2 Semiconductor Fabrication

Semiconductor devices have long been used in electronics. The first solid-state rectifiers were developed in the late nineteenth century. The galena crystal detector, invented in 1907, was widely used to construct crystal radio sets. By 1947, the physics of semiconductors was sufficiently understood to allow Bardeen and Brattain to construct the first bipolar junction transistor. In 1959, Kilby constructed the first integrated circuit, ushering in the era of modern semiconductor manufacture.

The impediments to manufacturing large quantities of reliable semiconductor devices were essentially technological, not scientific. The need for extraordinarily pure materials and precise dimensional control prevented early transistors and integrated circuits from reaching their full potential. The first devices were little more than laboratory curiosities. An entire new technology was required to mass produce them, and this technology is still rapidly evolving.

This chapter provides a brief overview of the process technologies currently used to manufacture integrated circuits. Chapter 3 then examines three representative process flows used for manufacturing specific types of analog integrated circuits.

2.1 SILICON MANUFACTURE

Integrated circuits are usually fabricated from *silicon*, a very common and widely distributed element. The mineral *quartz* consists entirely of silicon dioxide, also known as *silica*. Ordinary sand is chiefly composed of tiny grains of quartz and is therefore also mostly silica.

Despite the abundance of its compounds, elemental silicon does not occur naturally. The element can be artificially produced by heating silica and carbon in an electric furnace. The carbon unites with the oxygen contained in the silica, leaving more-or-less pure molten silicon. As this cools, numerous minute crystals form and grow together into a fine-grained gray solid. This form of silicon is said to be *polycrystalline* because it contains a multitude of crystals. Impurities and a disordered crystal structure make this *metallurgical-grade polysilicon* unsuited for semiconductor manufacture.

Metallurgical-grade silicon can be further refined to produce an extremely pure semiconductor-grade material. Purification begins with the conversion of the crude silicon into a volatile compound, usually trichlorosilane. After repeated distillation, the extremely pure trichlorosilane is reduced to elemental silicon using hydrogen gas. The final product is exceptionally pure, but still polycrystalline. Practical integrated circuits can only be fabricated from single-crystal material, so the next step consists of growing a suitable crystal.

2.1.1. Crystal Growth

The principles of crystal growing are both simple and familiar. Suppose a few crystals of sugar are added to a saturated solution that subsequently evaporates. The sugar crystals serve as seeds for the deposition of additional sugar molecules. Eventually the crystals grow to be very large. Crystal growth would occur even in the absence of a seed, but the product would consist of a welter of small intergrown crystals. The use of a seed allows the growth of larger, more perfect crystals by suppressing undesired nucleation sites.

In principle, silicon crystals can be grown in much the same manner as sugar crystals. In practice, no suitable solvent exists for silicon, and the crystals must be grown from the molten element at temperatures in excess of 1400°C. The resulting crystals are at least a meter in length and ten centimeters in diameter, and they must have a nearly perfect crystal structure if they are to be useful to the semiconductor industry. These requirements make the process technically challenging.

The usual method for growing semiconductor-grade silicon crystals is called the *Czochralski process*. This process, illustrated in Figure 2.1, uses a silica crucible charged with pieces of semi-grade polycrystalline silicon. An electric furnace raises the temperature of the crucible until all of the silicon melts. The temperature is then reduced slightly and a small seed crystal is lowered into the crucible. Controlled cooling of the melt causes layers of silicon atoms to deposit upon the seed crystal. The rod holding the seed slowly rises so that only the lower portion of the growing crystal remains in contact with the molten silicon. In this manner, a large silicon crystal can be pulled centimeter-by-centimeter from the melt. The shaft holding the crystal rotates slowly to ensure uniform growth. The high surface tension of molten silicon distorts the crystal into a cylindrical rod rather than the expected faceted prism.

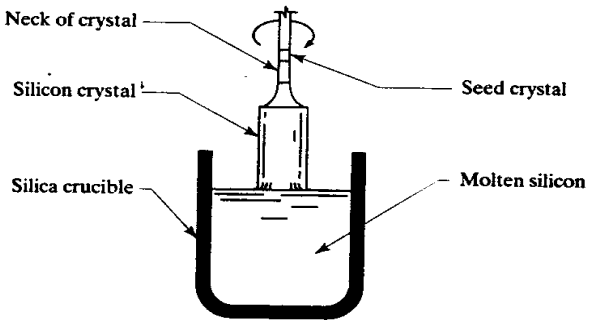


FIGURE 2.1 Czochralski process for growing silicon crystals.

The Czochralski process requires careful control to provide crystals of the desired purity and dimensions. Automated systems regulate the temperature of the melt and the rate of crystal growth. A small amount of doped polysilicon added to the melt sets the doping concentration in the crystal. In addition to the deliberately

introduced impurities, oxygen from the silica crucible and carbon from the heating elements dissolve in the molten silicon and become incorporated into the growing crystal. These impurities subtly influence the electrical properties of the resulting silicon. Once the crystal has reached its final dimensions, it is lifted from the melt and is allowed to slowly cool to room temperature. The resulting cylinder of monocrystalline silicon is called an *ingot.*

Since integrated circuits are formed upon the surface of a silicon crystal and penetrate this surface to no great depth, the ingot is customarily sliced into numerous thin circular sections called *wafers.* Each wafer yields hundreds or even thousands of integrated circuits. The larger the wafer, the more integrated circuits it holds and the greater the resulting economies of scale. Most modern processes employ either 150mm (6″) or 200mm (8″) wafers. A typical ingot measures between one and two meters in length and can provide hundreds of wafers.

2.1.2. Wafer Manufacturing

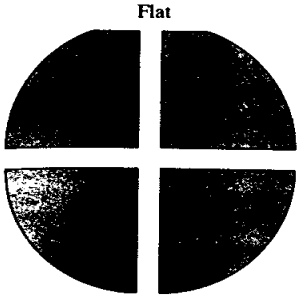
The manufacture of wafers consists of a series of mechanical processes. The two tapered ends of the ingot are sliced off and discarded. The remainder is then ground into a cylinder, the diameter of which determines the size of the resulting wafers. No visible indication of crystal orientation remains after grinding. The crystal orientation is experimentally determined and a flat stripe is ground along one side of the ingot. Each wafer cut from it will retain a facet, or *flat,* which unambiguously identifies its crystal orientation.

After grinding the flat, the manufacturer cuts the ingot into individual wafers using a diamond-tipped saw. In the process, about one-third of the precious silicon crystal is reduced to worthless dust. The surfaces of the resulting wafers bear scratches and pockmarks caused by the sawing process. Since the tiny dimensions of integrated circuits require extremely smooth surfaces, one side of each wafer must be polished. This process begins with mechanical abrasives and finishes with chemical milling. The resulting mirror-bright surface displays the dark gray color and characteristic near-metallic luster of silicon.

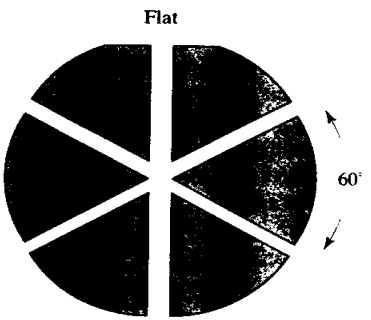
2.1.3. The Crystal Structure of Silicon

Each wafer constitutes a slice from a single silicon crystal. The underlying crystalline structure determines how the wafer splits when broken. Most crystals tend to part along *cleavage planes* where the interatomic bonding is weakest. For example, a diamond crystal can be cleaved by sharply striking it with a metal wedge. A properly oriented blow will split the diamond into two pieces, each of which displays a perfectly flat cleavage surface. If the blow is not properly oriented, then the diamond shatters. Silicon wafers also show characteristic cleavage patterns that can be demonstrated using a scrap wafer, a pad of note paper, and a wooden pencil. Place the wafer on the notepad, and place the pad in your lap. Take a wooden pencil and press down in the center of the wafer using the eraser. The wafer should split into either four or six regular wedge-shaped fragments, much like sections of a pie (Figure 2.2). The regularity of the cleavage pattern demonstrates that the wafer consists of monocrystalline silicon.

Figure 2.3 shows a small section of a silicon crystal drawn in three dimensions. Eighteen silicon atoms lie wholly or partially within the boundaries of an imaginary cube called a *unit cell.* Six of these occupy the centers of each of the six faces of the cube. Eight more atoms occupy the eight vertices of the cube. Two unit cells placed side-by-side share four vertex atoms and a single face-centered



(100) Wafer



(111) Wafer

FIGURE 2.2 Typical fracture patterns for (100) and (111) silicon wafers. Some wafers possess a second, smaller flat that denotes crystal orientation and doping. These *minor flats* have not been illustrated.

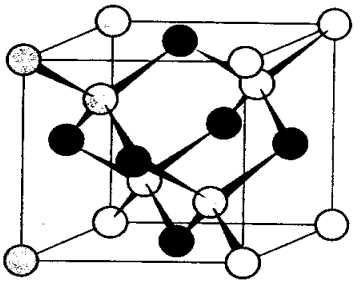


FIGURE 2.3 The diamond lattice unit cell displays a modified face-centered cubic structure. The face-centered atoms are shown in dark gray for emphasis.

atom. Additional unit cells can be placed on all sides to extend the crystal in all directions.

When the sawblade slices through a silicon ingot to form a wafer, the orientation of the resulting surface with respect to the unit cell determines many of the wafer's properties. A cut could, for example, slice across a face of the unit cell or diagonally through it. The pattern of atoms exposed by these two cuts differ, as do the electrical properties of devices formed into the respective surfaces. However, not all cuts made through a silicon crystal necessarily differ. Because the faces of a cube are indistinguishable from one another, a cut made across any face of the unit cell looks the same as cuts made across other faces. In other words, planes cut parallel to any face of a unit cube expose similar surfaces.

Because of the awkwardness of trying to describe various planes verbally, a trio of numbers called *Miller indices* are assigned to each possible plane passing through the crystal lattice (Appendix B). Figure 2.4 shows the two most important planar orientations. A plane parallel to a face of the cube is called a *(100) plane,* and a plane slicing diagonally through the unit cube to intersect three of its vertices is called a *(111) plane.* Silicon wafers are generally cut along either a (100) plane or a (111) plane. Although many other cuts exist, none of these have much commercial significance.

A trio of Miller indices enclosed in brackets denotes a direction perpendicular to the indicated crystal plane. For instance, a (100) plane has a [100] direction perpendicular to it and a (111) plane has a [111] direction perpendicular to it. Appendix B discusses how Miller indices are computed and explains the meaning of the different symbologies used to represent them.

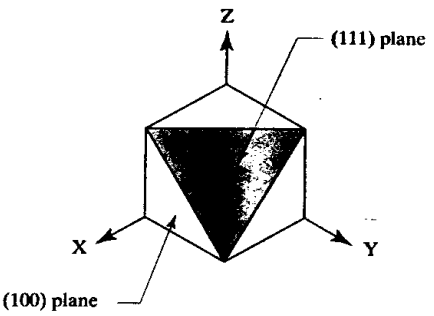


FIGURE 2.4 Identification of (100) and (111) planes of a cubic crystal.

2.2 PHOTOLITHOGRAPHY

The production of silicon wafers constitutes only the first step in the fabrication of integrated circuits. Many of the remaining steps deposit materials on the wafer or etch them away again. A variety of sophisticated deposition and etching techniques exist, but most of these are not *selective.* A nonselective, or *blanket,* process affects the entire surface of the wafer rather than just portions of it. The few processes that are selective are so slow or so expensive that they are useless for high-volume manufacturing. A technique called *photolithography* allows photographic reproduction of intricate patterns that can be used to selectively block depositions or etches. Integrated circuit fabrication makes extensive use of photolithography.

2.2.1. Photoresists

Photolithography begins with the application of a photosensitive emulsion called a *photoresist.* An image can be photographically transferred to the photoresist and a developer used to produce the desired masking pattern. The photoresist solution is usually *spun* onto the wafer. As shown in Figure 2.5, the wafer is mounted on a turntable spinning at several thousand revolutions per minute. A few drops of photoresist solution are allowed to fall onto the center of the spinning wafer, and centrifugal force spreads the liquid out across the surface. The photoresist solution adheres to the wafer and forms a uniform thin film. The excess solution flies off the edges of the spinning wafer. The film thins to its final thickness in a few seconds, the solvent rapidly evaporates, and a thin coating of photoresist remains on the wafer. This coating is baked to remove the last traces of solvent and to harden the photoresist to allow handling. Coated wafers are sensitive to certain wavelengths

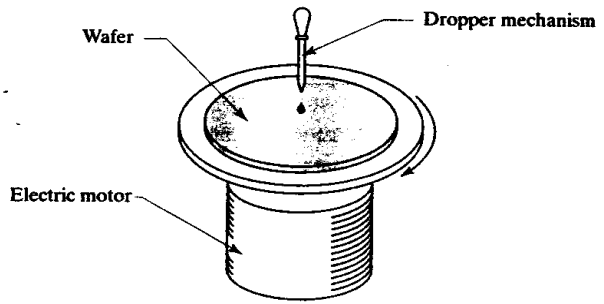


FIGURE 2.5 Application of photoresist solution to a wafer by spinning.

of light, particularly ultraviolet (UV) light. They remain relatively insensitive to other wavelengths, including those of red, orange, and yellow light. Most photolithography rooms therefore have special yellow lighting systems.

