3

Device Fabrication Technology[1]

CHAPTER OBJECTIVES

While the previous chapters explain the properties of semiconductors, this chapter will explain how devices are made out of the semiconductors. It introduces the basic techniques of defining physical patterns by lithography and etching, changing the doping concentration by ion implantation and diffusion, and depositing thin films over the semiconductor's substrate. One section describes the techniques of fabricating the important metal interconnection structures. It is useful to remember the names of the key techniques and their acronyms, as they are often used in technical discussions.

With rapid miniaturization and efficient high-volume processing, over 10^{19} transistors (or a billion for every person in the world) are produced every year. Massive integration of transistors has made complex circuits in the form of **integrated circuits** (ICs) inexpensive and a wide range of electronic applications practical and affordable. Semiconductor devices are responsible for the arrival of the "computer age" or the "second industrial revolution." At the heart of the information and communication technologies, ICs of all descriptions also find applications in consumer electronics, automobiles, medical equipment, and industrial electronics. As a result, semiconductor devices are making contributions to every segment of the global economy and every branch of human endeavors.

Many large semiconductor companies both design and fabricate ICs. They are called **integrated semiconductor companies**. An even larger number of companies only design the circuits. They are called **fabless** design companies. They leave the fabrication to silicon **foundries**, which specialize in manufacturing. So an IC company may or may not fabricate the chips that they design.

[1] Readers who are more interested in devices than fabrication technology may proceed to Chapter 4 after reading the introduction and Section 3.1 of this chapter. Some subsequent chapters will refer back to specific parts of Chapter 3 and afford the reader the opportunity to pick up the needed information on fabrication technology.

• VLSI! ULSI! GSI! •

The complexity or density of integration of ICs is sometimes described by the names LSI (large-scale integration, 10^4 transistors on a chip), **VLSI** (very large-scale integration, 10^6 transistors on a chip), **ULSI** (ultra-large-scale integration), and **GSI** (giga-scale integration). In actuality, all these terms are used to describe circuits and technologies of wide ranges of size and complexity and simply mean "large IC."

3.1 • INTRODUCTION TO DEVICE FABRICATION •

A handful of companies produce most of the silicon wafers (Fig. 1–3b) used in the world. Hundreds of silicon device fabrication lines purchase these wafers as their starting material. A large **wafer fab** can process 40,000 silicon wafers into circuits each month.

The simple example of the device fabrication process shown in Fig. 3–1 includes (a) formation of an SiO_2 layer, (b) its selective removal, (c) introduction of dopant atoms into the wafer surface, and (d) dopant diffusion into silicon.

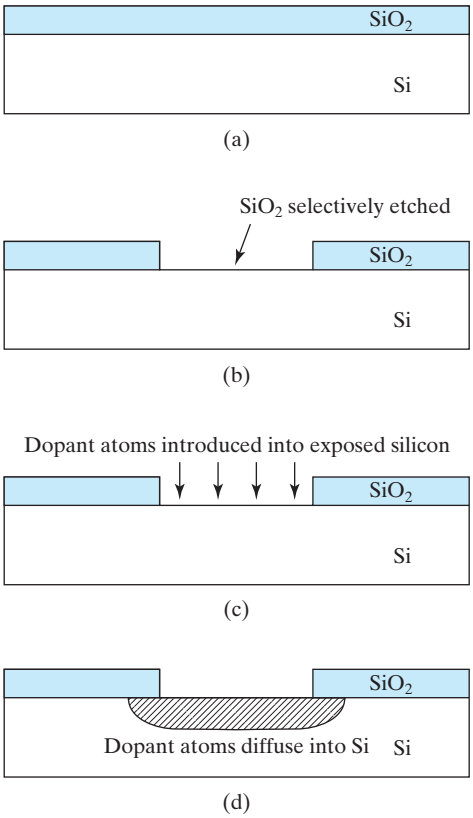


FIGURE 3–1 Some basic steps in the silicon device fabrication process: (a) oxidation of silicon; (b) selective oxide removal; (c) introduction of dopant atoms; and (d) diffusion of dopant atoms into silicon.

Combination of these and other fabrication steps can produce complex devices and circuits. This step-by-step and layer-upon-layer method of making circuits on a wafer substrate is called **planar technology**.

A major advantage of the planar process is that each fabrication step is applied to the entire silicon wafer. Therefore, it is possible to not only make and interconnect many devices with high precision to build a complex IC, but also fabricate many IC chips on one wafer at the same time. A large IC, for example, a central processor unit or **CPU**, may be 1–2 cm on a side, and a wafer (perhaps 30 cm in diameter) can produce hundreds of these chips. There is a clear economic advantage to reduce the area of each IC, i.e., to reduce the size of devices and metal interconnects because the result is more chip per wafer and lower cost per chip.

Since 1960, the world has made a huge investment in the planar microfabrication technology. Variations of this technology are also used to manufacture flat-panel displays, micro-electro-mechanical systems (MEMS), and even DNA chips for DNA screening. The rest of this chapter provides an introduction to the modern device processing technology. Perhaps the most remarkable advances have occurred in the fields of lithography (Section 3.3) and interconnect technology (Section 3.8). These are also the two areas that soak up the largest parts of the IC fabrication cost.

3.2 • OXIDATION OF SILICON •

In ICs, silicon dioxide is used for several purposes, ranging from serving as a mask against dopant introduction into silicon to serving as the most critical component in the metal-oxide-semiconductor transistor, the subject of Chapters 5–7.

SiO_2 layers of precisely controlled thickness are produced during IC fabrication by reacting Si with either oxygen gas or water vapor at an elevated temperature. In either case the oxidizing species diffuses through the existing oxide and reacts at the Si–SiO_2 interface to form more SiO_2. The relevant overall reactions are

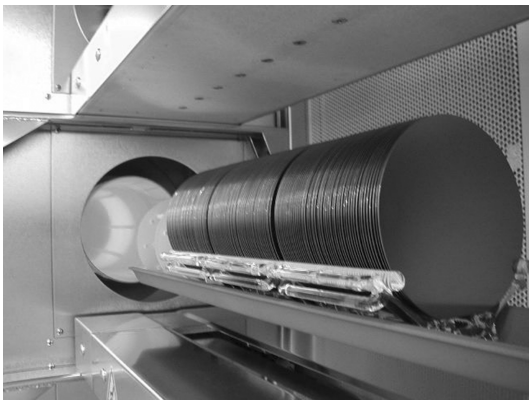
$$Si + O_2 \rightarrow SiO_2 \tag{3.2.1a}$$

$$Si + 2H_2O \rightarrow SiO_2 + 2H_2 \tag{3.2.1b}$$

Growth of SiO_2 using oxygen and water vapor is referred to as **dry** and **wet oxidation**, respectively. Dry oxidation is used to form thin oxide films. Wet oxidation, on the other hand, proceeds at a faster rate and is therefore preferred in forming the thicker oxides. Water vapor diffuses through SiO_2 faster than oxygen.

