 231, 232, 309, 310, 538, 540
Denuded zone, 352, 353
Depletion, 172–178, 186–189, 194, 195, 199, 200, 202, 204, 213, 214, 219–221
  approximation, 174, 178
  capacitance, 84, 100, 101, 115, 116, 127, 128, 147
  case, 172
  layer, 84, 92, 94–104, 110, 115, 121, 122, 127–129
  layer width, 92, 94–101, 104, 110, 121, 128, 129, 233–235, 239, 240, 244, 252
  mode, 243, 248, 252, 253
  mode device, 243, 248, 252
  region, 84, 89, 91–94, 97, 99–102, 104–106, 109–111, 115, 118, 119, 127, 128
Depth of focus, 449
Developer, 412–414, 416, 421, 444
Device building blocks, 2
Device leakage, 199
DH laser (see double heterostructure laser)
DI water, 428, 430
Diac (diode ac switch), 161, 162, 166
Diamond structure, 17, 22
Diaphragm, 443
Diborane, 356

Dichlorosilane, 355, 380, 382, 384
Dichlorosilane nitrous oxide, 382
Dielectric constant, 384–387, 396, 401, 402, 537
Dielectric isolation, 499–501
Dielectric layers, 369, 370, 400
Dielectric permittivity, 229, 247
Dielectric relaxation time, 266
Dielectric strengths, 385
Diethylzine $Zn(C_2H_5)_2$, 357
Diffraction, 447
  effect, 414, 426
Diffusion, 7, 8, 15, 86–90, 93, 101, 104, 107, 108, 110, 112–115, 117, 123, 127, 129, 427, 433, 434, 452–469, 471, 481, 483–486, 487
  barrier, 392
  capacitance, 101, 115, 129, 147
  coefficient, 58, 83, 338, 373, 377, 393, 394, 455, 464
  constant , 138, 140, 149, 167
  current, 57–59, 69, 70, 82, 83
  equation, 454, 455, 457, 459, 463, 466, 485,
  in silicon, 453, 458, 464, 487
  length, 69, 82, 106–108, 114, 115, 129, 138, 139, 165, 457, 460
  length of electrons, 107
  process, 57, 452, 453, 455, 460, 461, 466, 469, 471, 485, 486
  profiles, 457, 459, 462, 464–466
Diffusivity, 51, 58, 59, 82
Digital cameras, 6, 214, 510
Digital circuits, 497
Digital video disk, 6
Dipole layer, 268, 269
Direct exposure, 418
Direct recombination, 60, 63, 65
Direct tunneling, 524
Dislocation, 349–354, 362–367, 405
Dislocation loop, 364
Dissociation effect, 465, 485
Dissolution of silicon, 394
Distributed bragg reflector laser, 308
Distributed feedback (DFB), 308, 328
Distributed feedback laser, 308
D-MOSFET, 218
Donor, 17, 37–39, 41, 45, 46
Donor concentration, 230, 238, 252
Donor vacancy, 463
Dopant profile, 460, 461, 483
Doping distribution, 336, 338, 339, 366
Double barrier structure, 271
Double charged acceptor vacancy, 463
Double drift IMPATT diode, 262
Double-heterostructure (DH) laser, 299, 300, 303–307, 309, 310, 330
Drain, 186–194, 197–200, 202, 205, 209–213, 216–223
  -induced barrier lowering (DIBL), 200, 202, 516
DRAM (see dynamic random access memory)
Drift current, 53, 57, 82, 187
Drift region, 219, 262–264, 281
Drift velocity, 48, 49, 56, 59, 70, 75, 76, 78, 80, 81, 83, 519, 522
Dry etching, 10, 442, 447, 449
Dry oxidation, 86
Dry oxygen, 370, 373, 374, 376
Dual damascene process, 396
Dust particle, 405–408, 449
Dynamic random access memory (DRAM), 8, 9, 11, 13, 14, 387, 400, 439, 490, 507–509, 522, 526, 527

Earth's crust, 19
Early effect, 146
Early voltage, 146
E-beam evaporation, 390, 402
Edge dislocation, 349, 362, 365, 366
Effective density of states, 36, 39, 46
Effective mass, 30–32, 34, 37, 46, 73, 75, 76, 79, 232, 235, 236
Effective segregation coefficient, 338, 340
Effusion oven, 358, 360, 368
Einstein relation, 58, 59, 319
Elastic collision, 472
Elastic forces, 362
Electrically erasable programmable read-only memory (EEPROM), 216–218
Electrochemical anodization, 370
Electrochemical methods, 396
Electroluminescence phenomenon, 3
Electromagnetic spectrum, 282, 283
Electromigration, 392, 442
Electron affinity, 73, 124, 170, 195, 226, 252
Electron bombardment, 195
Electron cyclotron resonance (ECR), 390, 436–438
Electron gun, 417, 419
Electron mobility, 49, 50, 81, 82, 519
Electron resist, 411, 418, 421, 422
Electron temperature, 275, 277, 281
Electron-beam lithographic system, 411, 418–422, 426, 447– 449
Electron-hole pair generation, 111
Electronic industry, 1, 3, 13, 14
Electronic purse, 510
Electronic stopping, 471–473, 475
Electronic ticket, 510
Electronic-grade silicon, 333
Electrostatic micromotor, 444
Element semiconductor, 18, 19, 22
Embedded DRAM, 526, 528
Emission lines, 308
Emission processes, 109
Emission spectra of LEDs, 293
Emission wavelengths, 289, 298, 310
Emitter, 146–157, 165–167
  efficiency, 136, 137, 141, 142, 151, 153, 154, 166, 167
End-point control, 405
Energy barrier thickness, 270