The two basic types of photoresists are distinguished by what chemical reactions occur during exposure. A *negative resist* polymerizes under UV light. The unexposed negative resist remains soluble in certain solvents, while the polymerized photoresist becomes insoluble. When the wafer is flooded with solvent, unexposed areas dissolve and exposed areas remain coated. A *positive resist,* on the other hand, chemically decomposes under UV light. These resists are normally insoluble in the developing solvent, but the exposed portions of the resist are chemically altered in order to become soluble. When the wafer is flooded with solvent, the exposed areas wash away while the unexposed areas remain coated. Negative resists tend to swell during development, so process engineers generally prefer to use positive resists.

2.2.2. Photomasks and Reticles

Modern photolithography depends upon a type of projection printing conceptually similar to that used to enlarge photographic negatives. Figure 2.6 shows a simplified illustration of the exposure process. A system of lenses collimates a powerful UV light source, and a plate called a *photomask* blocks the path of the resulting light beam. The UV light passes through the transparent portions of the photomask and through additional lenses that focus an image on the wafer. The apparatus in Figure 2.6 is called an *aligner* since it must ensure that the image of the mask aligns precisely with existing patterns on the wafer.

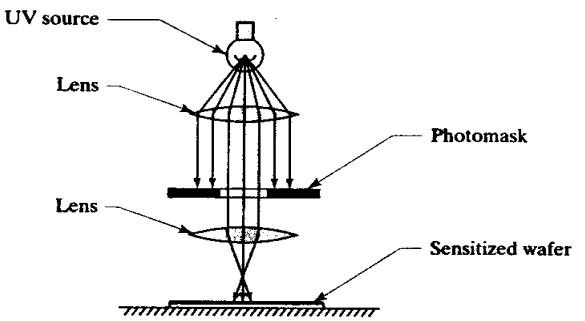


FIGURE 2.6 Simplified illustration of photomask exposure using an aligner.

The transparent plate used as the substrate of a photomask must be dimensionally stable or the pattern it projects will not align with those projected by previous masks. These plates most often consist of fused silica (often erroneously called quartz). After a thin layer of metal is applied to one surface of the plate, any one of various highly precise—but extremely slow and costly—methods are used to pattern the photomask. The image on the photomask is usually five or ten times the size of the image projected onto the wafer. Photographic reduction shrinks the size of any defects or irregularities in the photomask and therefore improves the quality of the final image. This type of enlarged photomask is called either a 5X or a 10X *reticle* depending on the degree of magnification employed.

A reticle can be used to directly pattern a wafer, but there are mechanical difficulties in doing so. The size of the photomask that an aligner can accept is limited by mechanical considerations, including the difficulty of constructing large lenses of the required accuracy. As a result, most commercial aligners accept a photomask about the same size as the wafer. A 5X reticle that could pattern an entire wafer in

one shot would be five times the size of the wafer, and would therefore not fit in the aligner. Practical 5X or 10X reticles are constructed to expose only a small rectangular portion of the final wafer pattern. The reticle must be stepped across the wafer and exposures made at many different positions in order to replicate the pattern across the entire wafer. This process is called *stepping*, and an aligner designed to step a reticle is called a *stepper*. Steppers are slower than ordinary aligners and are therefore more costly.

There is a faster method of exposing wafers that can be used for integrated circuits that do not require extremely fine feature sizes. The reticle can be stepped, not onto a sensitized wafer, but instead onto another photomask. This photomask now bears a 1X image of the desired pattern. The resulting photomask, called a *stepped working plate*, can expose an entire wafer in one shot. Stepped working plates make photolithography faster and cheaper, but the results are not as precise as directly stepping the reticle onto the wafer.

Even the tiniest dust speck is so large that it will block the transfer of a portion of the image and ruin at least one integrated circuit. Special air filtration techniques and protective garments are routinely used in wafer fabs, but some dust gets past all of these precautions. Photomasks are often equipped with *pellicles* on one or both sides to prevent dust from interfering with the exposure. Pellicles consist of thin transparent plastic films mounted on ring-shaped spacers that hold them slightly above the surface of the mask. Light passing through the plane of a pellicle is not in focus, so particles on the pellicle do not appear in the projected image. The pellicle also hermetically seals the surface of the mask and thereby protects it from dust.

2.2.3. Patterning

The exposed wafers are sprayed with a suitable developer, typically consisting of a mixture of organic solvents. The developer dissolves portions of the resist to uncover the surface of the wafer. A deposition or etch affects only these uncovered areas. Once the selective processing has been completed, the photoresist can be stripped away using solvents. Alternatively, the photoresist can be chemically destroyed by reactive ion etching in an oxygen ambient (Section 2.3.2). This procedure is called *ashing*.

Many important fabrication processes require masking layers that can withstand high temperatures. Since most practical photoresists are organic compounds, they are clearly unsuited to this task. Two common high-temperature masking materials are silicon dioxide and silicon nitride. These materials can be formed by the reaction of appropriate gases with the silicon surface. A photoresist can then be applied and patterned and an etching process used to open holes in the oxide or nitride film. Modern processing techniques make extensive use of oxide and nitride films for masking high-temperature depositions and diffusions.

2.3 OXIDE GROWTH AND REMOVAL

Silicon forms several oxides, the most important of which is *silicon dioxide* (SiO_2). This oxide possesses a number of desirable properties that together are so valuable that silicon has become the dominant semiconductor. Other semiconductors have better electrical properties, but only silicon forms a well-behaved oxide. Silicon dioxide can be grown on a silicon wafer by simply heating it in an oxidizing atmosphere. The resulting film is mechanically rugged and resists most common solvents, yet it readily dissolves in hydrofluoric acid. Oxide films are superb electrical insula-

tors and are useful not only for insulating metal conductor patterns but also for forming the dielectrics of capacitors and MOS transistors. Silicon dioxide is so important to silicon processing that it is universally known as *oxide.*

2.3.1. Oxide Growth and Deposition

The simplest method of producing an oxide layer consists of heating a silicon wafer in an oxidizing atmosphere. If pure dry oxygen is employed, then the resulting oxide film is called a *dry oxide.* Figure 2.7 shows a typical oxidation apparatus. The wafers are placed in a fused silica rack called a *wafer boat.* The wafer boat is slowly inserted into a fused silica tube wrapped in an electrical heating mantle. The temperature of the wafers gradually rises as the wafer boat moves into the middle of the heating zone. Oxygen gas blowing through the tube passes over the surface of each wafer. At elevated temperatures, oxygen molecules can actually diffuse through the oxide layer to reach the underlying silicon. There oxygen and silicon react, and the layer of oxide gradually grows thicker. The rate of oxygen diffusion slows as the oxide film thickens, so the growth rate decreases with time. As Table 2.1 indicates, high temperatures greatly accelerate oxide growth. Crystal orientation also affects oxidation rates, with (111) silicon oxidizing significantly faster than (100) silicon.[1] Once the oxide layer has reached the desired thickness (as gauged by time and temperature), the wafers are slowly withdrawn from the furnace.

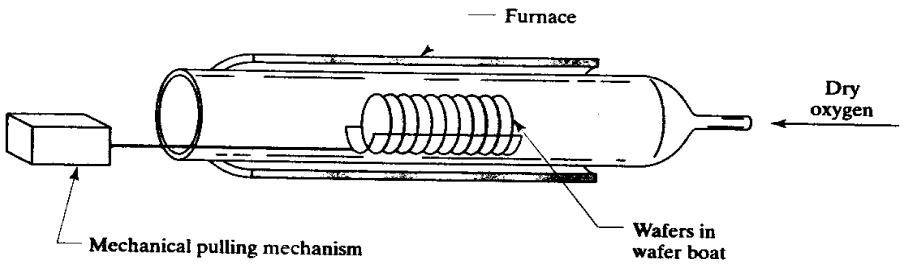


FIGURE 2.7 Simplified diagram of an oxidation furnace.

Ambient	800°C	900°C	1000°C	1100°C	1200°C
Dry O_2	30 hr	6 hr	1.7 hr	40 min	15 min
Wet O_2	1.7 hr	20 min	6 min		

TABLE 2.1 Times required to grow 0.1 μm of oxide on (111) silicon.[2]

Dry oxide grows very slowly, but it is of particularly high quality because relatively few defects exist at the oxide-silicon interface. These defects, or *surface states,* interfere with the proper operation of semiconductor devices, particularly MOS transistors. The density of surface states is measured by a parameter called the *surface state charge,* or Q_{ss}. Dry oxide films that are thermally grown on (100) silicon have especially low surface state charges and thus make ideal dielectrics for MOS transistors.

[1] W. R. Runyan and K. E. Bean, *Semiconductor Integrated Circuit Processing Technology* (Reading, MA: Addison-Wesley, 1994), p. 84ff.

[2] Calculated from R. P. Donovan, "Oxidation," in R. M. Burger and R. P. Donovan, eds., *Fundamentals of Silicon Integrated Device Technology* (Englewood Cliffs, NJ: Prentice-Hall, 1967), pp. 41, 49.

Wet oxides are formed in the same way as dry oxides, but steam is injected into the furnace tube to accelerate the oxidation. Water vapor moves rapidly through oxide films, but hydrogen atoms liberated by the decomposition of the water molecules produce imperfections that may degrade the oxide quality.[3] Wet oxidation is commonly used to grow a thick layer of *field oxide* where no active devices will be built. Dry oxidations conducted at higher-than-ambient pressures can also accelerate oxide growth rates.

Sometimes an oxide layer must be formed on a material other than silicon. For instance, oxide is frequently employed as an insulator between two layers of metallization. In such cases, some form of deposited oxide must be used rather than the grown oxides previously discussed. Deposited oxides can be produced by various reactions between gaseous silicon compounds and gaseous oxidizers. For example, silane gas and nitrous oxide react to form nitrogen gas, water vapor, and silicon dioxide. Deposited oxides tend to possess low densities and large numbers of defect sites, so they are not suitable for use as gate dielectrics for MOS transistors. Deposited oxides are still acceptable for use as insulating layers between multiple conductor layers, or as protective overcoats.

Oxide films are brightly colored due to *thin-film interference*. When light passes through a transparent film, destructive interference between transmitted and reflected wavefronts causes certain wavelengths of light to be selectively absorbed. Different thicknesses of films absorb different colors of light. Thin-film interference causes the iridescent colors seen in soap bubbles and films of oil on water. The same effect produces the vivid colors visible in microphotographs of integrated circuits. These colors are helpful in distinguishing various regions of an integrated circuit under a microscope or in a microphotograph. The approximate thickness of an oxide film can often be determined using a table of oxide colors.[4]

2.3.2. Oxide Removal

Figure 2.8 illustrates the procedure used to form a patterned oxide layer. The first step consists of growing a thin layer of oxide across the wafer. Next, photoresist is applied to the wafer by spinning. A subsequent oven bake drives off the final traces

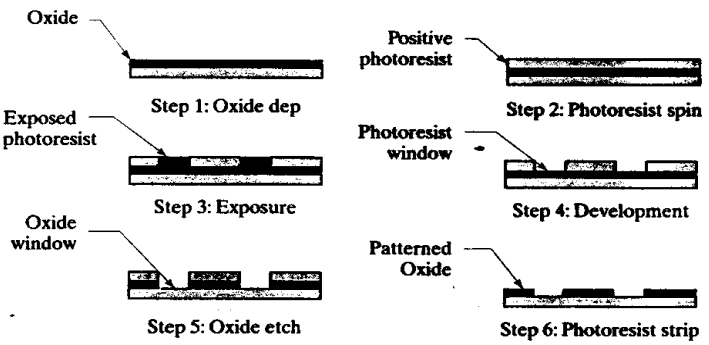


FIGURE 2.8 Steps in oxide growth and removal.

[3] Hydrogen incorporation due to wet oxidation conditions reduces the concentration of dangling bonds, but it increases the fixed oxide charge. The differences between wet and dry oxidation are therefore not as simplistic as the text may suggest.

[4] For a table, see W. A. Pliskin and E. E. Conrad, "Nondestructive Determination of Thickness and Refractive Index of Transparent Films," *IBM J. Research and Development*, Vol. 8, 1964, pp. 43–51.

of solvent and hardens the photoresist for handling. After photolithographic exposure, the wafer is developed by spraying it with a solvent that dissolves the exposed areas of photoresist to reveal the underlying oxide. The patterned photoresist serves as a masking material for an oxide etch. Having served its function, the photoresist is finally stripped away to leave the patterned oxide layer.

Oxide can be etched by either of two methods. *Wet etching* employs a liquid solution that dissolves the oxide, but not the photoresist or the underlying silicon. *Dry etching* uses a reactive plasma to perform the same function. Wet etches are simpler, but dry etches provide better linewidth control.

Most wet etches employ solutions of buffered hydrofluoric acid (HF). This highly corrosive substance readily dissolves silicon dioxide, but it does not attack either elemental silicon or organic photoresists. The etch process consists of immersing the wafers in a plastic tank containing the hydrofluoric acid solution for a specified length of time, followed by a thorough rinsing to remove all traces of the acid. Wet etches are *isotropic* because they proceed at the same rate laterally as well as vertically. The acid works its way under the edges of the photoresist to produce sloping sidewalls similar to those shown in Figure 2.9A. Since the etching must continue long enough to ensure that all openings have completely cleared, some degree of overetching inevitably occurs. The acid continues to erode the sidewalls as long as the wafer remains immersed. The extent of sidewall erosion varies depending upon etching conditions, oxide thickness, and other factors. Because of these variations, wet etching cannot provide the tight linewidth control required by modern semiconductor processes.

There are several types of dry etching processes.[5] One called *reactive ion etching* (RIE) employs plasma bombardment to erode the surface of the wafer. A silent electrical discharge passed through a low-pressure gas mixture forms highly energetic molecular fragments called *reactive ions*. The etching apparatus projects these ions downward onto the wafer at high velocities. Because the ions impact the wafer at a relatively steep angle, etching proceeds vertically at a much greater rate than laterally. The *anisotropic* nature of reactive ion etching allows the formation of nearly vertical sidewalls such as those shown in Figure 2.9B. Figure 2.10 shows a simplified diagram of a reactive ion etching apparatus.