Figures 3–2a and b show a **horizontal furnace**. Oxidation may also be carried out in a **vertical furnace** as shown in Fig. 3–2c. A simplified sketch of the furnace is presented in Fig. 3–3. Oxidation temperatures of 700– 1,200 °C are produced in the furnace by electrical resistance heating coils. The tube at the center of the furnace is usually made of clear fused quartz, although SiC and polycrystalline Si tubes are also used. The Si wafers to be oxidized are loaded onto a quartz boat and pushed into the center of the furnace. During dry oxidation, the oxygen gas is fed into the

(a)



(b)

(c)

FIGURE 3–2 Examples of furnace systems that may be used for oxidation and other processes. (a) is a horizontal furnace and (b) is a close-up photo showing sillicon wafers waiting to be pushed into the furnace. (© Steed Technology, Inc. Used by permission.) (c) shows a newer vertical furnace. (Copyright © ASM International N.V. Used by permission.) The vertical furnaces occupy less floor space.

tube. Wet oxidation is performed by bubbling a carrier gas (Ar or N_2) through water in a heated flask (see Fig. 3–3) or by burning O_2 and H_2 to form H_2O at the input to the tube. Generally, in a production system, processes such as wafer loading, insertion into the furnace, ramping of the furnace temperature, and gas control are all automated. The thickness of the oxide grown depends on the furnace temperature, the oxidation time, the ambient gas, and the Si surface orientation. Representative dry and wet oxidation growth curves are shown in Fig. 3–4. Wafers

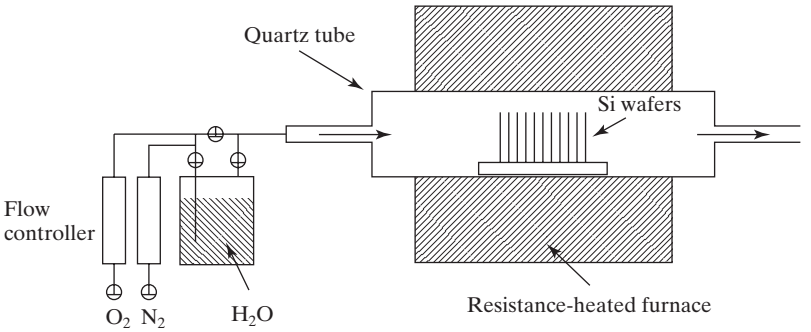


FIGURE 3–3 Schematic drawing of an oxidation system.

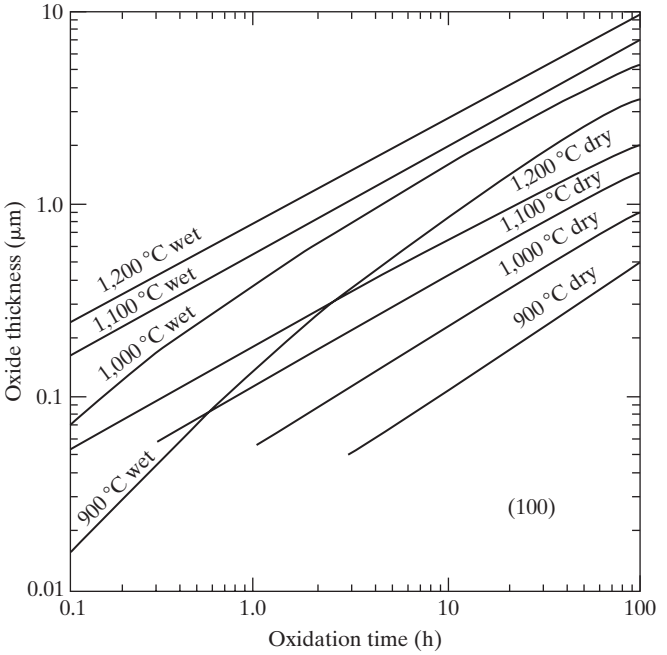


FIGURE 3–4 The SiO_2 thickness formed on (100) silicon surfaces as a function of time. (From [2]. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.)

used in IC productions are predominantly cut in the (100) plane because the interface trap density (see Section 5.7) is low due to the low density of unsaturated bonds in this plane relative to the other planes. Also, the electron surface mobility (see Section 6.3.1) is high.

EXAMPLE 3–1 Two-Step Oxidation

a. How long does it take to grow 0.1 μm of dry oxide at 1,000°C?

b. After step (a), how long will it take to grow an additional 0.2 μm of oxide at 900 °C in a wet ambient so that the total oxide thickness is 0.3 μm?

SOLUTION:

a. From the "1,000 °C dry" curve in Fig. 3–4, it takes 2.5 h to grow 0.1 μm of oxide.

b. In this part, use the "900 °C wet" curve only. First we determine that it would have taken 0.7 h to grow the 0.1 μm oxide at 900 °C in a wet ambient and 2.4 h to grow 0.3 μm of oxide from bare silicon. This means that it will take 2.4 – 0.7 h = 1.7 h in a wet 900 °C furnace to increase the oxide thickness from 0.1 to 0.3 μm. This is the correct answer regardless of how the first 0.1 μm oxide is produced (900 °C wet or 1,000 °C dry or any other condition). The answer is 1.7 h.

3.3 • LITHOGRAPHY •

How can we selectively remove oxide from those areas in which dopant atoms are to be introduced in Fig. 3–1b? Spatial selection is accomplished using a process called **photolithography** or **optical lithography**.

Major steps in the lithography process are illustrated in Fig. 3–5 using the patterning of an SiO_2 film as an example. The top surface of the wafer is first coated with an ultraviolet (UV) light sensitive material called **photoresist**. Liquid photoresist is placed on the wafer, and the wafer is spun at high speed to produce a thin, uniform coating. After spinning, a short bake at about 90 °C is performed to drive solvent out of the resist.

The next step is to expose the resist through a photomask and a high-precision reduction (for example 5 to 1 reduction) lens system using UV light as illustrated in Fig. 3–5b. The **photomask** is a quartz photoplate containing the patterns to be produced. Opaque regions on the mask block the UV light. Regions of the photoresist exposed to the light undergo a chemical reaction that varies with the type of resist being employed. In **negative resists**, the areas where the light strikes become polymerized and more difficult to dissolve in solvents. When placed in a developer (solvent), the polymerized regions remain, while the unexposed regions dissolve and wash away. The net result after development is pictured on the right-hand side of Fig. 3–5c. **Positive resists** contain a stabilizer that slows down the dissolution rate of the resist in a developer. This stabilizer breaks down when exposed to light, leading to the preferential removal of the exposed regions as shown on the left-hand side of Fig. 3–5c. Steps (a) through (c) make up the complete lithography process. To give a context for the purpose of lithography, we include step (d) for oxide removal. Buffered hydrofluoric acid (HF) may be used to dissolve unprotected regions of the oxide film. Lastly, the photoresist is removed in a step called **resist strip**. This is accomplished by using a chemical solution or by oxidizing or "burning" the resist in an oxygen plasma or a UV ozone system called an **asher**.