Energy momentum diagram, 30, 31
Enhancement-mode MOSFET, 249, 507, 510
Epitaxial film, 405
Epitaxial growth, 8, 10
  rate, 480
  technique, 104, 332, 354, 367
Epitaxial layer, 237, 238, 243, 252, 497, 498, 501, 519
Epitaxial process, 357, 361, 365
Epitaxy, 333, 357, 366
Equilibrium segregation coefficient, 335, 336, 346
Equivalent circuit, 115, 116, 146–148, 494, 495, 509, 510
Erfc (see error function)
Error function , 457, 458, 462, 463, 467, 482, 485
Etch anisotropy, 440
Etch chemistry, 436
Etch mechanism, 433, 434, 436
Etch rate, 380, 427–431, 434, 435, 438–443, 450
Etch selectivity, 434, 436, 439, 440, 448, 450
Etch solution, 427, 430
Etching, 427–444, 447–451
  equipment, 435
  operations, 499
  temperature, 392
EUV exposure tool, 424, 447
Evaporation, 394, 402
Exclusive NOR, 274, 275
Exclusive OR, 277
Exhaust system, 334
Exposure response curve, 413
Exposure tools, 404, 408, 410, 417, 420, 424, 447
Extreme-ultraviolet (EUV) lithography, 423
Extrinsic, 285, 311, 312
  base regions, 503
  diffusion, 452, 462–464, 466, 487
  photoconductor, 312
  semiconductor, 53
  stacking fault, 349, 351
  transition, 285
Eye response, 288, 293

F-doped oxide, 524
Fermi distribution, 34, 35, 40
Fermi level, 17, 34–36, 39–43, 46
Fick's diffusion equation, 7, 455
Field oxide, 369, 376, 379, 395, 401
Field transistor, 197, 222
Field-effect transistor (FETs), 250, 251
Film formation, 490, 499
Film thickness, 414, 431, 435
Fixed charge, 182, 196, 223
Fixed-oxide charge, 180, 182, 183, 187, 194
Flash memory, 216, 218
Flat doping profile, 481
Flat-band condition, 171, 173, 179, 180, 182–184, 192, 194, 221, 222
Flat-band voltage, 179, 182–184, 194, 222
Flat-diode configuration, 290
Flat-panel display, 294
Flip-flop structure, 215
Floating dot, 217, 218
Floating gate, 6, 216, 509, 510, 528
Float-zone process, 332, 339–341, 365, 367
Float-zone silicon, 342
Flow control, 334, 358
Focused electron beam, 390
Forward breakover, 157, 159, 160
Forward-blocking state, 160
Four-point probe method, 54, 351
Free-carrier flux density, 126
Frenkel defect, 349
Frequency locking circuit, 259
Frequency response, 130, 146, 149, 155, 165, 167, 168
Full width at half maximum (FWHM), 271, 293
Full-color displays, 293
Full-color indicators, 293
Fused quartz, 17
Fused silica, 411
Fused-silica crucible, 346

Ga, 425, 431
$Ga_{0.47}In_{0.53}As$, 246
$Ga_{0.51}In_{0.49}P$, 326
$Ga_{1-x}In_xAs$, 363, 368
GaAs (see gallium arsenide)
$GaAs_{1-y}P_y$, 19, 20
Gain factor, 306, 330
GaInAs/AlInAs material, 276, 277
GaInAs-AlGaAs DH, 307
$Ga_xIn_{1-x}N$, 289, 292
GaInP/GaAs tandem cell, 328
Gallium arsenide (GaAs), 3, 7, 18–20, 27, 31, 32, 36–39, 43–45, 54, 59, 60, 76, 78, 79, 227, 230, 232, 237, 243, 246, 247, 251–253, 265, 266, 269, 271, 272, 274, 276, 279–281, 289–292, 295, 298, 304, 332, 333, 343–345, 344, 345–347, 354, 356–363, 365–368, 370, 453–455, 462, 466, 469, 473, 475, 481, 483, 485, 519, 522, 537, 556
Gallium arsenide etching, 431
Gallium chloride, 356, 357
Gallium phosphide (GaP), 289, 290, 292, 293, 537
Gallium-arsenic system, 343, 344
GaN, 283, 284
GaP, 19, 20
Gas source, 334
Gas thyratron, 130
GaSb, 260, 280, 537, 538
Gate, 156, 160–166, 168, 172, 180, 186–189, 193–197, 199, 200, 202, 205, 210–212, 215–223
  capacitance, 189
  length, 237, 238, 245, 250, 251
  oxide, 505–509, 522–524, 526–528
  oxide thickness, 522, 523, 527
  turn-off, 164, 166
Gaussian distribution, 460, 462, 471, 481, 485, 499
Gaussian spot beam, 420
$Ga_xIn_{1-x}As_yP_{1-y}$, 287, 297, 298
$Ga_xIn_{1-x}As_ySb_{1-y}$, 298
$Ge_{0.3}Si_{0.7}$ on silicon, 363
Generation lifetime, 109
Generation-recombination centers, 109, 124, 151
Generation-recombination rates, 178

Germanium (Ge), 260, 278, 279, 314, 326, 537
Gettering, 351– 353
Ge_xSi_{1-x}, 362, 363
G-line, 410
Global planarization, 370, 397, 400
Gold, 21, 111, 117, 424, 425, 446, 465
  contact, 242
  doping, 209
  silicon photodetector, 317
Graded base, 154, 155
Graded layer, 155
Graded profile, 155
Graded-index fiber, 295, 296
Gradient of the diffusion profile, 458, 460
Gradual-channel approximation, 188
Grain boundary, 211, 349
Graphite boat, 345
Graphite crucible, 346, 354
Graphite susceptor, 334, 351, 357
Greek Alphabet, 535
GRIN-SCH structure, 310
Groove, 347, 348
Gross world product (GWP), 1
Gunn diode, 6