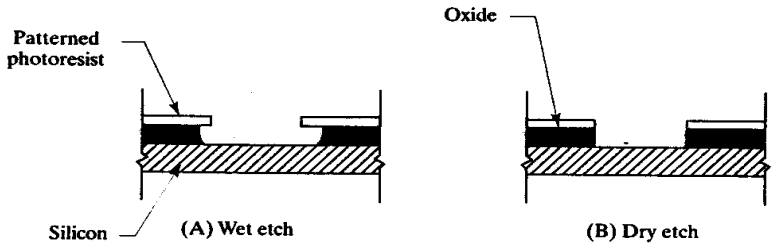


FIGURE 2.9 Comparison of isotropic wet etching (A) and anisotropic dry etching (B). Note the undercutting of the oxide caused by wet etching.

The etch gas employed in the RIE system generally consists of an organohalogen compound such as trichloroethane, perhaps mixed with an inert gas such as argon. The reactive ions formed from this mixture selectively attack silicon dioxide in preference to either photoresist or elemental silicon. Different mixtures of etch gases

[5] Reactive ion etching is actually only one of three forms of dry etching, the other two being plasma etching and chemical vapor etching. RIE is among the most useful because it produces highly anisotropic etching characteristics. See Runyan, *et al.*, pp. 269–272.

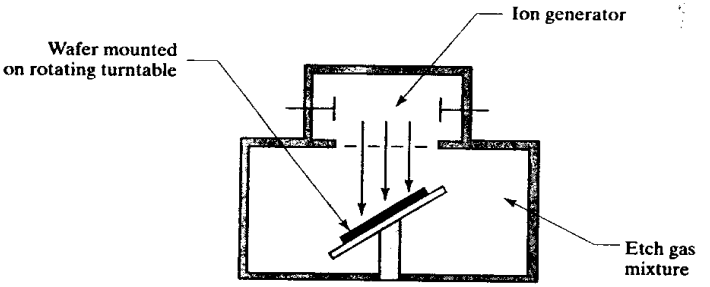


FIGURE 2.10 Simplified diagram of reactive ion etching apparatus.

have been developed that allow anisotropic etching of silicon nitride, elemental silicon, and other materials.

Modern processes rely on dry etching to obtain tight control of submicron geometries that cannot be fabricated in any other way. The increased packing density and higher performance of these structures more than compensate for the complexity and cost of dry etching.

2.3.3. Other Effects of Oxide Growth and Removal

During a typical processing sequence, the wafer is repeatedly oxidized and etched to form successive masking layers. These multiple masked oxidations cause the silicon surface to become highly nonplanar. The resulting surface irregularities are of great concern because modern fine-line photolithography has a very narrow depth of field. If the surface irregularities are too large, then it becomes impossible to focus the image of the photomask onto the resist.

Consider the wafer in Figure 2.11. A planar silicon surface has been oxidized, patterned, and etched to form a series of oxide openings (Figure 2.11A). Subsequent thermal oxidation of the patterned wafer results in the cross-section shown in Figure 2.11B. The opening that is left from the previous oxide removal initially oxidizes very rapidly, while the surfaces already coated with an oxide layer oxidize more slowly. The silicon surface erodes by about 45% of the oxide thickness grown.[6] The silicon under the previous oxide opening therefore recedes to a greater depth than the surrounding silicon surfaces. The thickness of oxide in the old opening will always be less than that of the surrounding surfaces since these already have some oxide on them when growth begins. The differences in oxide thickness and in the depths of the silicon surfaces combine to produce a characteristic surface discontinuity called an *oxide step.*

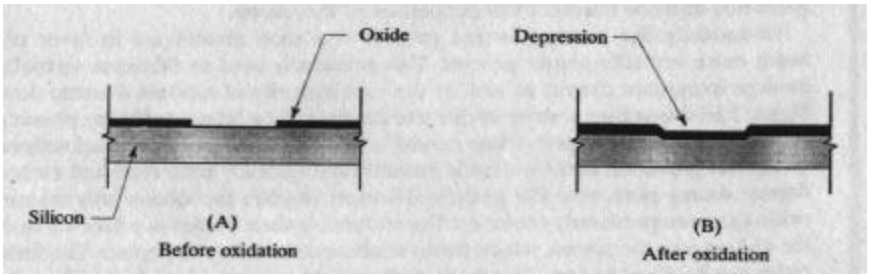


FIGURE 2.11 Effects of patterned oxidation on wafer topography.

[6] This value is the inverse of the *Pilling-Bedworth ratio,* which equals 2.2: G. E. Anner, *Planar Processing Primer* (New York: Van Nostrand Reinhold, 1990), p. 169.

The growth of a thermal oxide also affects the doping levels in the underlying silicon. If the dopant is more soluble in oxide than in silicon, during the course of the oxidation it will tend to migrate from the silicon into the oxide. The surface of the silicon thus becomes depleted of dopant. Boron is more soluble in oxide than in silicon, so it tends to segregate into the oxide. This effect is sometimes called *boron suckup.* Conversely, if the dopant dissolves more readily in silicon than in oxide, then the advancing oxide-silicon interface pushes the dopant ahead of it and causes a localized increase in doping levels near the surface. Phosphorus (like arsenic and antimony) segregates into the silicon, so it tends to accumulate at the surface as oxidation continues. This effect is sometimes called *phosphorus pileup* or *phosphorus plow.* The doping profiles of Figures 2.12A and 2.12B illustrate boron suckup and phosphorus plow, respectively. In both cases, the pre-oxidization doping profiles were constant and the varying dopant concentrations near the surface are solely due to segregation. The existence of these segregation mechanisms complicates the task of designing dopant profiles for integrated devices.

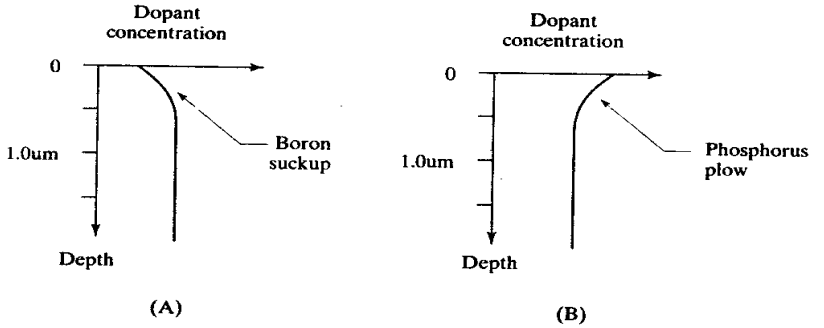


FIGURE 2.12 Oxide segregation mechanisms: (A) boron suckup and (B) phosphorus plow.[7]

The doping of silicon also affects the rate of oxide growth. A concentrated N+ diffusion tends to accelerate the growth of oxide near it by a process called *dopant-enhanced oxidation.* This occurs because the donors interfere with the bonding of atoms at the oxide interface, causing dislocations and other lattice defects. These defects catalyze oxidation and thus accelerate the growth of the overlying oxide. This effect can become quite significant when a heavily doped N+ deposition occurs early in the process, before the long thermal drives and oxidations. Figure 2.13 shows a wafer in which a long thermal oxidation has been

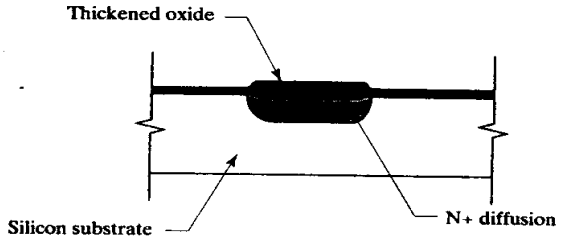


FIGURE 2.13 Effects of dopant-enhanced oxidation.

[7] A. S. Grove, O. Leistiko, and C. T. Sah, "Redistribution of Acceptor and Donor Impurities During Thermal Oxidation of Silicon," *J. Appl. Phys.*, Vol. 35, #9, 1964, pp. 2695–2701.

conducted after the deposition of an N+ region. The oxide over the N+ diffusion is actually thicker than the oxide over adjacent regions. Dopant-enhanced oxidation can be used to thicken the field oxide in order to reduce its capacitance per unit area. Thus, a capacitor formed over a deep-N+ diffusion will exhibit less parasitic capacitance between its bottom plate and the substrate than will a capacitor formed over lightly doped regions.

2.3.4. Local Oxidation of Silicon (LOCOS)

A technique called *local oxidation of silicon* (LOCOS) allows the selective growth of thick oxide layers.[8] The process begins with the growth of a thin *pad oxide* that protects the silicon surface from the mechanical stresses induced by subsequent processing (Figure 2.14). Chemical vapor deposition produces a nitride film on top of the pad oxide. This nitride is patterned to expose the regions to be selectively oxidized. The nitride blocks the diffusion of oxygen and water molecules, so oxidation only occurs under the nitride windows. Some oxidants diffuse a short distance under the edges of the nitride, producing a characteristic curved transition region called a *bird's beak*.[9] Once oxidation is complete, the nitride layer is stripped away to reveal the patterned oxide.

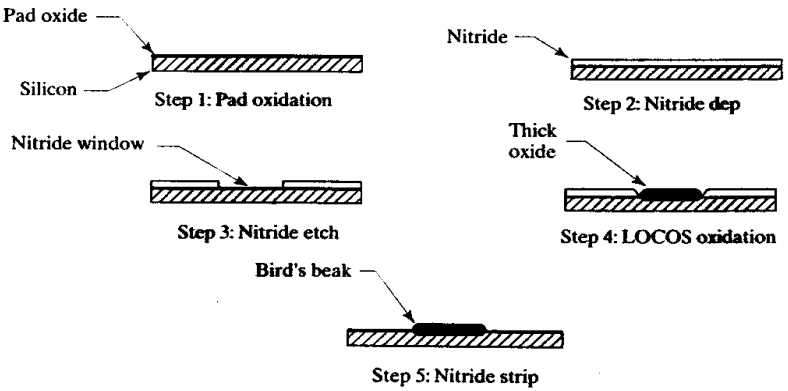


FIGURE 2.14 Local oxidation of silicon (LOCOS) process.

CMOS and BiCMOS processes employ LOCOS to grow a thick *field oxide* over electrically inactive regions of the wafer. The areas not covered by field oxide are called *moat* regions because they form shallow trenches in the topography of the wafer. A very thin, high-quality gate oxide subsequently grown in the moat regions forms the gate dielectric of the MOS transistors.

A mechanism called the *Kooi effect* complicates the growth of gate oxide.[10] The water vapor typically used to accelerate LOCOS oxidation also attacks the surface of the nitride film to produce ammonia, some of which migrates beneath the pad oxide near the edges of the nitride window. There it reacts with the underlying silicon to form silicon nitride again (Figure 2.15). Since these nitride deposits lie

---

[8] "LOCOS: A New I.C. Technology," *Microelectronics and Reliability*, Vol. 10, 1971, pp. 471–472.

[9] E. Bassous, H. N. Yu, and V. Maniscalco, "Topology of Silicon Structures with Recessed SiO_2," *J. Electrochem. Soc.*, Vol. 123, #11, 1976, pp. 1729–1737.

[10] E. Kooi, J. G. van Lierop, and J. A. Appels, "Formation of Silicon Nitride at a Si-SiO_2 Interface during Local Oxidation of Silicon and during Heat-Treatment of Oxidized Silicon in NH_3 Gas," *J. Electrochem. Soc.*, Vol. 123, #7, 1976, pp. 1117–1120.

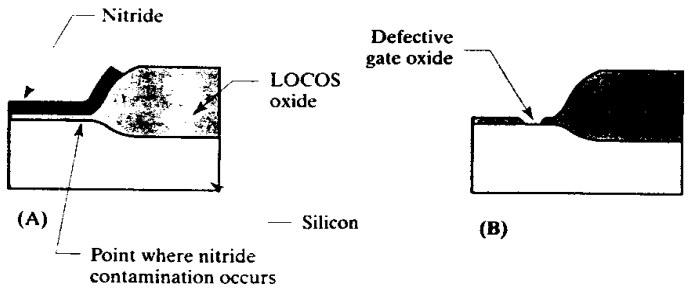


FIGURE 2.15 The Kooi effect is caused by nitride that grows under the bird's beak (A), preventing formation of gate oxide during subsequent oxidation (B).

beneath the pad oxide, they remain even after the LOCOS nitride is stripped. Removing the pad oxide prior to growing the gate oxide does not eliminate these deposits because this etch is selective to oxide, not to nitride. During gate oxidation, the nitride residues act as an unintentional LOCOS mask that retards oxide growth around the edges of the moat region. The gate oxide at these points may not be sufficiently thick to withstand the full operating voltage. The Kooi effect can be circumvented by first growing a thin oxide layer and then stripping it away. Because silicon nitride slowly oxidizes, this *dummy gate oxidation* removes the nitride residues and improves the integrity of the true gate oxide grown immediately afterward.

2.4 DIFFUSION AND ION IMPLANTATION

Discrete diodes and transistors can be fabricated by forming junctions into a silicon ingot during crystal growth. Suppose that the silicon ingot begins as a P-type crystal. After a short period of growth, the melt is counterdoped by the addition of a controlled amount of phosphorus. Continued crystal growth will now produce a PN junction embedded in the ingot. Successive counterdopings can produce multiple junctions in the crystal, allowing the fabrication of *grown-junction* transistors. Integrated circuits cannot be grown because there is no way to produce differently doped regions in different portions of the wafer. Even the manufacture of simple grown-junction transistors presents a challenge, because the thickness and planarity of grown junctions are difficult to control. Each counterdoping also raises the total dopant concentration. Some properties of silicon (such as minority carrier lifetime) depend upon the total concentration of doping atoms, not just upon the excess of one dopant species over the other. The repeated counterdopings therefore progressively degrade the electrical properties of the silicon.