Optical diffraction limits the minimum feature size that can be resolved to k times the wavelength of the light used in the optical exposure system.

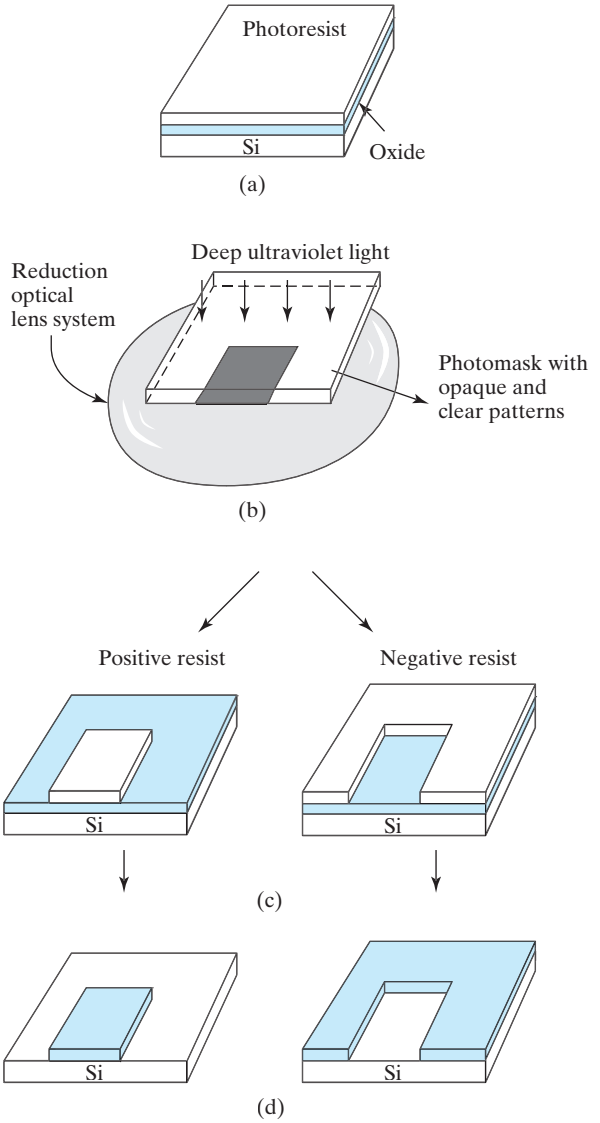


FIGURE 3–5 Major steps in the lithography process: (a) application of resist; (b) resist exposure through a mask and an optical reduction system; (c) after development of exposed photoresist; and (d) after oxide etching and resist removal. (After [2]. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.)

$$\text{Lithography Resolution} = k\lambda \tag{3.3.1}$$

A straightforward (but not easy) way to extend the resolution limit is to use UV light of shorter and shorter wavelengths that correspondingly reduce the resolution limit. Laser light sources of 248 and 193 nm (deep UV) are widely used. It is difficult to further reduce the wavelength (e.g., to 157 nm) owing to the lack of suitable transparent materials for lenses and mask plates at this wavelength. The factor k depends on the lens system and the photomask technology as described in the next paragraphs.

To obtain the best optical resolution, only a small area, about 10 cm^2, of the wafer is exposed in step (b). This area is called the **lithography field** and may contain a few to tens of IC chips. This exposure step is repeated for a neighboring area on the wafer and then another area by moving the wafer until the entire wafer has been exposed. For this reason, the lithography equipment is called a **stepper** for its **step-and-repeat** action.

Distortion of a pattern can result from the effect of the neighboring patterns surrounding it on the photomask. For example, a line may be successfully resolved but two lines close to each other may be bridged. This can be corrected by making the line slightly thinner on the photomask to begin with. This important technique is called **optical proximity correction** or **OPC**. Much computational resource is needed to perform OPC, i.e., to fine tune the photomask for a large IC pattern by pattern.

The k value in Eq. (3.3.1) can be reduced and the resolution limit can be pushed out with several other resolution enhancement techniques. For example, a **phase-shift photomask** might produce a 180° phase difference in the two clear regions on either side of a thin dark line by selective etching of the photomask substrate. Their diffractions into the dark region have electric fields of opposite signs (180° phase difference) and partially cancel each other out. As a result, thinner lines can be resolved. Some other examples of enhancement techniques are excluding certain ranges of the line-space pitch or allowing only certain ranges of it, shaped rather than uniform light source, and exposing only the vertical line patterns with one mask followed with exposing only the horizontal line patterns with another mask.

In addition to resolving small features, lithography technology also provides **alignment** between two lithography steps with an accuracy of about one-third the minimum feature size. *Lithography is the most difficult and expensive process among all the IC fabrication steps.* A typical IC fabrication flow applies the lithography technique over 20 times, each time using a different photomask.

3.3.1 Wet Lithography

Because of the difficulty of finding suitable materials for lenses and masks at wavelengths shorter than 193 nm, a clever technology has been developed to obtain better lithography resolution without requiring a shorter wavelength.

Figure 3–6a shows the objective lens of the optical lithography system and a wafer placed beneath it waiting to be exposed. The gap between the lens and the

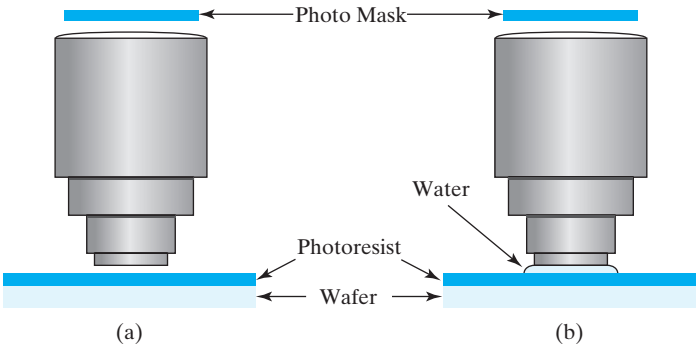


FIGURE 3–6 Schematics of (a) conventional dry lithography and (b) wet or immersion lithography. The wavelength of light source is 193 nm in both cases, but the effective wavelength in (b) is reduced by the refraction index of water, 1.43.

Extreme UV Lithography

A bold extension of optical lithography, **extreme ultraviolet lithography** or **EUVL** technology, would use a 13-nm wavelength. This is a huge leap in the reduction of the light source wavelength and the theoretically achievable resolution. Because extreme ultraviolet light is strongly absorbed by all materials, an all-reflective optical system using mirrors instead of lenses is used as shown in Fig. 3–7. Even the photomask is based on reflection rather than transmission. The optical surfaces need to be flat and smooth to 0.25 nm (the size of an atom). The EUV light may be generated by zapping a stream of Xe gas with laser pulses.