H_3PO_4, 430, 431, 516
Half life, 343
Hall coefficient, 56, 82
Hall effect, 55, 56
Hall field, 56
Hall measurement, 55, 82
Hall voltage, 56, 82
Haynes-Shockley experiment, 70, 71, 81
HDP, 433, 437, 441
He, 376
Heater, 160
Heavy holes, 309, 310
Helium, 298, 317, 318
  -neon gas laser, 298
  -neon laser, 317, 435
Heteroepitaxial growth, 357
Heteroepitaxy, 333, 362, 365, 366
Heterojunction, 84, 85, 124–129, 294, 296, 297, 299, 300, 304, 317
  bipolar transistor, 3, 4, 130, 131, 151, 153, 155, 165, 166, 168
  field-effect transistor, 361
  interface, 2, 3
Heterostructure, 246
  laser, 3, 4
Hexa-methylene-di-siloxane (HMDS), 414
HF, 428, 429, 449, 461, 487, 516
HFET, 246, 253
High current implantation, 453, 481, 483–485
High dielectric-constant materials, 215
High electron mobility transistor (HEMT), 246
High energy implantation, 453, 481, 483, 485
High frequency power conversion equipment, 164
High injection, 109, 112, 113, 127
High-aspect feature, 391
High-current densities, 113
High-density plasma (HDP), 432, 436–441, 448
  etching, 432
High-fidelity transfer, 431
High-field effects, 47, 75
High-frequency equivalent circuit, 146, 148
High-k dielectric materials, 524
High-k materials, 387
High-resolution gas spectroscopy, 298
High-speed laser printing, 298, 328
High-voltage direct current (HVDC) devices, 164, 166
Hi-lo structure, 261, 262
H-line, 430
HNO_3, 461
Hole, 27–29, 31–46
Hole injection, 232, 233
Hole ionization rates, 119
Hole mobility, 49
Homoepitaxial growth, 361
Homoepitaxy, 333, 361, 366
Homojunction laser, 303
Hot carrier injection, 181
Hot electron device, 275, 279, 280
Hydrofluoric acid, 428
Hydrogen (H_2), 318, 333, 355, 357, 384, 385, 391, 392, 402, 469, 474
  annealing, 181
  atom, 28, 37, 39
  chloride (HCl), 24, 333
Hydrogenation treatment, 211
Hydrophilic surface, 414
Hydrophobic surface, 414
Hyperabrupt junction, 103, 104

IC card, 510, 511
IC design, 411
IC processing, 417, 427, 444, 447
Ideal characteristics, 106, 127
Ideal diode current, 323
Ideal diode equation, 108–110, 114
Ideal MOS curves, 175
Ideality factor, 112
IGFET, 186, 221
III-V compound semiconductor, 237
III-V nitride LED, 292
I-line, 410
Impact ionization, 78, 80, 81
IMPATT diode, 254, 255, 261, 262–264, 267, 278–281
Implant damage, 452, 477, 484, 485, 488
Impurity distribution, 461, 468
Impurity redistribution, 452, 467, 469, 485
Impurity scattering, 49, 50
InAs, 537
Incandescent lamps, 294
Index of refraction, 380, 381, 409
Indirect recombination, 60, 63, 65
Indium phosphide (InP), 19, 20, 244–246, 251, 265, 266, 537
Inductively coupled plasma, 436, 437
Information display, 329
Infrared LED, 295, 328
Infrared sensors, 311
Infrared system, 255
InGaAsP-InP, 296
InP (see Indium phosphide)
InP/InGaAs structure, 153
InP-based HBT, 153, 154
Input conductance, 148
InSb, 537
Insulator, 17, 18, 32, 33

Integrated circuit, 3, 4, 6, 14, 85, 169, 170, 199, 202, 205, 294, 310, 369, 370, 376, 378, 391, 392, 395, 398, 400, 401, 404, 412, 424, 443, 448, 489–493, 511, 520, 522
technology, 370
capacitor, 493
era, 522
inductor, 494
manufacturing, 205
resistor, 491
interacting *p-n* junctions, 148
Interconnect, 87, 439, 441, 442, 447
Interconnections, 370, 385, 398, 401
Interface, 467, 469, 488
trap density, 181, 400
-trapped charges, 181, 183, 184
Interlayer dielectric (ILD), 396
International System of Units, 533
International Technology Roadmap for Semiconductors, 522
Interstitial diffusion, 455, 456
Interstitial site, 349, 367
Interstitials, 454
Intrinsic, 311, 312, 315
base region, 503
carrier density, 34, 36, 37, 43, 44, 114
diffusion, 462, 465
diffusivity, 462, 485
Fermi level, 36, 40, 41, 46, 52, 170
photoconductor, 285, 287
semiconductor, 34–36
stacking fault, 349, 351
transitions, 285, 287
vacancy density, 463
Inversion, 172–178, 187, 189, 190, 192–194, 204, 211, 218, 220, 221
case, 172, 173
layer, 173, 174, 178, 187, 189, 192, 193, 204, 211,
mode, 142
staggered structure, 210
Ion beam, 452, 462, 470, 473, 475, 476, 483, 484, 488
Ion channeling, 475
Ion distribution, 471, 483
Ion implantation, 8, 87, 195, 416, 452, 453, 469, 471, 483–486, 499
Ion stopping, 471
Ion-beam lithography, 418, 425, 447
Ionization coefficients, 285, 287
Ionization energy, 37–39
Ionization integrand, 261
Ionization rate, 79, 80
Isoelectronic center, 290
Isopropyl alcohol, 428
ITO (indium tin oxide), 294

Junction breakdown, 84, 88, 117, 118, 121, 127, 129
Junction curvature effect, 123, 467
Junction depth, 187, 200, 460, 461, 464, 466, 485–487, 491, 522, 524
Junction spiking, 392

Kinetic energy, 48, 52, 53, 73, 78, 79, 81
Kink, 213, 465
Kirchhoff's circuit laws, 132
KOH, 428, 430, 444, 450
KrF excimer laser, 410

Lamps, 293, 294
Lanthanum hexa-boride (LaB_6), 419
Lapping, 426
Large-scale integration, 490, 522
Laser crystallization of Si, 212
Laser interferometry, 435
Laser printing, 6
Lasing modes, 302
Latch-up, 212, 221, 512, 513
Lateral diffusion, 452, 467, 485, 506, 513
Lateral oxide isolation, 497, 504
Lateral straggle, 471, 473, 474
Lattice, 19, 21–23, 27, 30, 45, 537
constant, 21–23, 45, 124, 125
scattering, 49, 50, 81
-matched, 333, 361, 362, 366
-mismatched semiconductors, 365
Lead zirconium titanate (PZT), 387
Leakage current, 187, 188, 202, 205, 209, 214, 222, 507, 511, 516, 524
LED displays, 293
Lifetime, 47, 61–63, 71, 82, 83, 209, 478
of minority carriers, 138
Liftoff technique, 416
LIGA process, 443, 444, 446, 448, 449
Light activated thyristor, 164
Light bulb, 160
Light control, 160
Light holes, 309
Light-emitting diodes (LEDs), 282, 288
Lightly doped drain (LDD), 483
Limited-source diffusion, 499
Line defect, 349, 365
Linear rate constant, 375, 376
Linear region, 187–191, 199, 200, 223, 241, 242, 249, 251
Linearly graded junction, 93, 97–99, 103, 120, 121, 128
Liquid crystal displays (LCD), 210
Liquid encapsulant, 346, 365
Liquid encapsulation method, 346
Liquidus line, 343, 344
List of Symbols, 531
Lithographic operations, 499
Lithographic step, 225
Lithography, 85–87, 127, 447–449
Lithography process, 7, 9
Load devices, 216, 220
Load line, 146, 147, 149, 150
Load transistors, 215
Local oscillator, 259, 264, 279
Local oxidation of silicon (LOCOS), 498
LOCOS process, 500
Logic circuits, 205
Long through sputtering, 390, 391
Longitudinal field, 245, 249
Longitudinal modes, 302, 307
Long-wavelength cutoff, 312, 314
Lorentz force, 55, 56