Historically, the grown junction process was soon abandoned in favor of the much more versatile *planar process*. This process is used to fabricate virtually all modern integrated circuits as well as the vast majority of modern discrete devices. Figure 2.16 shows how a wafer of discrete diodes can be fabricated using planar processing. A uniformly doped silicon crystal is first sliced to form individual wafers. An oxide film grown on these wafers is photolithographically patterned and etched. A dopant source spun onto the patterned wafers touches the silicon only where the oxide has been previously removed. The wafers are then heated in a furnace to drive the dopant into the silicon, which forms shallow counterdoped regions. The finished wafer can be diced to form hundreds or thousands of individual diodes. The planar process does not require multiple counterdopings of the silicon ingot, thereby allowing more precise control of junction depths and dopant distributions.

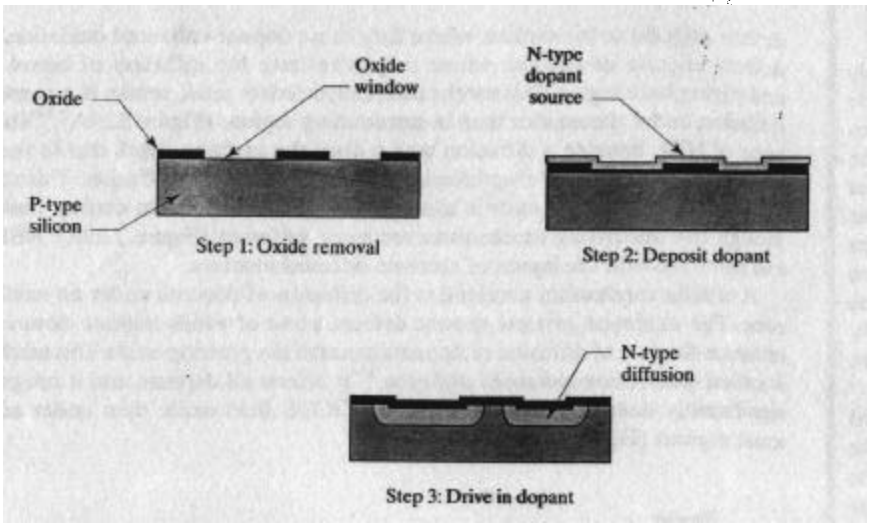


FIGURE 2.16 Formation of diffused PN-junction diodes using the planar process.

2.4.1. Diffusion

Dopant atoms can move through the silicon lattice by thermal diffusion in much the same way as carriers move by diffusion (Section 1.1.3). The heavier dopant atoms are more tightly bound to the crystal lattice, so temperatures of 800°C to 1250°C are required to obtain reasonable diffusion rates. Once the dopants have been driven to the desired junction depth, the wafer is cooled and the dopant atoms become immobilized within the lattice. A doped region formed in this manner is called a *diffusion.*

The usual process for creating a diffusion consists of two steps: an initial *deposition* (or *predeposition*) and a subsequent *drive* (or *drive-in*). Deposition consists of heating the wafer in contact with an external source of dopant atoms. Some of these diffuse from the source into the surface of the silicon wafer to form a shallow heavily doped region. The external dopant source is then removed and the wafer is heated to a higher temperature for a prolonged period of time. The dopants introduced during deposition are now driven down to form a much deeper and less concentrated diffusion. If a very heavily doped junction is required, then it is usually unnecessary to strip the dopant source from the wafer, and the deposition and subsequent drive can be conducted as a single operation.

Four dopants find widespread use in silicon processing: *boron, phosphorus, arsenic,* and *antimony.*[11] Only boron is an acceptor; the other three are all donors. Boron and phosphorus diffuse relatively rapidly, while arsenic and antimony diffuse much more slowly (Table 2.2). Arsenic and antimony are used where slow rates of

TABLE 2.2 Representative junction depths, in microns (10^{20} atoms/cm^3 source, 10^{16} atoms/cm^3 background, 15 min deposition, 1 hr drive).[12]

Dopant	950°C	1000°C	1100°C	1200°C
Boron	0.9	1.5	3.6	7.3
Phosphorus		0.5	1.6	4.6
Antimony			0.8	2.1
Arsenic			0.7	2.0

[11] These dopants were chosen because they readily ionize and because they are sufficiently soluble in silicon to form heavily doped diffusions. See F. A. Trumbore, "Solid Solubilities of Impurity Elements in Germanium and Silicon," *Bell Syst. Tech. J.*, Vol. 39, #1, 1960, pp. 205–233.

[12] Calculated using diffusivities from R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits,* 2nd ed. (New York: John Wiley and Sons, 1986), p. 85.

diffusion are advantageous—for example, when very shallow junctions are desired. Even boron and phosphorus do not diffuse appreciably at temperatures below 800°C, necessitating the use of special high-temperature diffusion furnaces.

Figure 2.17 shows a simplified diagram of a typical apparatus for conducting a phosphorus diffusion. A long fused silica tube passes through an electric furnace that is constructed to produce a very stable heating zone in the middle of the tube. After the wafers are loaded into a wafer boat, they are slowly pushed into the furnace by means of a mechanical arrangement that controls the insertion rate. Dry oxygen is blown through a flask containing liquid phosphorus oxychloride ($POCl_3$, often called "pockle"). A small amount of $POCl_3$ evaporates and is carried by the gas stream over the wafers. Phosphorus atoms released by the decomposition of the $POCl_3$ diffuse into the oxide film, forming a doped oxide that acts as a deposition source. When enough time has passed to deposit sufficient dopant in the silicon, the wafers are removed from the furnace and the doped oxide is stripped away (a process called *deglazing*). The wafers are then reloaded into another furnace, where they are heated to drive the phosphorus down to form the desired diffusion. If a very concentrated phosphorus diffusion is desired, then the wafers need not be removed for deglazing prior to the drive. With suitable modifications to the dopant source, this apparatus can diffuse any of the four common dopants.

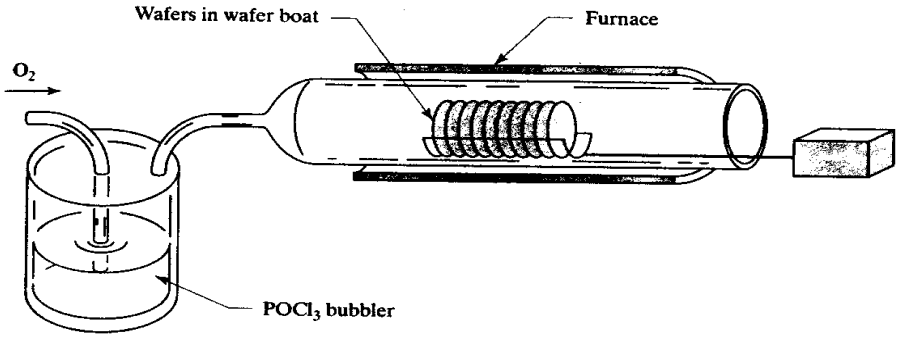


FIGURE 2.17 Simplified diagram of a phosphorus diffusion furnace using a $POCl_3$ source.

Many alternate deposition sources have been developed. A gaseous dopant such as diborane (for boron) or phosphine (for phosphorus) can be injected directly into the carrier gas stream. Thin disks of boron nitride placed between silicon wafers can serve as a solid deposition source for boron. In a high-temperature oxidizing atmosphere, a little boron trioxide outgases from these disks to the adjacent wafers. Various proprietary *spin-on glasses* are also sold as dopant sources. These consist of doped oxide dispersed in a volatile solvent. After the solution is spun onto a wafer, a brief bake drives out the solvent and leaves a doped oxide layer on the wafer. This so-called *glass* then serves as a dopant source for the subsequent diffusion.

None of these deposition schemes are particularly well controlled. Even with gaseous sources (which can be precisely metered) nonuniform gas flow around the wafer inevitably produces doping variations. For less-demanding processes such as standard bipolar, any of these schemes can give adequate results. Modern CMOS and BiCMOS processes require more accurate control of doping levels and junction depths than conventional deposition techniques can achieve. Ion implantation can provide the necessary accuracy at the expense of much more complex and costly apparatus.

2.4.2. Other Effects of Diffusion

The diffusion process suffers from a number of limitations. Diffusions can only be performed from the surface of the wafer, limiting the geometries that can be fabricated. Dopants diffuse unevenly, so the resulting diffusions do not have constant doping profiles. Subsequent high-temperature process steps continue the drive of previously deposited dopants, so junctions formed early in the process are driven substantially deeper during later processing. Dopants out-diffuse under the edges of the oxide windows, spreading the diffusion pattern. Diffusions interact with oxidizations due to segregation mechanisms, resulting in depletion or enhancement of surface doping levels. Diffusions even interact with one another since the presence of one doping species alters the diffusion rates of others. These and other complications make the diffusion process far more complex than it might at first appear.

Diffusion can produce only relatively shallow junctions. Practical drive times and temperatures limit junction depths to about fifteen microns. Most diffusions will be much shallower. Since diffusions are typically patterned using an oxide mask, the cross section of a diffusion generally resembles that shown in Figure 2.18. The dopant diffuses out in all directions at roughly the same rate. The junction moves laterally under the edges of the oxide window a distance equal to about 80% of the junction depth.[13] This lateral movement, known as *outdiffusion*, causes the final size of the diffused region to exceed the drawn dimensions of the oxide window. Outdiffusion is not visible under the microscope since the changes in oxide color caused by thin film interference correspond to the locations of oxide removals and not to the positions of the final junctions.

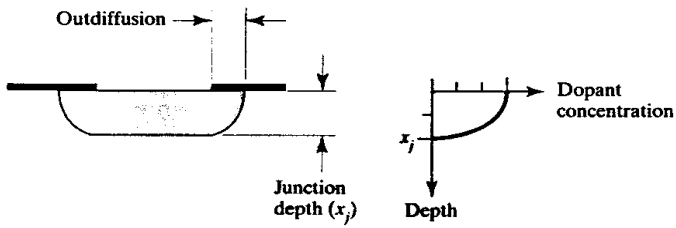


FIGURE 2.18 Cross section and doping profile of a typical planar diffusion.

The doping level of a diffusion varies as a function of depth. Neglecting segregation mechanisms, dopant concentrations are highest at the surface and gradually lessen with depth. The resulting ***doping profile*** can be theoretically predicted and experimentally measured. Figure 2.18 shows the theoretical doping profile for a point in the center of the oxide window. This profile assumes that oxide segregation remains negligible, which is not always the case. Boron suckup may substantially reduce the surface doping of a P-type diffusion and can cause a lightly doped diffusion to invert to become N-type. Phosphorus pileup will not cause surface inversion, but it still affects surface doping levels.

As mentioned above, the rate of diffusion can be altered by the presence of other doping species. Consider an NPN transistor with a heavily doped phosphorus emitter diffused into a lightly doped boron base. The presence of high concentrations of donors within the emitter causes lattice strains that spawn defects. Some of these

[13] See 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. of Research and Development*, Vol. 9, 1965, pp. 179–186.

defects migrate to the surface, where they cause dopant-enhanced oxidation. Other defects migrate downward, where they accelerate the diffusion of boron in the underlying base region. This mechanism, called *emitter push*, results in a deeper base diffusion under the emitter than in surrounding regions (Figure 2.19A).[14] The presence of NBL beneath a diffusion may reduce the junction depth due to the intersection of the tail of the updiffusing NBL with the base diffusion. This effect is sometimes called *NBL push* in analogy with the better-known emitter push, even though the underlying mechanisms are quite different (Figure 2.19B). NBL push can interfere with the layout of accurate diffused resistors.

A similar mechanism accelerates the diffusion of dopants under an oxidization zone. The oxidation process spawns defects, some of which migrate downward to enhance the rate of diffusion of dopants beneath the growing oxide. This mechanism is called *oxidization-enhanced diffusion*.[15] It affects all dopants, and it can produce significantly deeper diffusions under a LOCOS field oxide than under adjacent moat regions (Figure 2.19C).

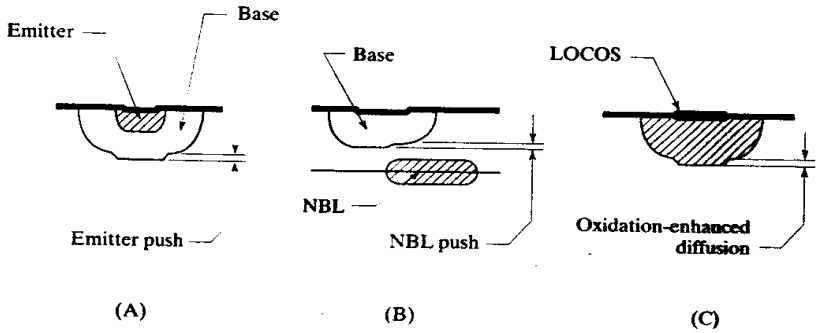


FIGURE 2.19 Mechanisms that can alter diffusion rates include emitter push (A), NBL push (B), and oxidation-enhanced diffusion (C).