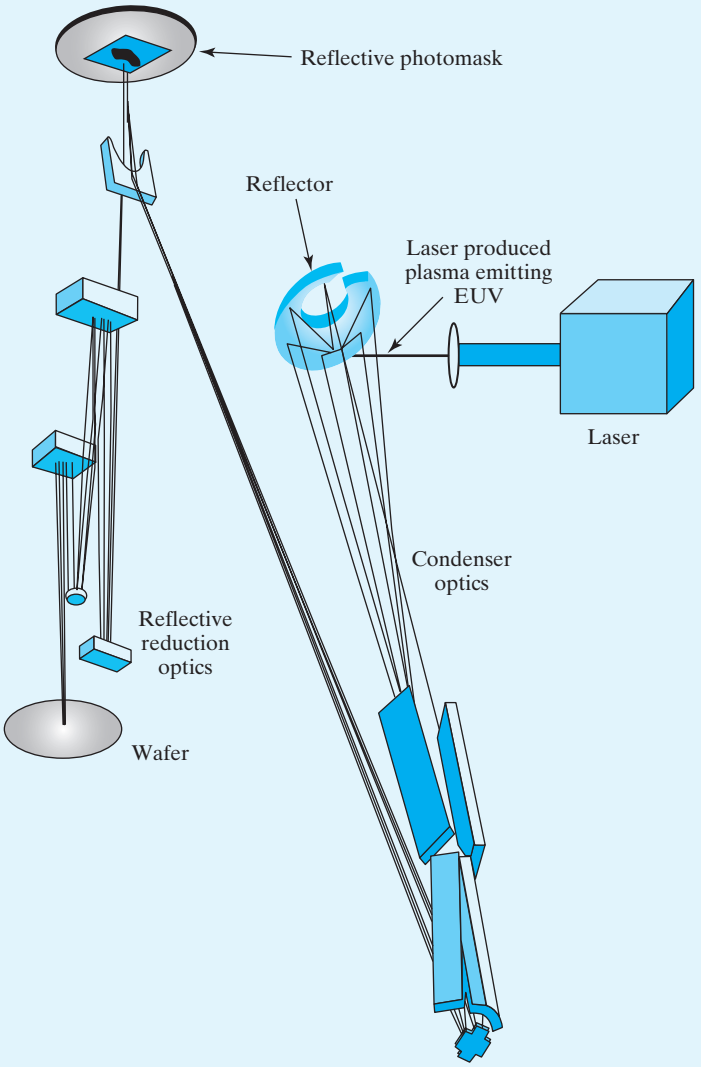


FIGURE 3–7 A schematic illustration of an extreme UV lithography system. (After Scott Hector, Motorola.)

wafer is a few millimeters. If this gap is filled with water as shown in Fig. 3–6b by immersing the system in water, we have the gist of **wet lithography** or **immersion lithography**. When light enters the water, its wavelength is reduced by the refraction index of water, 1.43, and therefore the lithography resolution is improved according to Eq. (3.3.1). Furthermore, the resolution can be improved even more by using a suitable liquid that has a larger index of refraction than water.

3.3.2 Electron Lithography

It is well known that electron microscopes have better resolution than optical microscopes. Electron lithography similarly is an alternative to optical lithography with resolution advantage. In **electron-beam lithography**, a focused stream of electrons delivers energy to expose the electron resist. The electron beam is scanned to write the desired pattern. The information necessary to guide the electron beam is stored in a computer and no mask is used. Electron-beam lithography has long been used to fabricate the photomasks used in optical lithography and for EUVL. For direct printing of patterns on wafers, electron lithography has slower exposure rates (in wafers per hour) than optical lithography. The exposure rate can be increased by employing multiple electron beams in each lithography machine.

There are schemes to expose a complex pattern simultaneously using a mask and a reduction electron-lens system (a carefully designed magnetic field), similar to optical lithography. This would improve the exposure rate.

3.3.3 Nanoimprint

High-resolution lithography, whether optical or electron lithography, is very expensive. Therefore, creating fine patterns without performing the expensive lithography is attractive. **Nanoimprint** is such a technique. Electron lithography is used to produce the fine patterns. The patterns are transferred (etched, see Section 3.4) into a suitable material to make a "stamp." This stamp is pressed into a soft coating over the wafer surface to create an *imprint* of the fine patterns. After the coating hardens, the desired fine patterns (see Fig. 3–5d) have been replicated on the wafer. The stamp can be used repeatedly to produce many wafers. In this sense, the stamp is the equivalent of the photomask in optical lithography.

3.4 • PATTERN TRANSFER—ETCHING •

After the pattern is formed in the resist by lithography, the resist pattern is often transferred to an underlying film, for example, the SiO_2 in Fig. 3–5d. If SiO_2 is removed with HF, this etching method is called **wet etching**. Since wet etching is usually **isotropic** (meaning without preference in direction, and proceeding laterally under the resist as well as vertically toward the silicon surface), the etched features are generally larger than the dimensions of the resist patterns as shown in Fig. 3–8a. **Dry etching** technique can overcome this shortcoming and is the dominant etching technology.

In dry etching, also known as **plasma etching** or **reactive-ion etching** or **RIE**, the wafer with patterned resist is exposed to a plasma, which is an almost neutral mixture of energetic molecules, ions, and electrons that is usually created by a radio frequency (RF) electric field as shown in Fig. 3–9a. The energetic species react

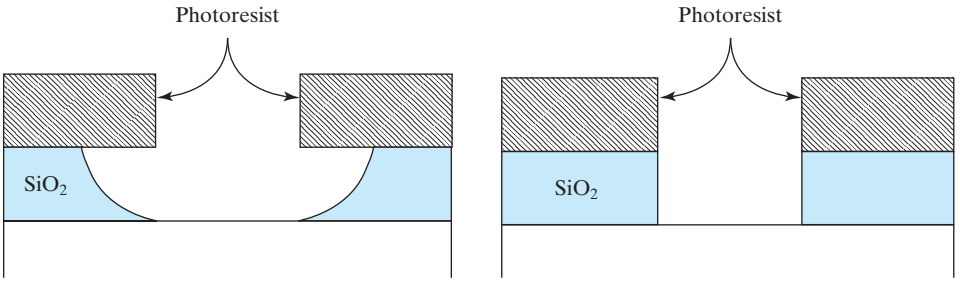
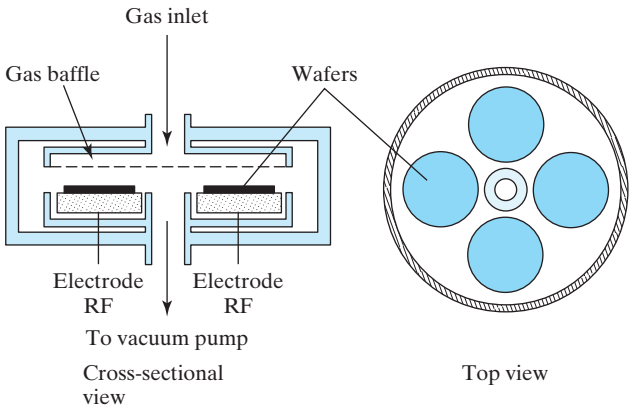
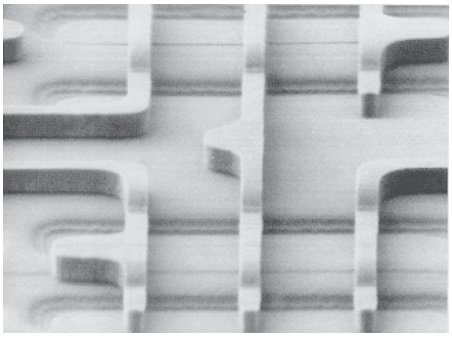