Loss, 292, 302, 306, 308, 315, 318, 319, 323
Low dielectric constant, 386, 401
Low-k polyimide, 396
Low pressure, 354
CVD (LPCVD), 378, 442
reactor, 388
Low-angle grain boundaries, 364, 366
Low-dielectric-constant (low-k) insulators, 524
Low-dielectric-constant materials, 494
Low-k materials, 386
Low-resistivity metals, 494

Magnetic field, 335
Magnetic tape card, 510
Magnetically enhanced RIE (MERIE), 432
Majority carrier current, 42, 233, 234
Mask, 439, 444, 446–450
beam system, 425
blank, 423, 424, 426, 447
thickness, 482
Masking, 453, 481–483, 485, 488
Maskless direct writing, 419
Mass action law, 105
Maximum output power, 322, 331
Maximum power, 322–324, 331
Maximum time to breakdown, 388
MBE (see molecular beam epitaxy)
Mean free path, 57, 360, 361, 367
Mean free time, 48, 49, 51, 57
Mean time to failure (MTF), 395
Medium-scale integration (MSI), 490
Melting point, 344, 346, 354, 365
Membrane, 421, 424, 443, 449
Memory devices, 489, 507, 510
Mercury-arc lamp, 410
MERIE plasma, 441
MESFET, 237–241, 243–245, 250–253
integrated circuit, 520
technology, 489, 519, 529
Metal, 32
film, 369, 370, 384, 400
interconnect, 439, 447
semiconductor contact, 224–226, 228, 229, 231, 237, 251, 252
semiconductor device, 261
Metallization, 370, 379, 390–392, 394, 396, 400, 401, 489, 499, 503, 506, 529
Metallurgical-grade silicon, 333
Metal-nitride-oxide-semiconductor (MNOS) , 509
Metalorganic chemical vapor deposition (MOCVD), 8, 10, 357, 365
Metalorganic compound, 357, 365,
Metalorganic molecular-beam epitaxy (MOMBE), 361
Metal-oxide-semiconductor (MOS) structure, 3
Metal-semiconductor contact, 3, 13
Metal-semiconductor interface, 2
Metal-semiconductor photodiode, 313–316
Metal-SiO_2–Si, 175
Metal work function, 226, 227, 252
Metric system, 405, 406
Microelectromechanical systems, 443
Microelectronics, 435, 443, 449
Microelectronic industry, 435
Micromechanic, 444
Microoptic, 444
Microprocessor, 8, 10, 11, 12, 15
Microprocessor chip, 490, 517, 518
Microprocessor-based control system, 334,
Microstrip, 256, 257
Microwave region, 254
Microwave technology, 254, 255, 280
Midgap, 174
Miller Indices, 22–24, 44, 45
Millimeter wave band, 255, 256, 279
Millimeter wave devices, 254
Millimeter-wave power amplifiers, 151
Minimum feature length, 11, 406, 504, 522, 525
Minimum linewidth, 408, 423
Minority carrier, 42, 178, 186, 209
Minority carrier current, 233, 234
Minority-carrier lifetime, 351
Minority-Carrier Storage, 114, 115, 127
MISFET, 186
Misorientation, 475, 476
Mix-and-match approach, 426, 447
Mixers, 151
Mobile ionic charge, 180, 182, 183
Mobile radio, 510
Mobility, 59, 70, 75, 76, 80–83, 153, 155, 187, 190, 210–212, 220, 461, 478, 485, 537
MOCVD, 357, 358, 361, 365
MODFET, 224, 225, 246–251
Modulation bandwidth, 297
Modulation frequency, 315
Modulation-doped field-effect transistor (MODFET), 4, 7
Mole fraction, 355
Molecular beam epitaxy (MBE), 8, 10, 271, 354, 357–361, 363, 365, 367, 368
Molecular impingement rate, 358
Molecular weight, 358–360, 367, 368
Molecular-beam epitaxy, 354, 357, 359, 361, 365
Monolithic IC, 8
Monolithic microwave integrated circuits (MMIC), 494, 522
MOS capacitors, 493
MOS devices, 440, 484, 485
MOS diode, 220, 221
MOS memory cells, 420
MOSFET, 3, 4, 6, 9, 10, 13, 220–223, 369, 395, 398, 401, 452, 484, 488
MOSFET scaling, 169, 199, 223
MOSFET technology, 489, 503, 528

$MOSi_2$, 399
MOST, 186
Motor drives, 165
Multilevel metallization, 370, 379
Multiple implantation , 481, 485
Multiple-quantum-well (MQW), 310, 311, 329

n-channel MOSFET, 186, 194, 197, 199, 204–206, 213, 215, 216
Near-infrared region, 314
Negative differential mobility, 265, 267
Negative photoresist, 411–413, 416, 422, 449
Negative resistance, 5, 260, 261, 263, 279
Neutron, 342, 343
Neutral vacancy, 462
Neutron irradiation, 209, 342, 343, 366,
NH_3, 384, 392
NiSi, 524
Nitric acid, 428, 430
Nitrided oxide, 517
Nitrogen, 290, 291, 329
Nitrous oxide, 380, 382
Noise factor, 317, 318
Nonconformal step coverage, 381
Nonequilibrium situation, 60
Nondegenerate semiconductor, 39
Nonvolatile semiconductor memory (NVSM), 6, 216, 218, 510
Notebook computer, 6
n-tub, 512, 513, 529
Nuclear collisions, 477, 478
Nuclear fusion reaction, 318
Nuclear spectroscopy, 255
Nuclear stopping, 471–473, 477, 478
Nuclear stopping power, 471–473
Numerical aperture, 409
n-well, 197, 206