Even the most sophisticated computer programs cannot always predict actual doping profiles and junction depths because of the many interactions that occur. Process engineers must experiment carefully to find the proper recipe for manufacturing a given combination of devices on a wafer. The more complicated the process, the more complex these interactions become and the more difficult it is to find a suitable recipe. Since process design takes so much time and effort, most companies use only a few processes to manufacture all of their products. The difficulty of incorporating new process steps into an existing recipe also explains the reluctance of process engineers to modify their processes.

2.4.3. Ion Implantation

Due to the limitations of conventional diffusion techniques, modern processes make extensive use of *ion implantation*. An ion implanter is essentially a specialized particle accelerator used to accelerate dopant atoms so that they can penetrate the silicon crystal to a depth of several microns.[16] Ion implantation does not require high

[14] A. F. W. Willoughby, "Interactions between Sequential Dopant Diffusions in Silicon—A Review," *J. Phys. D: Appl. Phys.*, Vol. 10, 1977, pp. 455–480.

[15] K. Taniguchi, K. Kurosawa, and M. Kashiwagi, "Oxidation Enhanced Diffusion of Boron and Phosphorus in (100) Silicon," *J. Electrochem. Soc.*, Vol. 127, #10, 1980, p. 2243–2248.

[16] The depth of the implant depends on implant energy. The implants discussed in this section involve energies of no more than several hundred keV. Some modern CMOS processes now employ multi-MeV implants to achieve significantly deeper profiles (5–10μm).

temperatures, so a layer of patterned photoresist can serve as a mask against the implanted dopants. Implantation also allows better control of dopant concentrations and profiles than conventional deposition and diffusion. However, large implant doses require correspondingly long implant times. Ion implanters are also complex and costly devices. Many processes use a combination of diffusions and implantations to reduce overall costs.

Figure 2.20 shows a simplified diagram of an ion implanter. An ion source provides a stream of ionized dopant atoms that are accelerated by the electric field of a miniature linear accelerator. A magnetic analyzer selects the desired species of ion, and a pair of deflection plates scans the resulting ion beam across the wafer. A high vacuum must be maintained throughout the system, so the entire apparatus is enclosed in a steel housing.

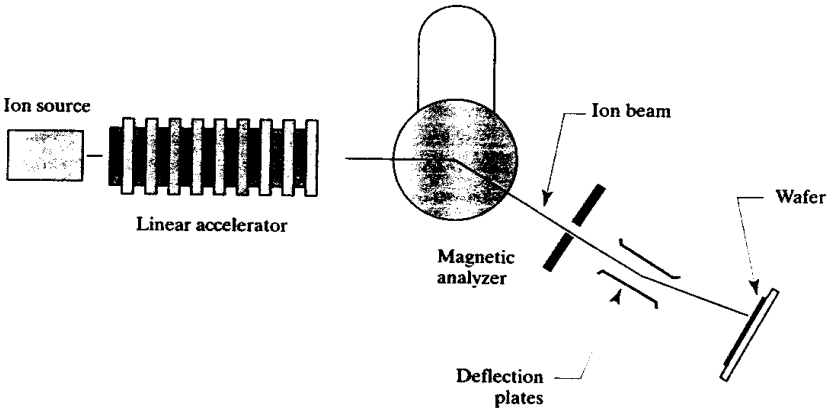


FIGURE 2.20 Simplified diagram of an ion implanter.[17]

Once the ions enter the silicon lattice, they immediately begin to decelerate due to collisions with surrounding atoms. Each collision transfers momentum from a moving ion to a stationary atom. The ion beam rapidly spreads as it sheds energy, causing the implant to spread out (*straggle*) in a manner reminiscent of outdiffusion. Atoms are also knocked out of the lattice by the collisions, causing extensive lattice damage that must be repaired by *annealing* the wafer at moderate temperatures (800°C to 900°C) for a few minutes. The silicon atoms become mobile and the intact silicon crystal structure around the edges of the implant zone serves as a seed for crystal growth. Damage progressively anneals out from the sides of the implant zone toward the center. Dopants added by ion implantation will redistribute by thermal diffusion if the wafer is subsequently heated to a sufficiently high temperature. Therefore, a deep lightly doped diffusion can be created by first implanting the required dopants and subsequently driving them down to the desired junction depth.

The dopant concentration provided by ion implantation is directly proportional to the *implant dose*, which equals the product of ion beam current and time. The dose can be precisely monitored and controlled, which allows for much better repeatability than conventional deposition techniques do. The doping profile is determined by the energy imparted to individual ions, a quantity called the *implant energy*. Low-energy implants are very shallow, while high-energy implants actually place most of the dopant atoms beneath the surface of the silicon. Ion implantation

[17] The scheme shown is but one of several; see Anner, p. 313ff.

can be used to counterdope a subsurface region to form a *buried layer.* Because of practical limitations on implant energy, these buried layers are usually quite shallow.

Ion implantation is somewhat anisotropic. The edges of an implant, especially a shallow low-energy one, do not spread as much as those produced by thermal diffusion. This aids in the manufacture of *self-aligned* structures that greatly improve the performance of MOS transistors. Figure 2.21 illustrates the creation of self-aligned MOS source/drain regions by ion implantation. A layer of polysilicon has been deposited and patterned on top of a thin gate oxide. The polysilicon not only forms the gate electrodes for MOS transistors but also simultaneously serves as an implant mask. The polysilicon blocks the implant from the region beneath the gate electrode, forming precisely aligned source and drain regions. The alignment of the source and drain with the gate is limited only by the small amount of straggle caused by the spreading of the ion beam. If self-aligned implants were not used, then photolithographic misalignments would occur between the gate and the source/drain diffusions, and the resulting overlap capacitances would substantially reduce the switching speed of the MOS transistors.

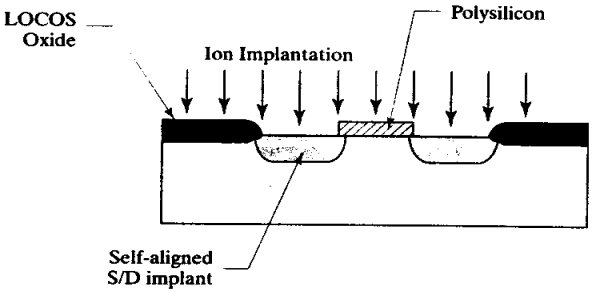


FIGURE 2.21 Self-aligned source and drain regions formed by ion implantation.

When the silicon lattice is viewed from certain angles, interstices between columns of silicon atoms, called *channels,* become visible. These disappear from view when the crystal is turned slightly. Channels are visible in both the (100) and the (111) silicon surfaces when these are viewed perpendicularly. If the ion beam were to impinge perpendicularly upon a (100) or a (111) surface, then ions could move deep into the crystal before scattering would commence. The final dopant distribution would depend critically upon the angle of incidence of the ion beam. To avoid this difficulty, most implanters project the ion beam onto the wafer at an angle of about 7°.

2.5 SILICON DEPOSITION

Films of pure or doped silicon can be chemically grown on the surface of a wafer. The nature of the underlying surface determines whether the resulting film will be monocrystalline or polycrystalline. If the surface consists of exposed monocrystalline silicon, then this serves as a seed for crystal growth and the deposited film will also be monocrystalline. If the deposition is conducted on top of an oxide or nitride film, then no underlying crystalline lattice will exist to serve as a seed for crystal nucleation, and the deposited silicon will form a fine-grained aggregate of polycrystalline silicon (*poly*). Modern integrated circuits make extensive use of both monocrystalline and polycrystalline deposited silicon films.

2.5.1. Epitaxy

The growth of a single-crystal semiconductor film upon a suitable crystalline substrate is known as *epitaxy*. The substrate normally consists of a crystal of the same material as the semiconductor that is to be deposited, but this need not always be the case. High-quality monocrystalline silicon films have been grown on wafers of synthetic sapphire or spinel, as these materials possess a crystal structure that is enough like silicon to allow crystal nucleation. The cost of synthetic sapphire or spinel wafers so greatly exceeds the cost of similar-sized silicon wafers that the vast majority of epitaxial depositions consist of silicon films grown on silicon substrates.

There are several different methods of growing epitaxial (*epi*) layers. One relatively crude method consists of pouring molten semiconductor material on top of the substrate, allowing it to crystallize for a short period of time, and then wiping the excess liquid off. The wafer surface can then be reground and polished to form an epitaxial layer. Obvious drawbacks to this *liquid-phase epitaxy* include the high cost of regrinding the wafer and the difficulty of producing a precisely controlled epi thickness.

Most modern epitaxial depositions use *low pressure chemical vapor deposited* (LPCVD) epitaxy. Figure 2.22 shows a simplified diagram of an early type of LPCVD epi reactor. The wafers are mounted on an inductively heated carrier block, and a mixture of dichlorosilane and hydrogen passes over them. These gases react at the surface of the wafers to form a slow-growing layer of monocrystalline silicon. The rate of growth can be controlled by adjusting the temperature, pressure, and gas mixture used in the reactor. No polishing is required to render the epitaxial surface suitable for further processing, as vapor-phase epitaxy faithfully reproduces the topography of the underlying surface. The epitaxial film can also be doped *in situ* by adding small amounts of gaseous dopants such as phosphine or diborane to the gas stream.

There are several benefits of growing an epitaxial layer on the starting wafer. For one, the epi layer need not have the same doping polarity as the underlying wafer. For example, an N– epitaxial layer can be grown on a P– substrate—an arrangement commonly employed for standard bipolar processes. Multiple epitaxial layers can be grown in succession and the resulting stack can be used to form transistors or other devices. The potential of epitaxy is limited chiefly by the slow rate of epi growth and by the expense and complexity of the required equipment, which are much greater than Figure 2.22 suggests.

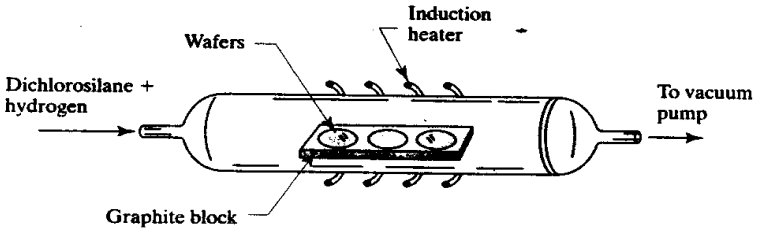


FIGURE 2.22 Simplified diagram of an epi reactor.[18]

[18] The horizontal tube reactor shown here has long been obsolete; see C. W. Pearce, "Epitaxy," in S. M. Sze, ed., *VLSI Technology* (New York: McGraw-Hill, 1983), pp. 61–65.

Epitaxy also allows the formation of *buried layers*. An N+ buried layer constitutes a key step in most bipolar processes since it allows the construction of vertical NPN transistors with low collector resistances. Figure 2.23 depicts the growth of such an N-buried layer (NBL). Arsenic and antimony are the preferred dopants for forming an NBL because their slow diffusion rates minimize the outdiffusion of the buried layer during subsequent high-temperature processing. Antimony is often chosen instead of arsenic because it exhibits less tendency to spread laterally during epitaxy (an effect called *lateral autodoping*).[19] Buried layer fabrication begins with a lightly doped P-type wafer. This wafer is oxidized, and windows are patterned in the resulting oxide layer. Either arsenic or antimony is implanted through the windows, and the wafer is briefly annealed to eliminate the resulting implant damage. Thermal oxidation occurs during this anneal, and discontinuities form around the edges of the oxide windows. Next, all oxide is stripped from the wafer, and an N– epitaxial layer is deposited. The resulting structure consists of patterned N+ regions buried under an epitaxial layer.

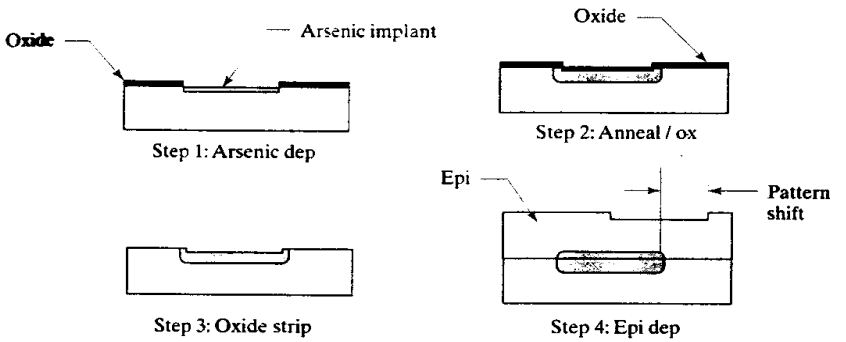


FIGURE 2.23 Formation of an N-buried layer (NBL), showing pattern shift.

As mentioned previously, oxidization during the anneal of the NBL causes slight surface discontinuities to form around the edges of the oxide window. The epitaxial layer faithfully reproduces these discontinuities in the final surface of the wafer. Under a microscope, the resulting step forms a faintly visible outline called the *NBL shadow*. Subsequent photomasks are aligned to this discontinuity. An alternative alignment method uses infrared light to image the NBL doping through the overlying silicon, but this requires more complicated equipment.

The accretion of silicon atoms at the edge of the NBL shadow during epitaxy displaces it laterally, an effect called *pattern shift* (Figure 2.23).[20] The magnitude of shift depends on many factors, including temperature, pressure, gas composition, substrate orientation, and tilt (See Section 7.2.3). When other layers are aligned to the NBL shadow, these must be offset to compensate for pattern shift.

---

[19] M. W. M. Graef, B. J. H. Leunissen, and H. H. C. de Moor, "Antimony, Arsenic, Phosphorus, and Boron Autodoping in Silicon Epitaxy," *J. Electrochem. Soc.*, Vol. 132, #8, 1985, pp. 1942–1954.