FIGURE 3–8 Comparison between (a) isotropic etching and (b) anisotropic etching.

chemically with the exposed regions of the material to be etched, while the ions in the plasma bombard the surface vertically and knock away films of the reaction products on the wafer surface. The latter action is directional so that the etching is preferentially vertical because the vertical surfaces can be covered with films of the reaction products. Hence the etch rate is **anisotropic**.



(a)



(b)

FIGURE 3–9 (a) A reactive-ion etching chamber and (b) scanning electron microscope view of a 0.16 μm pattern etched in polycrystalline silicon film. Excellent line width control is achieved even though the underlying surface is not flat [3].

By proper choices of the reactor design and etching chemistry, nearly vertical walls are produced in the etched material as shown in Fig. 3–9b. Low pressures and highly one-directional electric field tend to make etching anisotropic. Dry etching can also be designed to be isotropic or partially anisotropic if that is desired.

Suitable gas(es) is (are) introduced into the etch chamber based on the material to be etched. Silicon and its compounds can be etched by plasmas containing fluorine (F), whereas aluminum is etched with chlorine-containing plasmas.

The material **selectivity** of dry etching is usually not as high as that of wet etching. The material to be etched and the underlying material (e.g., SiO_2 and the underlying silicon) can both be significantly attacked during the etching process. Therefore, the dry etching process must be terminated as soon as the desired layer has been removed. This can be done with an **end-point detector**, which monitors the telltale light emission from the various etching products. There is often a trade-off between selectivity and anisotropy. For example, bromine (Br) provides better selectivity between Si and SiO_2 but poorer anisotropy than Cl.

Processing using plasma can potentially cause damage to the devices on the wafer. This is known as **plasma process-induced damage** or **wafer charging damage**. The main damage mechanism is the charging of conductors by the ions in the plasma, leading to an overly high voltage across a thin oxide and causing oxide breakdown. The worst condition is a small, thin oxide area connected to a large conductor, which collects a large amount of charge and current from the plasma and funnels them into the small-area oxide. The sensitivity of the damage to the size of the conductor is called the **antenna effect**.

Of course, pattern transfer is not limited to transferring a resist pattern onto another material. A pattern in an oxide may be transferred to Si, for example.

3.5 • DOPING •

The density profile of the dopant atoms in the silicon (**dopant profile**) is generally determined in two steps. First, the dopant atoms are placed on or near the surface of the wafer by **ion implantation**, **gas-source doping**, or **solid-source diffusion**. This step may be followed by an intentional or unintentional **drive-in diffusion** that transports the dopant atoms further into the silicon substrate.

3.5.1 Ion Implantation

Ion implantation is the most important doping method because of the precise control it provides. In ion implantation, an impurity is introduced into the semiconductor by creating ions of the impurity, accelerating the ions to high energies ranging from subkiloelectronvolt to megaelectronvolt, and then literally shooting the ions onto the semiconductor surface (Fig. 3–10). As one might suspect, the implanted ions displace semiconductor atoms along their paths into the crystal. Moreover, the ions themselves do not necessarily come to rest on lattice sites. A follow-up **anneal** (heating) of the wafer is therefore necessary for damage removal and for **dopant activation** (placing the dopant atoms on lattice sites as shown in Fig. 1–6) so that implanted impurities behave as donors and acceptors.

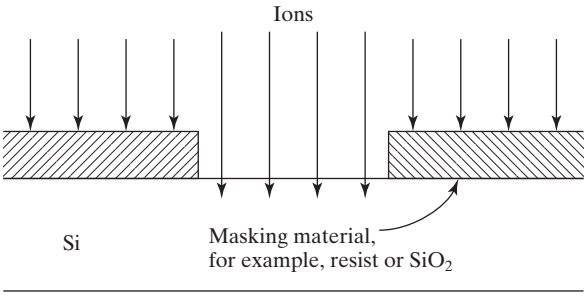


FIGURE 3–10 In ion implantation, a beam of high-energy ions penetrates into the unprotected regions of the semiconductor.

A schematic of an ion implantation system is presented in Fig. 3–11. Ions of the desired impurity are produced in the ion source shown at the extreme left. The ions are next accelerated into the mass analyzer where only the desired ions pass through a slit in the ion selection aperture. The resulting ion beam is then accelerated to the implantation energy, and finally the inch-size ion beam is scanned over the surface of the wafer, which is mounted on a massive metal plate. Scanning is accomplished by electrostatically scanning the ion beam, by mechanically moving the wafer, or by a combination of the two methods. An electrical contact to the wafer allows a flow of electrons to neutralize the implanted ions. A very precise determination of the total number of implanted ions per square centimeter (called the **implantation dose**, N_i) is

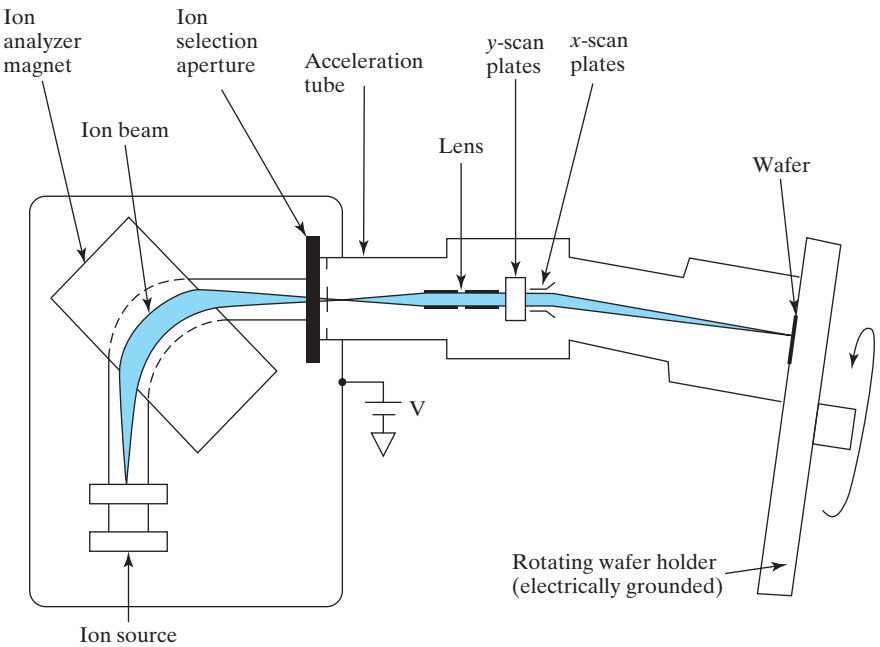


FIGURE 3–11 Simplified schematic of an ion implantation system. (After [4].)