Ohmic contact, 2, 4, 87, 224, 225, 234, 236, 237, 251, 292, 311, 324, 400
On condition, 149
Open tube, 454, 478
Optical absorption, 282, 283, 285–287, 315, 330
Optical concentration, 326, 328
Optical diffraction, 407
Optical emission spectroscopy, 434
Optical fiber, 288, 295, 296, 298, 303, 307, 308, 311, 328, 329
Optical fiber communication systems, 6, 288, 296, 298, 303, 308, 309, 328
Optical gain, 301, 302
Optical lithographic system, 417–419, 422, 424–426, 447, 448, 450
Optical proximity correction(OPC), 417
Optical reading, 298, 328
Optical resonant cavity, 285
Opto-isolator, 288, 295, 311, 328
Orangic LED, 294
Organic semiconductors, 294
Organic solvent, 412
Orientation dependent, 375
  etching, 404, 428, 429, 448, 450
  wet chemical etching, 443
OR-NAND, 277, 278
Oscillator, 151, 517, 525
Oxidation, 85–87, 127
Oxide layer, 225
Oxide masking, 8
Oxide thickness, 187, 197, 201, 222, 223
Oxide trapped charge, 180–184, 222
Oxygen, 335, 351–354, 359
Oxygen ion implantation, 500, 501
Oxygen molecule, 373
Ozone (O_3), 380

P^+-polysilicon, 179, 180, 195, 196
Pb, 17, 20, 32
P_2O_5, 453
P-glass flow, 382–384, 402
Periodic table, 18, 19
Phase diagram, 343, 344
  of the Al-Si system, 392, 393
Phase-shifting mask (PSM), 417
Phosphine (PH_3), 356, 453
Phosphoric acid, 430
Phosphorus (P), 290, 335, 342, 353, 366, 379, 380, 383, 384, 389, 401, 453, 454, 464, 465, 469, 472, 473, 478, 479, 483, 485, 487
  doped silicon dioxide, 379, 383, 401
  doped oxide (P-glass), 506
Photo-acid generator, 414
Photoconductivity method, 351
Photoconductor, 311–313, 328, 331
Photocurrent gain, 313
Photodetector, 311, 328
Photodiode, 312–315, 317, 318, 328, 329, 331
Photomask, 405, 407, 411, 414–416, 419, 424
Photon energy, 282, 283, 285, 288, 312, 314
Photonic application, 352, 363
Photonic devices, 3, 10
Photoresist, 7, 8, 14, 15, 87, 414–416, 421, 422, 429, 435, 442, 447, 449, 482, 488
Photoresist adhesion, 429
Photosensitive compound, 430
Physical constants, 536
Physical etching, 433, 434
Physical vapor deposition, 87, 390
p-i-n photodiode, 315, 316, 331
Pinch-off voltage, 241–243, 253
Pinch-off point, 187, 188
Planarization, 390
Planar process, 8, 160, 225
Planar transmission lines, 256–257
Planes, 301–303, 325
Plasma damage, 397
Plasma diagnostic, 433
Plasma etching, 431, 433, 435–436, 439, 442, 449, 451
Plasma enhanced chemical vapor deposition (PECVD), 378
Plasma nitride, 385
Plasma reaction, 385
Plasma reactor, 211
Plasma spray deposition, 390
Plasma-assisted etching, 404, 431, 447

Platinum silicide (PtSi), 236
Platinum (Pt), 18, 21, 464
p-n junction, 3, 4, 9, 14, 78, 80, 81, 169, 174, 176, 259, 261, 273, 279, 288, 328, 462, 486, 487, 492
  solar cell, 320–321, 331
$POCl_3$, 453
Point defect, 349–350, 365
Point-contact transistor, 4
Poisson's equation, 68, 90, 174, 183, 266, 268–269
Polishing, 426
Polonium, 21
Poly-butene-1 sulfone(PBS), 421
Polycide, 398–400, 440, 506
Polycrystalline GaAs, 365
Polycrystalline rod, 339
Polycrystalline silicon, 24, 199, 210, 334, 339, 365, 370
Poly-glycidyl methacrylate-co-ethyl acrylate (COP), 421
Polyimide, 521, 524, 529
Polymer linking, 413
Polymers, 421
Poly-methyl methacrylate (PMMA), 436
Polysilicon, 186, 210–212, 215–216, 219, 221–223, 370, 379, 380, 383, 388–391, 395, 398–402, 421
Polysilicon TFT, 211, 212, 215
Population inversion, 285, 299–301
Portable electronics system, 510
Positive resist, 427, 430, 435, 440, 443, 445, 450
Power amplifier, 264
Power consumption, 202, 205, 214, 220
Power limitations, 524
Power supplies, 164–166, 522, 523
Precipitate, 351, 352, 364, 366
Predeposition, 460, 483, 484, 485
Pressure sensor, 435, 443
Primary flat, 347–349
Principal quantum number, 28
Principle of detailed balance, 542
Probability of occupying, 34
Projected range, 470, 471, 473–475, 477, 478, 483–485
Projected straggle, 471, 473, 474, 485
Projection printing, 407, 409, 410
Properties of silicon dioxide, 380
Proximity effect, 423, 426, 447
Proximity printing, 424, 435, 437
p-tub, 511–513, 529
Pt film, 503
Pyrometer, 479
Pull rate, 334, 335, 339
Punch through, 153, 513, 518
  diode, 122
PVD, 390–393, 396, 400
p-well, 206, 208
Pyramids, 325, 326

Quality factor, 494
Quantum effect devices, 254, 270, 275, 281
Quantum efficiency, 290, 291, 304, 312, 315, 317, 318, 331
Quantum tunneling, 254, 259, 270
Quantum well thickness, 270
Quantum-well lasers, 309
Quartz, 210–212, 217, 495
  tube furnace, 453
  windows, 479
Quartzite, 24, 333
Quaternary compound, 19, 20
Quaternary, 289, 295, 298, 317
QW lasers, 310