[20] M. R. Boydston, G. A. Gruber, and D. C. Gupta, "Effects of Processing Parameters on Shallow Surface Depressions During Silicon Epitaxial Deposition," in *Silicon Processing*, American Society for Testing and Materials STP 804, 1983, pp. 174–189.

2.5.2. Polysilicon Deposition

If silicon is deposited on an amorphous material, then no underlying lattice exists to align crystal growth. The resulting silicon film consists of an aggregate of small intergrown crystals. This *poly* film has a granular structure with a grain size dependent upon deposition conditions and subsequent heat treatment. The grain boundaries of the poly represent lattice defects, which can provide sneak paths for leakage currents. Therefore, PN junctions are not normally fabricated from poly. Polysilicon is often used to construct the gate electrodes of self-aligned MOS transistors because, unlike aluminum, it can withstand the high temperatures required to anneal the source/drain implants. In addition, the use of poly has led to better control of MOS threshold voltages due to the ability of phosphorus-doped polysilicon to immobilize ionic contaminants (Section 4.2.2). Suitably doped poly can be used to fabricate very narrow resistors that exhibit fewer parasitics than diffused devices. Heavily doped polysilicon can also be used as an additional metallization layer for signals that can tolerate the insertion of considerable resistance in the signal path.

A patterned poly layer is produced by first depositing polysilicon across the wafer using an apparatus similar to that employed for epitaxy (see Figure 2.22). The wafer is then coated with photoresist, patterned, and etched to selectively remove the polysilicon. Modern processes usually employ dry etching rather than wet etching because of the importance of precisely controlled gate dimensions.

2.6 METALLIZATION

The active components of an integrated circuit consist of diffusions, ion implantations, and epitaxial layers grown in or on a silicon substrate. When this processing is complete, the resulting components are connected to form the integrated circuit, using one or more layers of patterned wiring. This wiring consists of layers of metal and polysilicon separated by insulating material, usually deposited oxide. These same materials can also be used to construct passive components such as resistors and capacitors.

Figure 2.24 illustrates the formation of a typical *single-level-metal* (SLM) interconnection system. After the final implantations and diffusions, a layer of oxide is grown or deposited over the entire wafer, and selected areas are patterned and etched to create oxide windows exposing the silicon. These windows will form *contacts* between the metallization and the underlying silicon. Once these contacts have been opened, a thin metallic film can be deposited and etched to form the interconnection pattern.

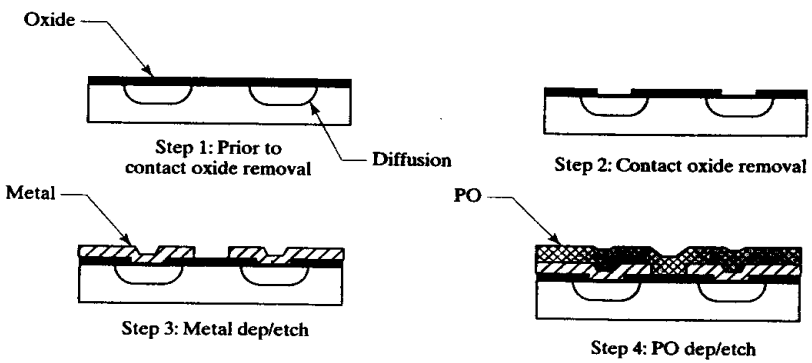


FIGURE 2.24 Formation of a single-level metal system.

Exposed aluminum wiring is vulnerable to mechanical damage and chemical corrosion. An oxide or nitride film deposited over the completed wafer serves as a *protective overcoat* (PO). This layer acts as a conformal seal similar in principle to the plastic conformal coatings sometimes applied to printed circuit boards. Windows etched through the overcoat expose selected areas of the aluminum metallization so that bondwires can be attached to the integrated circuit.

The process illustrated in Figure 2.24 fabricates only a single aluminum layer. Additional layers of metallization can be sequentially deposited and patterned to form a multilevel metal system. Multiple metal layers increase the cost of the integrated circuit, but they allow denser packing of components and therefore reduce the overall die size. The savings in die area often compensate for the cost of the extra processing steps. Multiple metal layers also simplify interconnection and reduce layout time.

CMOS processes frequently employ low-resistivity polysilicon to form the gate electrodes of self-aligned MOS transistors. This material can serve as a free additional layer of interconnect. Even the lowest-sheet poly still has many times the resistance of aluminum, so the designer must take care to avoid routing high-current or high-speed signals in poly. Advanced processes may add a second and even a third layer of polysilicon. These additional layers are used to fabricate different types of MOS transistors, to form the plates of capacitors, and to construct polysilicon resistors. Each of these additional poly layers can be pressed into service as another layer of interconnect.

2.6.1. Deposition and Removal of Aluminum

Most metallization systems employ aluminum or aluminum alloys to form the primary interconnection layers. Aluminum conducts electricity almost as well as copper or silver, and it can be readily deposited in thin films that adhere to all of the materials used in semiconductor fabrication. A brief period of heating will cause the aluminum to alloy into the silicon to form low-resistance contacts.

Aluminum is usually deposited by *evaporation* using an apparatus similar to the one shown in Figure 2.25. The wafers are mounted in a frame that holds their exposed surfaces toward a crucible containing a small amount of aluminum. When the crucible is heated, some of the aluminum evaporates and deposits on the wafer surfaces. A high vacuum must be maintained throughout the evaporation system to prevent oxidation of the aluminum vapor prior to its deposition upon the wafers. The illustrated evaporation system can only handle pure aluminum, but somewhat more complex systems can also evaporate selected aluminum alloys.

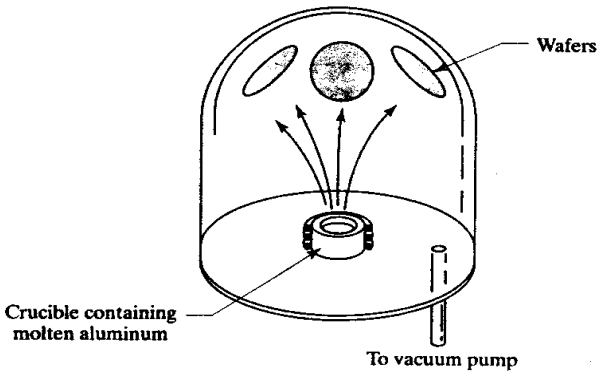


FIGURE 2.25 Simplified diagram of an aluminum evaporation apparatus.

Aluminum and silicon alloy at moderate temperatures. A brief period of heating will form an extremely thin layer of aluminum-doped silicon beneath the contact openings. This process, called *sintering,* can achieve Ohmic contact to P-type silicon because aluminum acts as an acceptor. The aluminum-silicon alloy forms a shallow, heavily doped P-type diffusion that bridges between the metal and the P-type silicon. Less obviously, Ohmic contact also occurs when aluminum touches heavily doped N-type silicon. Junctions form beneath these contacts, but their depletion regions are so thin that carriers can surmount them by quantum tunneling. Rectification will occur if the donor concentration falls too low, so Ohmic contact cannot be established directly between aluminum and lightly doped N-type silicon. The addition of a shallow N+ diffusion will enable Ohmic contact to these regions.

Sintering causes a small amount of aluminum to dissolve in the underlying silicon. Some silicon simultaneously dissolves in the aluminum metal, eroding the silicon surface. Some diffusions are so thin that erosion can punch entirely through them, causing a failure mechanism called *contact spiking.* Historically this was first observed in conjunction with the emitter diffusion of NPN transistors, so it is also called *emitter punchthrough.*[21] Contact spiking can be minimized by replacing pure aluminum metallization with an aluminum-silicon alloy. If the deposited aluminum is already saturated with silicon, then—at least in theory—it cannot dissolve any more. In practice, the silicon content of the alloy tends to separate during sintering to leave an unsaturated aluminum matrix. Careful control of sinter time and temperature will minimize this effect.

Another failure mechanism was encountered in high-density digital logic. As the dimensions of the integrated circuits were progressively reduced, the current density flowing through the metallization increased. Some devices eventually exhibited open-circuit metallization failures after many thousands of hours of operation at elevated temperatures. When the faulty units were examined, some of their leads contained unexpected breaks. These were eventually found to result from a failure mechanism called *electromigration.*[22] Carriers flowing through the metal collide with the lattice atoms. At current densities in excess of several million amps per square centimeter, these impacts become so frequent that the metal atoms begin to move. The displacement of the atoms causes voids to form between individual grains of the polycrystalline metal aggregate. Eventually these voids grow together to form a gap across the entire lead, causing an open-circuit failure (Section 4.1.2). The addition of a fraction of a percent of copper to the aluminum alloy improves electromigration resistance by an order of magnitude. Most modern metal systems therefore employ either aluminum-copper-silicon or aluminum-copper alloys.

2.6.2. Refractory Barrier Metal

The feature sizes of integrated circuits have steadily shrunk as ever-increasing numbers of components have been packed into approximately the same amount of silicon real estate. In order to obtain the necessary packing density, the sidewalls of contact and via openings have become increasingly steep. Evaporated aluminum does not deposit isotropically; the metal thins where it crosses oxide steps (Figure 2.26A). Any reduction in the cross-sectional area of a lead raises the current density and accelerates electromigration. A variety of techniques have been developed to improve step coverage on very steep sidewalls like those formed by reactive ion etching of thick oxide films.

---

[21] M. D. Giles, "Ion Implantation," in S. M. Sze, ed., *VLSI Technology* (New York: McGraw-Hill, 1983), pp. 367–369.

[22] J. R. Black, "Physics of Electromigration," *Proc. 12th Reliability Phys. Symp.*, 1974, p. 142.

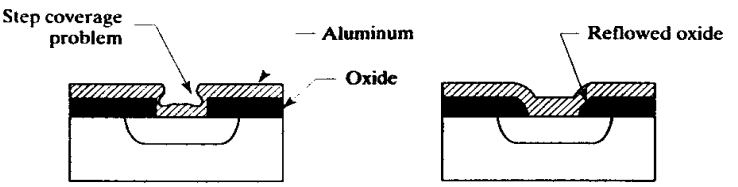


FIGURE 2.26 Step coverage of evaporated aluminum without reflow (A) and with reflow (B).

The step coverage of evaporated aluminum can be greatly increased by moderating the angle of the sidewalls. This can be achieved by heating the wafer until the oxide melts and slumps to form a sloped surface. This process is called *reflow* (Figure 2.26B). Pure oxide melts at too high a temperature to allow reflow, so phosphorus and boron are added to the oxide to reduce its melting point. The resulting doped oxide film is called either a *phosphosilicate glass* (PSG) or a *borophosphosilicate glass* (BPSG), depending on the choice of additives.

Reflow cannot be performed after aluminum has been deposited, because it cannot tolerate the temperatures required to soften PSG or BPSG. Therefore, while reflow can help improve the step coverage of first-level metal, it must be supplemented by other techniques in order to successfully fabricate a multilevel metal system. One option consists of using metals that deposit isotropically upon steeply sloped sidewalls, such as molybdenum, tungsten, or titanium. These *refractory barrier metals* possess extremely high melting points and are thus unsuited for evaporative deposition. A low-temperature process called *sputtering* can successfully deposit them. Figure 2.27 shows a simplified diagram of a sputtering apparatus. The wafers rest on a platform inside a chamber filled with low-pressure argon gas. Facing them is a plate of refractory barrier metal forming one of a pair of high-voltage electrodes. Argon atoms bombard the refractory metal plate. This knocks atoms loose that then deposit on the wafers to form a thin metallic film.

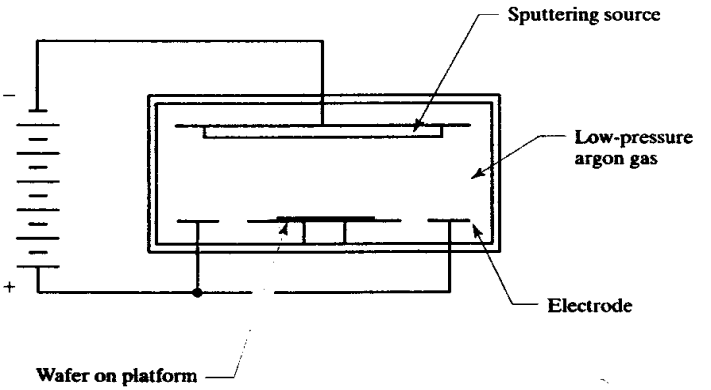


FIGURE 2.27 Simplified diagram of a sputtering apparatus.

The sputtered refractory barrier metal film not only provides superior step coverage, but also virtually eliminates emitter punchthrough.[23] If step coverage were the only criterion for choosing a metal system, then aluminum could be entirely replaced by refractory barrier metal. Unfortunately, refractory metals have relatively

[23] T. Hara, N. Ohtsuka, K. Sakiyama, and S. Saito, "Barrier Effect of W-Ti Interlayers in Al Ohmic Contact Systems," *IEEE Trans. Electron Devices*, Vol. ED-34, #3, 1987, pp. 593–597.

high resistivities and cannot be deposited in thick films as easily as aluminum can. Most metal systems therefore employ a sandwich of both materials. A thin layer of refractory metal beneath the aluminum ensures adequate step coverage in the contacts where the aluminum metal drastically thins. Elsewhere the aluminum reduces the electrical resistance of the metal leads. The relatively short sections of refractory barrier metal in the contacts do not contribute much resistance to the overall interconnection system.

Refractory barrier metals are extremely resistant to electromigration, so the thinning of aluminum on the sidewalls of contacts and vias does not represent an electromigration risk. Refractory barrier metal also tends to suppress classical electromigration failures by bridging any open circuits that develop in the aluminum metallization. Aluminum displaced by electromigration can still short adjacent leads, so refractory barrier metal cannot be relied on to supplement the current-carrying capacity of aluminum wiring except on the sidewalls of contact and via openings.