obtained by simply integrating the beam current over the time of the implant. The concentration profile produced by ion implantation has the general form of a Gaussian function and is described by the peak location below the surface (R, called the **implantation range**), and the spread (ΔR, called **implantation straggle**).

$$N(x) = \frac{N_i}{\sqrt{2\pi}(\Delta R)} \cdot e^{-(x-R)^2/2\Delta R^2} \tag{3.5.1}$$

These parameters vary with the implant ion and substrate material and are roughly proportional to the ion energy as shown in Fig. 3–12. Computed distributions for phosphorus implanted into Si at various energies are shown in Fig. 3–13. Ion implantation processes can sometimes cause wafer charging damage. To alleviate this problem, electrons may be introduced near the wafer to neutralize the charge on the wafer.

3.5.2 Gas-Source Doping

In practice, gas-source doping is used to dope Si with phosphorus only. There are no convenient gas sources for As or B. It is carried out in a furnace similar to that used for oxidation (see Figs. 3–2 and 3–3). The N_2 carrier gas in Fig. 3–3 would pass through a bubbler containing phosphorus oxychloride ($POCl_3$, often pronounced "pockle") that is a liquid at room temperature. The N_2 carries the vapor of the source into the furnace tube. The reaction with Si or other gases liberates phosphorus atoms, which diffuse into the silicon.

3.5.3 Solid-Source Diffusion

In solid-source diffusion, the Si surface is first coated with a thin film (of a SiGe alloy, for example) containing dopants as deposited or due to subsequent implant of dopants into this film (and leave the damages in it). Dopants are diffused into Si. The SiGe film may be removed by wet etching.

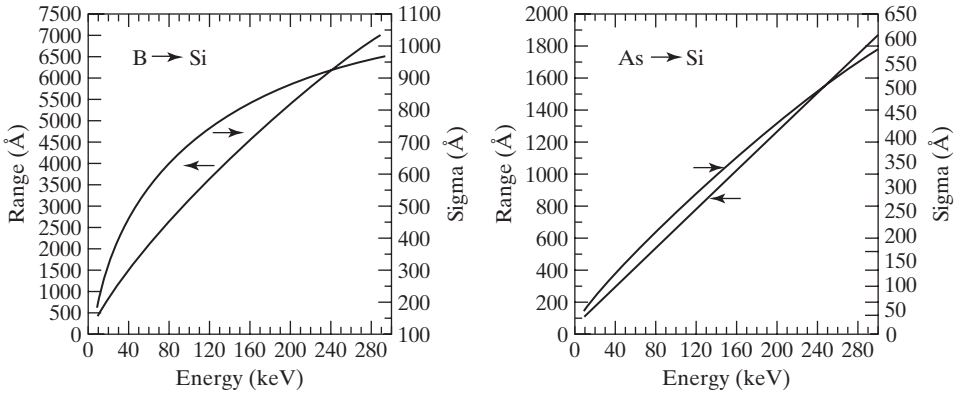


FIGURE 3–12 R and ΔR of implantation of (a) B and (b) As in silicon, versus energy [5].

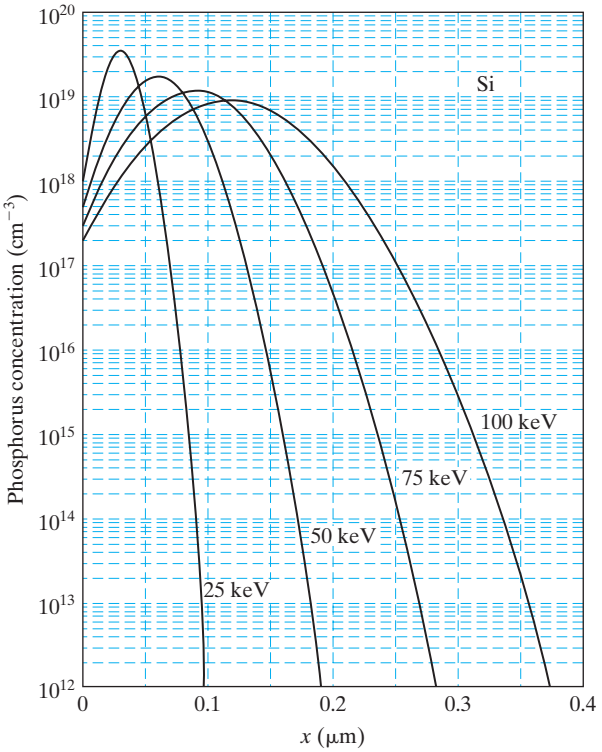


FIGURE 3–13 Computed implantation profiles of phosphorus assuming a constant dose of $10^{14}/cm^2$ [6].

3.6 • DOPANT DIFFUSION •

After dopant introduction by implantation or gaseous deposition, we may want to drive the dopant deeper into silicon. This is accomplished by **diffusion**. Unwanted diffusion also may occur during the post-implant anneal. The diffusion process is illustrated in Fig. 3–14. The dopant impurity diffuses with time at high temperature. If the diffusing dopant is of the opposite doping type to the substrate, as shown in Fig. 3–14, a line may be drawn to indicate the boundary where $N_a = N_d$. This structure is known as a **PN junction**, and the thickness of the diffusion layer is called the **junction depth**. For some applications, very deep junctions are desired. For other important applications, the shallowest possible junction is desired. Excessive diffusion is often the undesirable side effect of the necessary post-implantation anneal. In either case, it is important to control diffusion tightly.

Regardless of whether the shallow dopant addition is carried out by implantation or gaseous predeposition, the impurity concentration versus position inside the semiconductor after sufficient diffusion can be shown to be Gaussian [4].

$$N(x, t) = \frac{N_0}{\sqrt{\pi D t}} e^{-x^2/4Dt} \tag{3.6.1}$$

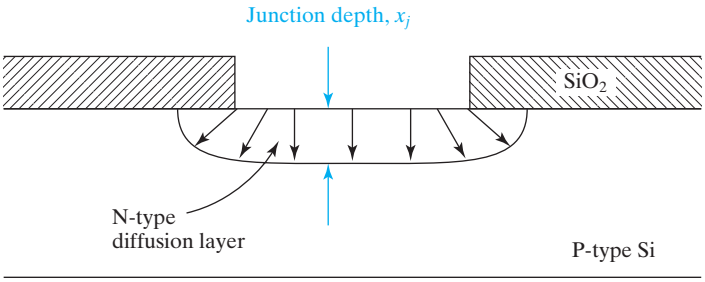


FIGURE 3–14 The basic diffusion process.