Radar system, 255, 261
Radiation-damage toleration, 213
Radiative transitions, 282, 283, 289, 330
Radio astronomy, 255
Radio frequency, 435, 437, 494, 527
Radius of curvature, 467
Random access memory, 490, 507, 508, 522
Range, 452–453, 455, 466, 470–475, 477, 478, 481, 483–488
Rapid thermal annealing (RTA), 478, 503, 517
Raster scan, 419
RC time constant, 385, 396
RC time delay, 524
Reactive ion etching (RIE), 396, 432, 436
Reactor, 370, 378, 379, 385, 388–389
Real space transfer (RST), 255
Recombination, 47, 60–61, 63–67, 69, 70, 80–82
  center, 60, 63–66, 82
  current, 292
  rate, 60, 61, 64–66, 312, 541, 542
Recrystallization, 479
Rectification, 84
Rectifying contact, 2
Rectifying metal semiconductor barrier, 370
Redistribution diffusion, 460
Redistribution process, 468
Reflection loss, 292
Reflectivity, 301, 330
Refractive index, 292, 295, 299, 301, 308, 330, 417
Registration, 407
Reliability, 388, 401
Resin, 412, 414
Resist, 416–419, 421–424, 426, 429, 431, 432, 439, 442–444, 449, 450
Resistance, 224–225, 234, 236–239, 244, 492–495, 497, 500, 505–507, 516, 518–520, 523–524, 526, 528
  heating, 390
Resistivity, 17, 18, 33, 47, 51, 53–55, 81, 339, 343, 351, 353, 354, 365, 367, 442
  of silicon nitride, 384
Resolution, 404, 407, 408, 443, 447–449
  enhancement techniques, 404, 410
Resonant circuit, 104, 257
Resonant frequency, 104, 258, 280
Resonant tunneling diode (RTD), 7, 254, 279, 543
Response speed, 311, 314–315, 317

Retrograde well, 209, 513, 523
Reverse substrate source bias, 195
Reverse-blocking state, 157, 160
Reverse-breakdown region, 157
Reverse-breakdown voltage, 160
RF heating, 379, 390
Richardson constant, 232
RIE, 432–437, 439, 441
Ring oscillator, 525
Rolling mills, 165
Rotation mechanism, 334
RST transistor (RSTT), 276
Ruby laser, 298

Sacrificial layers, 444
Salicide, 398, 399, 402
Sand, 24, 333, 365
Sapphire substrate, 292
Satellite, 318, 325
Saturation channel current, 245
Saturation current, 192, 232–233, 238, 249, 252–253
Saturation current density, 108, 114, 232, 252
Saturation mode, 142, 144, 149, 151, 160
Saturation region, 187, 190–192, 200, 242–243
Saturation velocity, 75, 76, 79, 81, 262–264
Saturation voltage, 242, 249
Scaling of MOSFET, 204
Scaling rules, 202
SCALPEL system, 420, 421, 426, 448
Scanning electron microscope, 383
Scanning focused-beam system, 434, 436, 437, 439–440, 443, 448, 450
Schottky barrier, 225, 231, 237, 238, 246, 248, 251–253
  diodes, 227, 230–232, 462
Sealed ampules, 454
Secondary flat, 347–349
Secondary ion mass spectroscope (SIMS), 351, 462
Seed, 25, 45, 334, 339, 340, 346, 347, 354, 365– 367,
  crystal, 334, 339, 347, 354, 365, 367
  holder, 334, 340
Segregation coefficient, 338, 340, 346, 365, 468–487
Selective transistor, 217–218
Selectivity, 435, 443–444, 447, 449
Selenium, 346, 357, 454
Self-aligned metal silicide, 398
Self-aligned gate process, 9
Self-aligned structure, 500–501, 503
Semiconductor controlled rectifier (SCR), 156
Semiconductor industry, 1, 7, 10, 12, 14
Semiconductor lasers, 285, 298, 299, 303
Semiconductor material, 17–20, 24, 44
Semiconductor power devices, 128
Semiconductor surface, 170–175, 183, 187, 193
Semiconductor technologies, 1, 7, 8, 14
Semiconductor work function, 226, 231
Semiinsulating GaAs 346, 519
Semiinsulating substrate, 238, 243
Sensor, 407
Separation by implantation of oxygen (SIMOX), 485, 500
Series resistance, 112, 127, 237, 323, 362, 493–494, 497, 519, 526
Shadow effect, 483
Shadow printing, 407, 408, 424
Shallow trench, 514, 515, 529
Shallow trench isolation, 439
Sheet charge, 182, 183
Sheet resistance, 461, 485, 486, 492, 494, 497, 505, 506, 524, 528
Short-channel effects, 169, 199, 202
Si, 332, 334, 336, 342, 343, 345, 346, 355, 356, 361, 363, 368, 537, 538
Si/SiGe, 277–278, 280
Si/SiGe HBTs, 154
Si/SiGe material, 154
Si_3N_4, 482, 492
SiC, 288–289, 291–292, 537
Sidewall passivation, 441
SiGe MODFET, 250–251
Silane, 355, 357, 379, 380, 382, 384, 388, 389, 391, 392, 401, 402
Silane oxygen deposition, 380, 382
Silane reduction process, 392
Silicia, 19, 24, 25
Silicide, 395, 398–400
Silicide formation, 523–524
Silicide process, 524, 526
Silicide sintering, 399
Silicon (Si), 50, 54, 59, 60, 63, 75, 76, 79, 81, 82, 294, 314–318, 320, 325, 326, 328–330, 332– 339, 342–346, 348, 349, 355, 356, 361, 363, 365– 368, 441, 443, 445, 447, 449, 450, 454, 456, 458, 461–462, 464–469, 472–473, 475–479, 481, 483– 488, 537, 538
  APD, 318
  carbide, 424, 479
  crystal, 24–26, 29
  dioxide, 19, 33, 45, 86, 333, 346, 370–377, 379, 380, 382–383, 387, 392, 401–402, 468, 469, 476, 485
  etching, 428
  ingots, 335–337
  nitride, 370, 384, 399, 402, 430, 441, 454, 497, 505, 524
  properties, 538
  reduction process, 392
  tetrachloride, 355–356
  trench etching, 439
Silicone rubber membrane, 443, 444, 449
Silicon-on-glass, 495
Silicon-on-insulator (SOI), 210, 212–214, 220, 223, 485, 488
Silicon-on-nitride, 212
Silicon-on-oxide, 212
Silicon-on-sapphire, 495
Silicon-on-spinel, 212