As mentioned, refractory barrier metal virtually eliminates emitter punchthrough. The degree of alloying between silicon and refractory metals is negligible and the aluminum cannot penetrate the barrier metal to contact the silicon. Most refractory barrier metal systems therefore employ aluminum-copper alloys rather than aluminum-copper-silicon, because aluminum-silicon alloying cannot occur.

2.6.3. Silicidation

Another modification of the standard metallization flow involves the addition of a silicide. Elemental silicon reacts with many metals, including platinum, palladium, titanium, and nickel, to form compounds of definite composition. These *silicides* can form both low-resistance Ohmic contacts and, in the case of certain silicides, stable rectifying Schottky barriers. Thus silicidation not only improves contact resistance—which can be a problem with barrier metal systems—but also allows the formation of Schottky diodes at no extra cost. Silicides have much lower resistivities than even the most heavily doped silicon, so they can also be used to reduce the resistance of selected silicon regions. Many MOS processes employ silicided poly (also called *clad poly*) to form the gates of high-speed MOS transistors. Some of these processes also clad the source/drain regions of the transistors to reduce their resistance. Since most silicides are relatively refractory, their deposition does not preclude subsequent high-temperature processing. Silicided gates can thus be used to form self-aligned source/drain regions.

Figure 2.28 shows the steps required to deposit a platinum silicide layer in selected regions of the wafer. Immediately after the contacts are opened, a thin film of

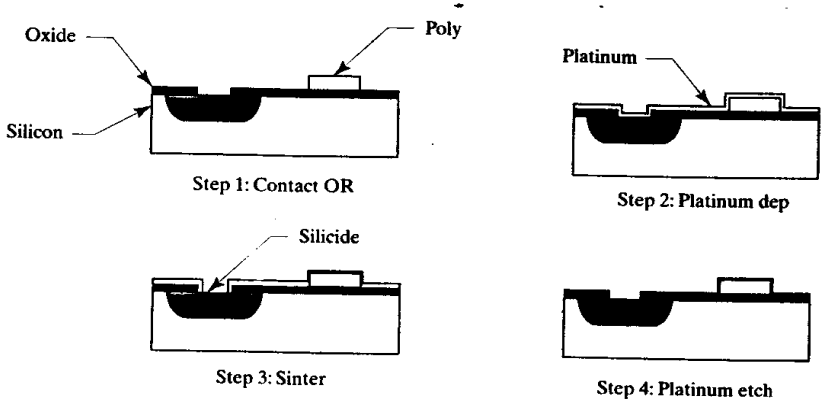


FIGURE 2.28 Silicidation process, showing both silicided contacts and silicided poly.

platinum metal is deposited across the entire wafer. The wafer is then heated to cause the portions of the platinum film in contact with silicon to react to form platinum silicide. The unreacted platinum can be removed using a mixture of acids called aqua regia. This procedure silicides both contact openings and any exposed polysilicon. If desired, an additional masking step can select which regions should receive silicide. Processes employing clad poly must incorporate a silicide block mask to fabricate polysilicon resistors. If this were not done, silicidation would turn all of the poly into a low-resistance material.

A typical silicided metal system consists of a lowermost layer of platinum silicide, an intermediate layer of refractory barrier metal,[24] and a topmost layer of copper-doped aluminum. The resulting sandwich exhibits low electrical resistance, high electromigration immunity, stable contact resistance, and precisely controlled alloying depths. The three layers required to obtain all of these benefits are more costly than a simple aluminum alloy metallization, but the performance benefits are substantial.

2.6.4. Interlevel Oxide, Interlevel Nitride, and Protective Overcoat

Figure 2.29 shows a cross section of a typical modern metallization system. The first layer of material above the silicon consists of thermally grown oxide. Upon this oxide lies a patterned polysilicon layer that will eventually form the gates of MOS transistors. On top of this poly lies a thin deposited oxide layer called a *multilevel oxide* (MLO) that serves to insulate the poly and to thicken the thermal oxide layer. Contact openings are etched through the MLO and thermal oxide to contact the silicon, and through the MLO to contact the poly. Following reflow, the contact openings are silicided to reduce contact resistance. Above the MLO lies the first layer of metal, consisting of a thin film of refractory barrier metal and a much thicker layer of copper-doped aluminum. Above the first metal layer lies another deposited oxide layer called an *interlevel oxide* (ILO), which insulates the first metal from the overlying second metal. Vias are etched through the ILO. On top of this lies the second layer of metal, again consisting of refractory barrier metal and copper-doped aluminum. The topmost and final layer consists of a compressive nitride film, which serves as a *protective overcoat* (PO). This metallization system has a total of six layers (one poly, two metals, MLO, ILO, and PO) and requires five masking steps (poly, contact, metal-1, via, metal-2, and PO). Some advanced processes may employ as many as three layers of polysilicon and five layers of metal.

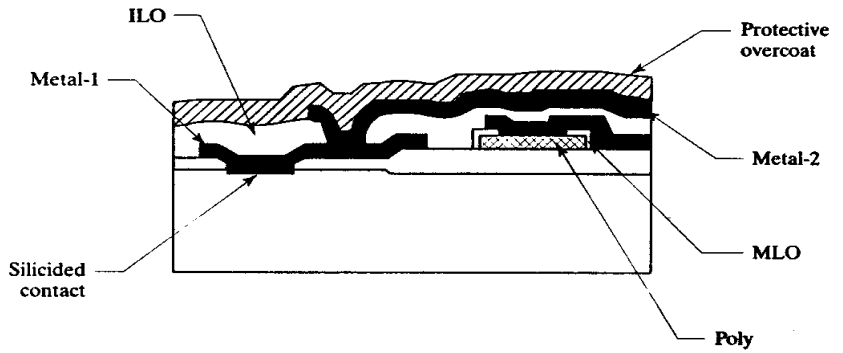


FIGURE 2.29 Cross section of a double-metal, single-poly metallization system.

[24] The addition of a refractory barrier metal prevents the platinum silicide from reacting with the aluminum. This is not required for most refractory silicides; see Sze, pg. 409.

Interlevel oxide layers are normally produced by low-temperature deposition—for example, by the reaction of silane and nitrous oxide or by the decomposition of tetraethoxysilane (TEOS). A relatively thick ILO helps minimize parasitic capacitances between layers of the conductor sandwich, but it can cause step coverage problems in via openings. As previously discussed, reflow is not possible once aluminum has been deposited, so a refractory barrier metal is often used to improve the step coverage of the second metal layer.

An excellent capacitor can be formed between two layers of metal or polysilicon. A thin insulating dielectric deposited between the plates completes the capacitor. The thinner this dielectric, the greater the resulting capacitance per unit area. One technique for forming a capacitor consists of depositing one polysilicon layer, oxidizing this to form a thin dielectric, and depositing a second polysilicon layer to complete the capacitor. Any region where the two poly layers overlap will form a capacitor consisting of two poly plates separated by the thin oxide dielectric. Oxide forms an ideal capacitor dielectric because it is a nearly perfect insulator, and very thin oxide films can be grown with little risk of pinholes or other defects. The capacitance achievable with oxide dielectrics is limited by the rupture voltage of the oxide; the thicker oxide layers required to withstand higher voltages have proportionately smaller capacitances per unit area.

One way to boost the capacitance per unit area for a given operating voltage consists of using a material with a higher dielectric constant. Silicon nitride, with a dielectric constant that is 2.3 times that of oxide, is a common choice for fabricating high capacitance-per-unit-area films. Unfortunately, nitride films are more prone to pinhole formation than are oxide films of equivalent thickness. Therefore oxide and nitride films are sometimes combined to form a stacked dielectric with a dielectric constant between that of oxide and nitride. A typical oxide-nitride-oxide stacked dielectric can achieve a dielectric constant about twice that of oxide.

The protective overcoat consists of a thick deposited oxide or nitride film coating the entire integrated circuit. It insulates the uppermost metal layer from the outside world, so that (for example) a particle of conductive dust will not short two adjacent leads. The overcoat also helps to ruggedize the integrated circuit, a necessary precaution since the aluminum metallization is soft and deforms under pressure. The protective overcoat also helps to block the entrance of contaminants. Both the aluminum metallization and the underlying silicon are vulnerable to certain types of contaminants that can penetrate the plastic encapsulation. A properly formulated protective overcoat can form a barrier to these contaminants. Heavily doped phosphosilicate glasses are sometimes used as protective overcoats, but most modern processes have switched to compressive nitride films, which offer superior mechanical hardness and chemical resistance.

2.7 ASSEMBLY

Wafer fabrication ends with the deposition of a protective overcoat, but there remain a number of manufacturing steps required to complete the integrated circuits. Since most of these steps require less-stringent cleanliness than wafer fabrication, they are usually performed in a separate facility called an *assembly/test site.*

Figure 2.30 shows a diagram of a typical finished wafer. Each of the small squares on the wafer represents a complete integrated circuit. This wafer contains approximately 300 integrated circuit dice arrayed in a rectangular pattern by the step-and-repeat process that created the stepped working plates. A few locations in the array are occupied by process control structures and test dice rather than actual integrated circuits.

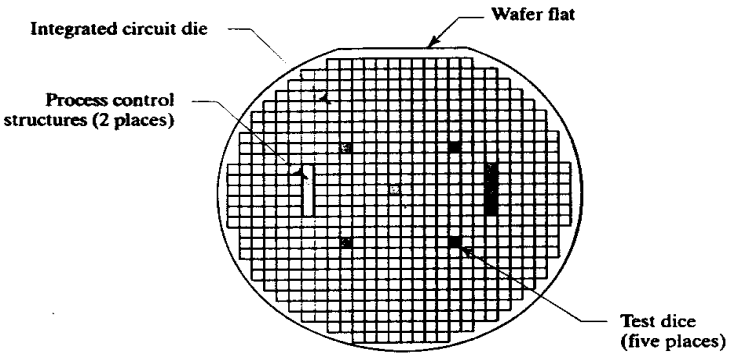


FIGURE 2.30 Pattern for a typical wafer created from a stepped working plate.

Process control structures consist of extensive arrays of transistors, resistors, capacitors, and diodes, as well as more specialized structures such as strings of contacts and vias. The wafer fab uses these structures to evaluate the success or failure of the manufacturing process. Automated testing equipment gathers data on every wafer, and any that fail to meet specifications are discarded. The data are also analyzed for statistical trends so that corrections can be implemented before the variances become large enough to cause yield losses. The process control structures are standardized, and the same ones are used for a wide range of products.

Test dice are used by design engineers to evaluate prototypes of an integrated circuit. Unlike process control structures, test dice are specific to a given product and in most cases are actually variations on the layout of the integrated circuit. A dedicated test metal mask allows probing of specific components and subcircuits that would be difficult to access on the finished die. Sometimes a test contact or protective overcoat mask is also used, but in almost every case the test die shares the same diffusion masks as the integrated circuit. Test dice are normally created by adding a few more layers (e.g., test metal, test nitride) to the database containing the layout of the integrated circuit. These layers create a separate set of reticles that are used to expose a few selected spots on the stepped working plate. The wafer in Figure 2.30 contains only five test die locations. These locations become unnecessary when testing has been completed. Sometimes a new set of masks is created that replaces the test dice with product dice to gain an extra percent or two of yield. In other cases, the tiny increase in die yield cannot justify fabrication of new masks, so the test dice remain on production material.

Figure 2.30 depicts a wafer produced from a set of stepped working plates. Wafers created by *direct-step-on-wafer* (DSW)[25] processing rarely include any test dice because at least one test die must be included in every exposure. This would result in twenty or more test dice per wafer, which would consume a corresponding amount of area. If test dice are included in a DSW design, then the production mask set will almost certainly be modified to replace them with product dice to improve the die yield.

As mentioned previously, all completed wafers are tested to determine whether the processing was performed correctly. If the wafers pass this test, then each die is

[25] The acronym *DSW* has also been used to stand for *direct slice write*, a process by which a scanned electron beam directly exposes the photoresist on a wafer. This process, more commonly called *direct-write-on-wafer* (DWW), is strictly of academic interest because it is too slow to have any practical application in silicon processing. However, it is frequently used to fashion photomasks.

individually tested to determine its functionality. The high-speed automated test equipment typically requires less than three seconds to test each die. The percentage of good dice depends on many factors, most notably the size of the die and the complexity of the process used to create it. Most products yield better than 80% and some yield in excess of 90%. High yields are obviously desirable because every discarded die represents lost profit. The equipment that tests the wafers also marks those that fail the test. Marking is usually done by placing a drop of ink on each defective die, but some modern systems eliminate the need for inking by remembering the location of the bad dice electronically.

Wafer-level testing, or *wafer probing,* requires contact to specific locations on the interconnection pattern of the integrated circuit. These locations are exposed through holes in the protective overcoat, allowing contact to be made with the help of an array of sharp metal needles, or *probes.* These probes are mounted on a board called a *probe card.* The automated test machine lowers the probe card until electrical continuity is established. The integrated circuit is tested, the card is lifted, and positioning servos move the wafer to align the next die underneath the probes.

Once the wafer has been completely tested, the individual dice are sawn apart using a diamond-tipped sawblade. Another automated system then selects the good dice from the scribed wafer for mounting and bonding. The rejected dice (including the remains of the process control structures and the test dice) are discarded.

2.7.1. Mount and Bond

Many manufacturers now offer unmounted integrated circuit dice, but the sales of such *bare dice* are seldom large. Most customers simply do not have the equipment or expertise needed to handle bare dice, let alone to package them. Packaging therefore falls in the province of integrated circuit manufacturing.