N_0 is the number of dopants per square centimeter and is determined by the dopant addition step, x is the distance into the semiconductor measured from the semiconductor surface, $N(x, t)$ is the impurity concentration at a depth x after a given time t, D is the **diffusivity** for the given impurity and furnace temperature, and t is the time for the diffusion step. Figure 3–15 shows the diffusivities of some common dopants in silicon. The diffusion rate increases with increasing temperature.

Diffusion is commonly performed in an open tube system similar in construction to that used for oxidation (Figs. 3–2 and 3–3). Diffusion temperatures range from roughly 900 °C to 1,200 °C. Sometimes the term **diffusion** refers to the combined process of gaseous dopant deposition and diffusion. The gaseous dopant deposition step is followed by a second step where the gaseous dopant source is

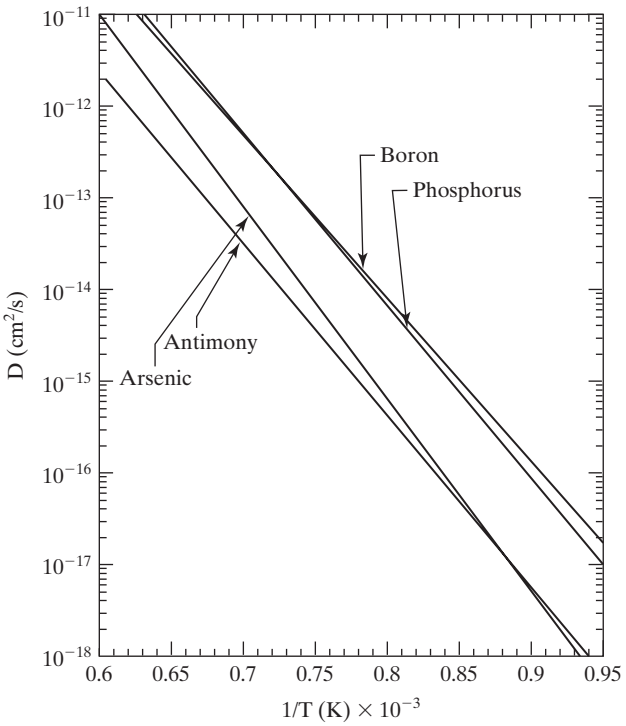


FIGURE 3–15 Diffusivity versus $1/T$ for Sb, As, B, and P in silicon. (From [5].)

• Dopant Diffusion and Carrier Diffusion •

The dopant diffusivity has the same dimension as the electron or hole diffusion constant, square centimeter per second. Their values, however, differ by a huge factor. Even at a high temperature, dopants only diffuse a small distance in an hour. Fortunately, the dopant diffusivities are negligibly small at room temperature. Otherwise, the device structures would change with time after they have been fabricated!

shut off, and the impurities are driven deeper into the semiconductor. The portion of the process step with the source present is called the **predeposition**, and the latter portion with the source shut off is called the **drive-in**.

• Shallow Junctions and Rapid Thermal Annealing •

High-performance devices often require that the junction depth (see Fig. 3–14) be kept shallow. This in turn requires that the Dt product in Eq. (3.6.1) be minimized. However, in order to activate the dopant and repair the crystal damage after ion implantation, thermal annealing is required. Unfortunately, **furnace annealing** may need 30 min in a furnace at 900 °C. This condition causes too much diffusion of the dopant, especially with B.

As it turns out, annealing can be completed at 1,050 °C in 20 s, which condition causes much less diffusion. In order to heat the silicon wafer up (and to cool it off) rapidly for short-duration annealing, a special heating technique is required. In **rapid thermal annealing (RTA)**, the silicon wafer is heated to high temperature in seconds by a bank of heat lamps. Cooling off is also fast because the thermal mass of the entire system is small. Similar systems can be used for **rapid thermal oxidation** and **rapid thermal chemical vapor deposition (CVD)** (see Section 3.7). Together, they are called **rapid thermal processing (RTP)**.

Pushing RTA further to 0.1 s annealing, one can obtain even shallower junctions. Such short annealing is called **flash annealing**. For even shorter durations (less than a microsecond) of heating, the silicon wafer can be heated with very short laser pulses. The process is called **laser annealing**, which may or may not involve melting a very thin layer of silicon.

As it turns out, crystal damage caused by ion implantation raises the dopant diffusivity at lower temperatures to values much larger than those shown in Fig. 3–15. This is called **transient enhanced diffusion** or **TED**. As a result, it is difficult to make ultra-shallow junctions using furnace annealing. The term ***transient*** denotes the fact that the enhancement of diffusion disappears after a short time during which the crystal damage is annealed out.

3.7 • THIN-FILM DEPOSITION •

Silicon nitride, silicon dioxide, Si, and many types of metal thin films are deposited during IC fabrication. Deposited films are usually not single crystalline.

Three Kinds of Solid

A solid material may be **crystalline**, **polycrystalline**, or **amorphous**. They are illustrated in Fig. 3–16. A crystalline structure has nearly perfect periodic structure as described in Section 1.1. Silicon wafers and epitaxially deposited films (see Section 3.7.3) fall in this category as do high-quality gemstones such as ruby and sapphire (Al_2O_3 with impurities that produce the characteristic colors) as well as diamond.

Often, materials are **polycrystalline**, which means the material is made of densely packed crystallites or grains of single crystal. Each grain has a more or less random orientation. The interface between crystallites is called a **grain boundary**. Each grain may be 10–10,000 nm in size. Metal films and Si films deposited at higher temperatures fall in this category, as do all metal objects that we encounter in daily life. Because each grain contains a large number of atoms, *polycrystalline materials have basically the same properties as single crystalline materials. In particular, polycrystalline and crystalline silicon have qualitatively similar electronic properties.*

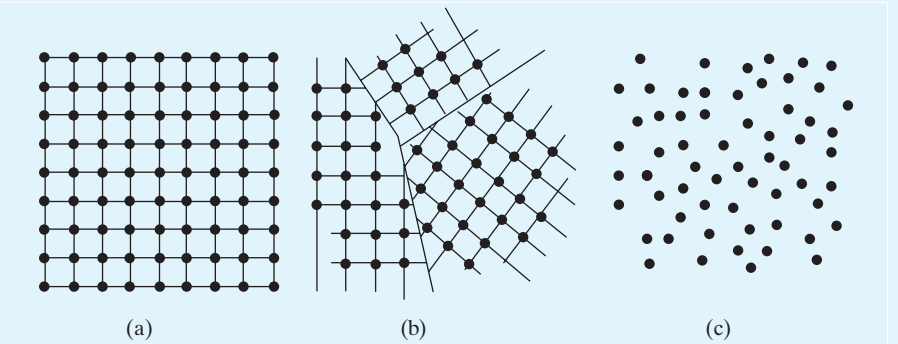


FIGURE 3–16 Crystalline material (a) has perfect ordering. Polycrystalline material (b) is made of tiny crystalline grains. (c) Amorphous material has no significant ordering.