Silver, 18
SIMOX, 485, 500
Simple cubic lattice, 21
Single drift IMPATT, 261–262
Single-electron memory cell (SEMC), 6, 7, 218
Single-frequency lasers, 308
Si-O bonds, 182
SiO_2, 450, 454, 482, 483, 485, 487, 488
SiO_2–Si MOS Diode, 179, 221
Si-Si bond , 182
Slurry, 397, 398
Small-scale integration (SSI), 490
Smart card, 6
Smart IC, 214
Sn, 346, 352, 361
Snell's law, 292
SnO_2, 325
Sodium, 21, 182
SOI (see silicon on insulator)
SOI devices, 212, 213
SOI integration, 525
SOI technology, 525, 528
Solar cell, 3, 4, 318–331
Solar constant, 319
Solar radiation, 318–320, 328, 329
Solar spectral irradiance, 319
Solid solubility, 392
Solid-phase epitaxy process, 479
Solubility, 351–352, 453
SOS, 212
Source, 186–190, 193–195, 197, 199–202, 205–206, 210–213, 216–220, 222
Space charge, 88–95, 100–102, 107, 120–121, 128
  density, 229
  neutrality, 94
  effect, 425
  region, 92, 229
Specific contact resistance, 224, 234–236
Spectrum splitting, 323, 324
Spiking, 392, 394, 395
Spinning micromotor, 443
Spiral inductor, 494, 495, 528
Spontaneous emission, 282–285, 302, 305
Spontaneous radiation, 288, 328
Spray etching, 427
Sputter etching, 432, 433
Sputtering, 390–392
Stacking fault, 349, 351, 364
Stainless steel, 479
Static random access memory (SRAM), 214, 215, 220, 507
Steel industry, 1
Step coverage, 379, 381, 382, 391, 392, 400
Step index fiber, 295–296
Step-and-repeat projection, 409
Step-and-scan projection lithography, 409
Stimulated emission, 299, 301
Stochastic space charge, 425, 426, 447
Stoichiometry, 481
Storage capacitor, 170, 214, 223
Storage cell, 214, 218
Storage time delay, 151
Strained layer, 365, 366
  epitaxy, 361–363, 366
  superlattice, 363–364
Stripe geometry laser, 308, 309
Strong inversion, 173, 175, 176, 178, 221
Submillimeter wave band, 255, 279
Substitutional sites, 349–351, 367, 478
Substrate bias, 187, 194, 198, 201, 220
Substrate damage, 437
Substrate doping, 173, 187, 192, 195, 200, 201, 220
Subthreshold characteristic, 194, 202–204, 210–211
Subthreshold current, 194, 202–204, 210–211
Subthreshold swing, 193, 202, 222
Superlattice, 361, 363, 364
Surface concentration, 456–458, 460, 461, 463–464, 466, 467, 469, 486, 487
Surface depletion layer, 173, 175
Surface depletion region, 175, 178, 188, 189, 199, 221, 223
Surface micromachining, 443, 444, 448
Surface orientation, 347, 349, 365
Surface potential, 173, 174, 178, 184, 193, 200, 201
Surface recombination, 65, 69, 70, 82, 153
Surface recombination velocity, 69, 82
Surface states, 65, 227, 252
Surface wave-coupled plasma (SWP) source, 437
Surface-emitting infrared InGaAsP LED, 296
Switched-mode power supplies (SMPS), 164–165
Switching, 116, 117, 127, 128, 130, 131, 146, 149–151, 155–157, 161, 164–166, 168
  circuit, 149, 150
  time, 131, 149
  transients, 149
Synchrotron, 423, 426, 444
  radiation, 423
  storage ring, 424
System-on-a-chip, 525, 526, 529

Ta_2O_5, 493, 494, 525
Tandem solar cell, 324–328
Tantalum, 424
Tantalum oxide (Ta_2O_5), 387
$TaSi_2$, 399
Taylor series, 243
Technology drivers, 11, 13
TED, 266–269, 279, 281
Television, 510, 525
Tellurium, 346, 454, 473
Temperature effect, 113, 306
Tensile stress, 384, 385
TEOS, 379–382, 384
  deposited SiO_2, 384
Ternary compound, 19, 20
Terrestrial energy source, 318, 328
Tetraethylorthosilicate, 379
Tetrahedron bond, 27
Thermal energy, 48, 57, 60, 83
Thermal equilibrium condition, 84, 88–91, 94–98, 105, 125–127, 129
Thermal oxidation, 369–371, 373, 400–401, 453, 467, 468, 478
Thermal oxide, 369, 379, 400
Thermal velocity, 48, 57, 64, 75, 83