The first step in packaging an integrated circuit is mounting it on a *leadframe.* Figure 2.31 shows a somewhat simplified diagram of a leadframe for an eight-pin *dual-in-line package* (DIP), complete with a chip mounted on it. The leadframe itself consists of a rectangular *mount pad* that holds the die and a series of lead fingers that will eventually be trimmed to form the eight leads of the DIP. Leadframes usually come in strips, so several dice can be handled as a single assembly. They are either stamped out of thin sheets of metal, or they are etched using photographic techniques similar to those used to pattern printed circuit boards. The lead frame usually consists of copper or a copper alloy, often plated with tin or a tin-lead alloy. Copper is not an ideal material for leadframes because it has a different coefficient

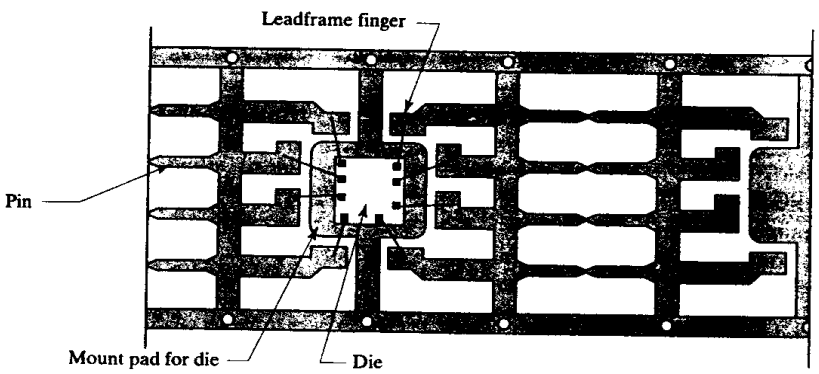


FIGURE 2.31 Simplified diagram of a leadframe for an 8-pin DIP.

of thermal expansion than silicon. As the packaged part is heated and cooled, differential expansion of the die and the leadframe sets up mechanical stresses injurious to the performance of the die. Unfortunately, most of the materials that possess coefficients of expansion similar to silicon have inferior mechanical and electrical properties. Some of these materials are occasionally used for low-stress packaging of specialty parts; a nickel-iron alloy called *Alloy-42* is probably the most commonly encountered (Section 7.2.6).

The die is usually mounted to the leadframe using an epoxy resin. In some cases, the resin may be filled with silver powder to improve thermal conductivity. Epoxy is not entirely rigid, and this helps reduce the stresses produced by thermal expansion of the leadframe and die. Alternate methods exist that provide superior thermal union between the silicon and the leadframe, but at the cost of greater mechanical stress. For example, the backside of the die can be plated with a metal or metal alloy and soldered to the leadframe. Alternatively, a rectangle of gold foil called a *gold preform* can be attached to the leadframe; heating the die causes it to alloy with the gold preform to create a solid mechanical joint. Solder connections and gold preforms both allow excellent thermal contact between the die and the leadframe. Both also produce an electrical connection that can be used to connect the substrate of the die to a pin. Conductive epoxies improve thermal conductivity, but they cannot always be trusted to provide electrical connectivity.[26]

After the dice are mounted on leadframes, the next step is attaching bondwires to them. Bonding can only be performed in areas of the die where the metallization is exposed through openings in the protective overcoat; these locations are called *bondpads*. The probe card used for wafer probing makes contact to the bondpads for purposes of testing, but the probes may also make contact to a few pads that will not receive bondwires. Those pads reserved for testing purposes are usually called *testpads* to distinguish them from actual bondpads. Testpads are often made smaller than bondpads since probe needles can usually land in a smaller zone than can bondwires.

Bonding is performed by high-speed automated machines that use optical recognition to determine the locations of the bondpads. These machines typically employ one-mil (25μm) gold wire for bonding, although gold wires as small as 0.8 mil or as large as 2.0 mil are in common use. Aluminum wires up to ten mils in diameter can also be employed, although these require rather different bonding machinery. Only one diameter and type of wire can be bonded at a time, so few dice use more. The most common arrangement consists of one-mil gold bondwires on all pads. Multiple one-mil wires bonded in parallel can carry higher currents or provide lower resistances without requiring a second bonding pass for larger-diameter wire.

The most common technique for bonding gold wire is called *ball bonding*.[27] Since aluminum wire cannot be ball-bonded, an alternate technique called *wedge bonding* has been developed for it. Figure 2.32 shows the essential steps of the ball-bonding process.

The bonding machine feeds the gold wire through a slender tube called a *capillary*. A hydrogen flame melts the end of the wire to form a small gold sphere, or *ball* (Figure 2.32, Step 1). Once a ball has been formed, the capillary presses down against the bondpad. The gold ball deforms under pressure, and the gold and aluminum fuse together to form a weld (Step 2). Next, the capillary lifts and moves to the vicinity of

---

[26] R. L. Opila and J. D. Sinclair, "Electrical Reliability of Silver Filled Epoxies for Die Attach." *23rd International Reliability Physics Symp.*, 1985, pp. 164–172.

[27] B. G. Streetman, *Solid State Electronic Devices*, 2nd ed. (Englewood Cliffs, NJ: Prentice-Hall, 1980), pp. 368–370.

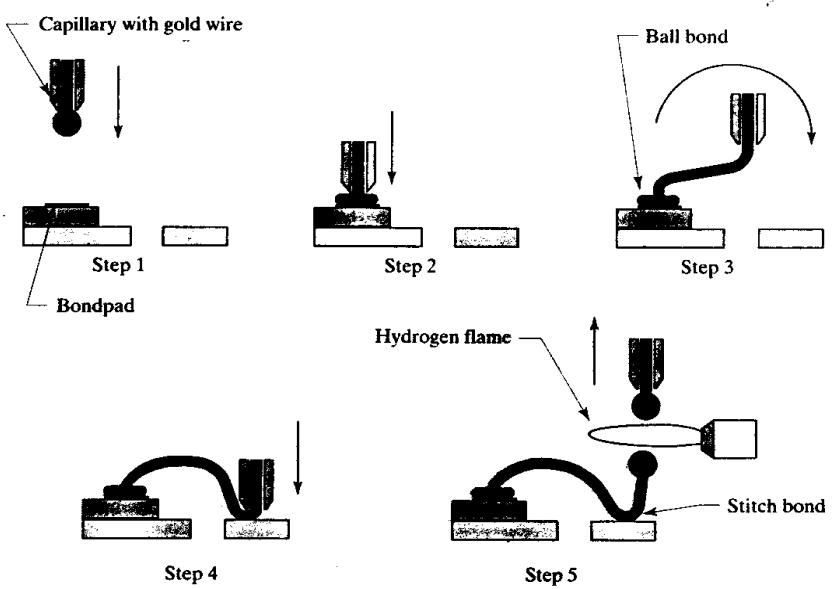


FIGURE 2.32 Steps in the ball-bonding process.

the lead finger (Step 3). The capillary again descends, smashing the gold wire against the lead finger. This causes the gold to alloy to the underlying metal to produce a weld (Step 4). Since no ball is present at this point, the resulting bond is called a *stitch bond* rather than a ball bond. Finally, the capillary lifts up from the lead finger and the hydrogen flame passes through the wire, causing it to fuse into two (Step 5). The bond is now complete and another ball has been formed on the wire protruding from the capillary, allowing the process to be repeated. An automated bonding machine can perform these steps ten times a second with great precision. The extreme speed and unerring accuracy of these machines produce huge economies of scale, and the entire bonding process costs no more than a penny or two.

Aluminum wire cannot be ball-bonded because the hydrogen flame would ignite the fine aluminum wire. Instead, a small, wedge-shaped tool is used to supplement the capillary. When the capillary brings the wire into proximity with the bondpad, the tool smashes it against the pad to create a stitch bond. The process is repeated at the lead finger, and the tool is then held down against the lead finger while the capillary moves up. The tension created in the aluminum wire snaps it at its weakest point, immediately adjacent to the weld. The process can then be repeated as many times as necessary.

The ball-bonding process requires a square bondpad approximately three times as wide as the diameter of the wire. Thus a one-mil gold wire can be attached to a square bondpad about three mils across. Wedge bonding is more selective, and usually requires bondpads that are rectangular in shape. These bondpads must lie in the same direction as the wedge tool. They are typically twice as wide and four times as long as the wire is thick. The exact rules for wedge bondpads can become quite complex, particularly for thicker aluminum wires.

Figure 2.31 shows a die mounted on a leadframe after the bonding process is complete. The bondwires connect various bondpads to their respective leads. Although the wires are quite small compared to the pins, each is still capable of carrying an amp of current.

2.7.2 Packaging

The next step in the assembly process is injection molding. A mold is clamped around the leadframe and heated plastic resin is forced into the mold from below. The plastic wells up around the die, lifting the wires away from it in gentle loops. Injection from the side or from the top usually smashes the wires against the integrated circuit and is therefore impractical. The plastic resin employed for integrated circuits cures rapidly at the temperatures used in molding and, once cured, it forms a rigid block of plastic.

When the molding process is complete, the leads are trimmed and formed to their final shapes. This is done in a mechanical press using a pair of specially shaped dies that simultaneously trim away the links between the individual leads and bend them to the required shape. Depending on the material of the leadframe, solder dipping or plating may be required to prevent oxidation and contamination of the pin surfaces. The completed integrated circuits are now labeled with part numbers and other designation codes (these usually include a code identifying the date of manufacture and the lot number). The completed integrated circuits are tested again to ensure that they have not been damaged by the packaging process. Finally, the completed devices are packaged in tubes, trays, or reels for distribution to customers.

2.8 SUMMARY

Modern semiconductor processing takes advantage of the properties of silicon to manufacture inexpensive integrated circuits in huge volumes. Photolithography allows the reproduction of intricate patterns hundreds or thousands of times across each wafer, leading to enormous economies of scale.

Junctions can be formed by one of three means: epitaxial deposition, diffusion, or ion implantation. Low-pressure chemical-vapor-deposited (LPCVD) epitaxial layers can produce extremely high-quality silicon films with precisely controlled dopant concentrations. Diffusion of dopants from a surface source allows the formation of vast numbers of junctions using only a single photolithographic masking step. Ion implantation allows similar but more costly patterning of junctions with superior control of doping levels and distributions.

Many materials can also be deposited on the surface of the wafer. These include polycrystalline silicon (poly), oxide, nitride, and any of numerous metals and metal alloys. Typical semiconductor processes combine several diffusions into the bulk silicon with several depositions of materials onto the resulting wafer. The next chapter examines how the various techniques of semiconductor fabrication are combined to manufacture three of the most successful integrated circuit processes.

2.9 EXERCISES

2.1. When pressure is applied to the center of an unknown wafer, it splits into six segments. What can be definitely concluded from this observation? What may be reasonably conjectured?

2.2. Draw a diagram similar to that in Figure 2.4 illustrating the relationship between the (100) and (110) planes of a cubic crystal (refer to Appendix B, if necessary).

2.3. Suppose a photomask consists of a single opaque rectangle on a clear background. A negative resist is used in combination with this mask to expose a sensitized wafer. Describe the pattern of photoresist left on the wafer after development.

2.4. Suppose a wafer is subjected to the following processing steps:

a. Uniform oxidation of the entire wafer surface.

b. Opening of an oxide window to the silicon surface.

c. An additional period of oxidation.
d. Opening of a smaller oxide window in the middle of the region patterned in step b.
e. An additional period of oxidation.

Draw a cross section of the resulting structure, showing the topography of both the silicon and the oxide surfaces. The drawing need not be made to scale.

2.5. Suppose a wafer is uniformly doped with equal concentrations of boron and phosphorus atoms. After a prolonged oxidation, will the surface of the silicon be N-type or P-type? Why?

2.6. Suppose that a wafer is uniformly doped with 10^{16} atoms/cm^3 of boron. This wafer is then subjected to the following processing steps:
a. Uniform oxidation of the entire wafer surface.
b. Opening of an oxide window to the silicon surface.
c. Deposition of boron and phosphorus, each at a source concentration of 10^{20} atoms/cm^3, using a fifteen-minute deposition and a one-hour drive at 1000°C.

Assuming that the two dopants do not interact with each other or with the oxide, draw a cross section of the resulting structure. Indicate the approximate depths of any junctions formed.

2.7. Phosphorus is diffused into a lightly doped wafer through an oxide window 5μm square. If the resulting junction is 2μm deep, then what is the width of the phosphorus diffusion at the surface?

2.8. Most ion implantation systems position the accelerator so that the ion beam impacts the wafer surface at a slight angle (often 7°). Explain the reason for this feature.

2.9. If the surface of the oxide layer covering a wafer is ground perfectly smooth, different regions of the wafer still exhibit different colors, but the NBL shadow vanishes. Explain these observations.

2.10. Draw a cross section of the following metallization system:
a. 1μm-wide contacts through 5kÅ oxide silicided with 2kÅ platinum silicide.
b. First-level metal consisting of 2kÅ RBM and 6kÅ copper-doped aluminum.
c. 1μm-wide vias 3kÅ deep through highly planarized ILO.
d. Second-level metal consisting of 2kÅ RBM and 10kÅ copper-doped aluminum.
e. 10kÅ protective overcoat.

Assume that the silicide surface is level with the surrounding silicon surface, and that the aluminum metal thins 50% on the sidewalls of the contacts and vias. The drawing should be made to scale.

2.11. Suppose that a die measures 60 by 80 mils, where one mil is one thousandth of an inch. Approximately how many of these dice could be fabricated on a 150mm-diameter wafer? Assuming that 70% of these potential dice actually work, and that a finished wafer costs $250, compute the cost of each functional die.