Thermionic current, 273
Thermionic emission, 231, 232, 234, 235, 251, 275, 279
Thermionic emission process, 47, 72, 83
Thickness, 347, 349, 358, 360–363, 365, 366, 368
Thickness variation, 347, 349, 358
Thin film transistors (TFT), 209–212, 215, 220, 223
Thin oxide growth, 376
Threshold current density, 303, 305–306, 310, 328
Threshold energy, 413, 414
Threshold field, 265, 267
Threshold voltage, 169, 178, 190–200, 202, 211, 213, 214, 217, 218, 220–223, 243, 247–248, 251, 253, 469, 484, 485, 487, 488, 505, 509–511, 516, 522, 523, 528, 529
  adjustment, 469, 484, 487
  control, 194, 199
  shift, 199, 217, 223
Throughput, 407, 414, 418–420, 423, 426, 439, 447–449
Thyristor, 3, 4, 156–166, 168, 206, 342
THz radiation detection system, 273
Ti, 442
$TiCl_4$, 392
Tilt angle ion implantation, 483, 484
Tin, 441–443
TiO_2, 524, 525
$TiSi_2$, 399, 524
Titanium nitride (TiN), 395
TiW etch chamber, 439
Top-gate structure, 210, 211
Total capacitance, 176, 178
Total conduction current density, 59
Total internal reflection, 292
Trace impurity, 351
Transceiver, 517
Transconductance, 148, 191, 192, 202, 222, 242–245, 249, 279, 280
Transfer process, 404
Transferred electron devices, 254, 264, 267, 281
Transferred-electron diode, 6
Transformer-coupled plasma (TCP)source, 437
Transient behavior, 84, 114, 116, 127, 129
Transient time, 116, 117
Transistor action, 4, 130, 131, 134
Transit time, 149, 153, 155, 313–315, 331
Transit time domain mode, 255, 268, 269
Transmission coefficient, 74, 75, 83, 270–272, 543
Transmutation, 342
Transverse field, 188
Transverse magnetic mode, 257
Traps, 389
Tree of disorder, 477
Trench, 500, 501, 507, 513, 526, 528
  capacitor, 439
  isolation, 8, 11, 209, 221, 501, 512, 514, 515, 523, 527, 529
  type capacitor, 507, 526
  type DRAM, 514
Triac (triode ac switch), 162
Trichlorosilane ($SiHCl_3$), 24, 333, 355
Trimethylgallium, 361
Trimethylaluminum $Al(CH_3)_3$, 357, 358
Triode-configuration RIE etch, 436
Tris (8–hydroxy-quinolinato) aluminum, 294
Tuned cavity, 257
Tungsten, 21, 391, 441–443
  filaments, 479
  thermionic-emission cathode, 419
  -silicon Schottky diode, 230, 233
Tunnel diode, 3, 4, 254, 259, 260, 273, 278–280
Tunneling current, 235
Tunneling effect, 117–120
Tunneling phenomenon, 73
Tunneling probability, 235
Tunneling process, 217, 218
Turn-off time, 116, 117
Turn-off transient, 150, 151
Turn-on transient, 151
Twin tub, 512, 513, 529
Twin well, 206, 452
Two zone cathode contact, 267
Types of MOSFET, 187, 193, 195

ULSI, 490, 496, 503, 522, 527, 528
  chips, 490
  circuit, 385, 386, 387, 396, 400, 442, 503
Ultralarge scale integration (ULSI), 351, 377, 417, 431, 437
Ultrahigh vacuum system, 365
Ultrahigh-vacuum chamber, 358, 359
Ultrapure polycrystalline, 339
Ultrashallow Junction Formation, 524
Ultrasonic generations, 150
Ultrathin oxide, 489, 524
Ultraviolet, 282, 283, 288, 314, 319
U-MOSFET, 218
Unipolar resonant tunneling transistor, 254, 274, 279
Unit cell, 19–22, 45
Unit prefixes, 534
UV irradiation, 217
UV light, 416, 423

Vacancy, 349, 350, 367, 454–456, 462–465
  diffusion, 455, 456
  mechanism, 454, 463
Vacuum level, 124, 125, 170, 171, 179–181
Vacuum system, 225
Valence-band discontinuity, 152
Valence bond, 17, 27
Valence electrons, 27–29, 33, 37
Valley current, 259, 272, 273
Vapor pressure, 344, 345, 357, 366
Vapor-phase epitaxy, 354
Vapor-phase HF, 429
Varactor, 103, 104
Variable reactor, 103
Variable-shaped beam, 420
Vector scan, 419, 420
Velocity saturation region, 245
Vertical isolation, 504
Vertical oxidation furnace, 377

Via plug, 396
Video recording, 298
Viscosity, 415
Visible LED structures, 290, 293, 328
VLSI, 490, 496
V-MOSFET, 218
Volatile compound, 433
Volatile memories, 503
Volume charge density, 183, 184, 222
Volume defects, 349, 351, 365
V-shaped groove, 218

Wafer, 490, 522
Wafer shaping, 333, 347, 365
Water molecules, 374
Water vapor, 374, 376
Wavelength , 408
Weak inversion, 173
Wet chemical etching, 404, 426–428, 431, 432, 434, 443, 447, 450
Wet oxidation, 85, 86
WF_6, 391
White LED, 294, 328
Wireless-communication equipment, 517
Work function, 72, 83, 124, 175, 179, 180, 183, 187, 194, 195, 199, 221, 226, 227, 231, 252
Work function difference, 170, 171, 175, 179, 180, 183, 187, 199
WSi_2, 399

X-ray lithography (XRL), 424
X-ray radiation, 444
X-ray resist, 424, 444
Yellow light, 414
Yield, 411, 412, 434, 439, 449
Yield strength, 351, 367

Zinc (Zn), 17, 20, 32, 38, 346
Zinc diffusion, 465
Zincblend lattice, 22, 27
Zinc sulfide, 317
$ZnAs_2$, 454
Zn-Ga-As alloys, 454
Zn (see Zinc)
ZnO, 537
ZnS, 537
$ZnO\text{-}SiO_2$, 454
ZnSe, 291
Zone-refining technique, 341

This eagerly-anticipated revision offers more than 50% new or revised material that reflects the multitude of important recent discoveries and advances in device physics and integrated circuit processing.

The book offers a thorough introduction to physical principles of modern semiconductor devices and their fabrication technology. Readers are presented with theoretical and practical aspects of every step in device characterizations and fabrication, with an emphasis on integrated circuits.

The material is divided into three parts:

1. the basic properties of semiconductor materials, emphasizing silicon and gallium arsenide
2. the physics and characteristics of semiconductor devices bipolar, unipolar special microwave and photonic devices
3. the latest processing technologies, from crystal growth to lithographic pattern transfer

Each chapter is presented in a logical manner enabling readers to learn all important devices from a single source. Plus, the book covers historical developments of devices and technology in the last 100 years. Readers gain a sound perspective on the past and a foundation for projecting future trends.

ABOUT THE AUTHOR

S. M. Sze is UMC Chair Professor of the National Chiao Tung University and President of the National Nano Device Laboratories, Taiwan, R.O.C. For many years he was a member of the technical staff at Bell Laboratories. Professor Sze is the co-inventor of the nonvolatile semiconductor memory. He has written numerous texts on devices physics, including PHYSICS OF SEMICONDUCTOR DEVICES, considered a reference classic. In 1991, he received the IEEE J. J. Ebers award for his "fundamental and pioneering contributions..." He received his PhD in solid-state electronics from Stanford University in 1963.

ISBN 0-471-33372-7

X0046N7YN9
Semiconductor Devices: Physics and Technolog
New

JOHN WILEY & SONS, INC.
New York / Chichester
Weinheim / Brisbane
Singapore / Toronto
http://www.wiley.com/college/sze