An **amorphous** material has no atomic or molecular ordering to speak of. It may be thought of as a liquid with its molecules frozen in space. Thermally grown or deposited SiO_2, silicon nitride, and Si deposited at low temperature fall in this category. At high temperature, Si atoms have enough mobility to move and form crystallites on the substrate.

Carrier mobilities are lower in amorphous and polycrystalline Si than in single-crystalline Si. However, transistors of lower performance levels can be made of amorphous or polycrystalline Si, and are widely used in flat-panel computer monitors and other displays. They are called **thin-film transistors** or **TFTs**. The are also used in solar cells presented in Chapter 4.

3.7.1 Sputtering

Sputtering is performed in a vacuum chamber. The source material, called the **sputtering target**, and the substrate holding the Si wafer form opposing parallel plates connected to a high-voltage power supply. During deposition, the chamber is

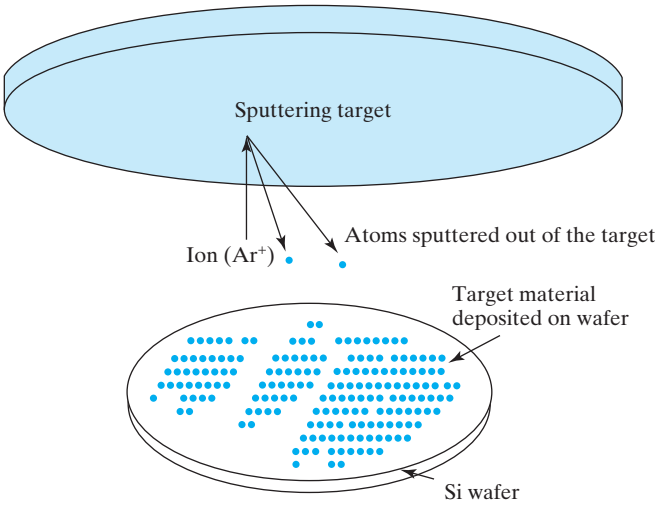


FIGURE 3–17 Schematic illustration of the sputtering process.

first evacuated of air and then a low-pressure amount of sputtering gas (typically Ar) is admitted into the chamber. Applying an interelectrode voltage ionizes the Ar gas and creates a plasma between the plates. The target is maintained at a negative potential relative to the substrate, and Ar ions are accelerated toward the sputtering target. The impacting Ar ions cause target atoms or molecules to be ejected from the target. The ejected atoms or molecules readily travel to the substrate, where they form the desired thin film. A simplified illustration of the sputtering process is shown in Fig. 3–17. A DC power supply can be used when depositing metals, but an RF supply is necessary when depositing insulating films. Sputtering may be combined with a chemical reaction in **reactive sputtering**. For example, when Ti is sputtered in a nitrogen-containing plasma, a TiN (titanium nitride) film is deposited on the Si wafer. Sputtering is the chief method of depositing Al and other metals. Sputtering is sometimes called a method of **physical vapor deposition (PVD)**.

3.7.2 Chemical Vapor Deposition (CVD)

While sputtering is a relatively simple and satisfactory way of depositing thin film over flat surfaces, it is directional and cannot deposit uniform films on the vertical walls of holes or steps in the surface topography. This is called a **step coverage** problem. CVD, on the other hand, deposits a much more ***conformal*** film, which covers the vertical and horizontal surfaces with basically no difference in the film thickness.

In CVD, the thin film is formed from gas-phase components. Either a compound decomposes to form the thin film or a reaction between gas components takes place to form it. A schematic of the CVD process is shown in Fig. 3–18. The CVD process is routinely used to deposit films of SiO_2, Si_3N_4 (a dielectric with excellent chemical and electrical stability), and polycrystalline silicon or poly-Si (see the sidebar "Three Kinds of Solid").

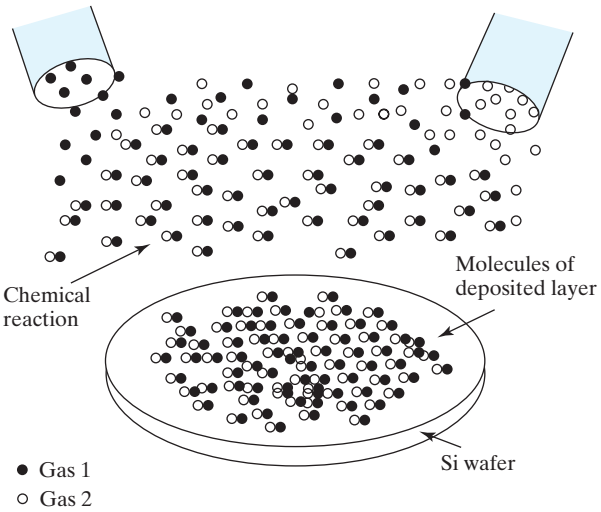


FIGURE 3–18 Chemical vapor deposition process.

These are some commonly used chemical reactors in the CVD deposition process:

Poly-Si: SiH_4 (silane) $\rightarrow Si + 2H_2$

Si_3N_4: $3SiH_2Cl_2$ (dichlorosilane) $+ 4NH_3 \rightarrow Si_3N_4 + 6HCl + 6H_2$

SiO_2: $SiH_4 + O_2 \rightarrow SiO_2 + 2H_2$

High-temperature SiO_2: $SiH_2Cl_2 + 2H_2O \rightarrow SiO_2 + 2HCl + 2H_2$

A **high-temperature oxide (HTO)** is particularly conformal because the high deposition temperature promotes particle movement on the surface so that even sidewall coverage is excellent. Commonly used CVD processes include **low-pressure chemical vapor deposition** or **LPCVD**, and **plasma-enhanced chemical vapor deposition** or **PECVD** processes. Low pressure offers better thickness uniformity and lower gas consumption. A simple LPCVD deposition system is illustrated in Fig. 3–19a. In PECVD, the electrons in the plasma impart energy to the reaction gases, thereby enhancing the reactions and permitting lower deposition temperatures. Figure 3–19b shows the schematic of a PECVD reactor.

Dopant species can be introduced during the CVD deposition of Si. This doping process is called ***in situ*** **doping** and is a method of heavily doping the Si film.

3.7.3 Epitaxy

Epitaxy is a very special type of thin-film deposition technology [7]. Whereas the deposition methods described in the preceding section yield either amorphous or polycrystalline films, **epitaxy** produces a crystalline layer over a crystalline substrate. The film is an extension of the underlying crystal. In a CVD reactor with special precautions to eliminate any trace of oxide at the substrate surface and at sufficiently high temperature, an arriving atom can move over the surface till it stops at a correct location to perfectly extend the lattice pattern of the substrate crystal. Figure 3–20a illustrates the epitaxy process. Selective epitaxy (Fig. 3–20b) is

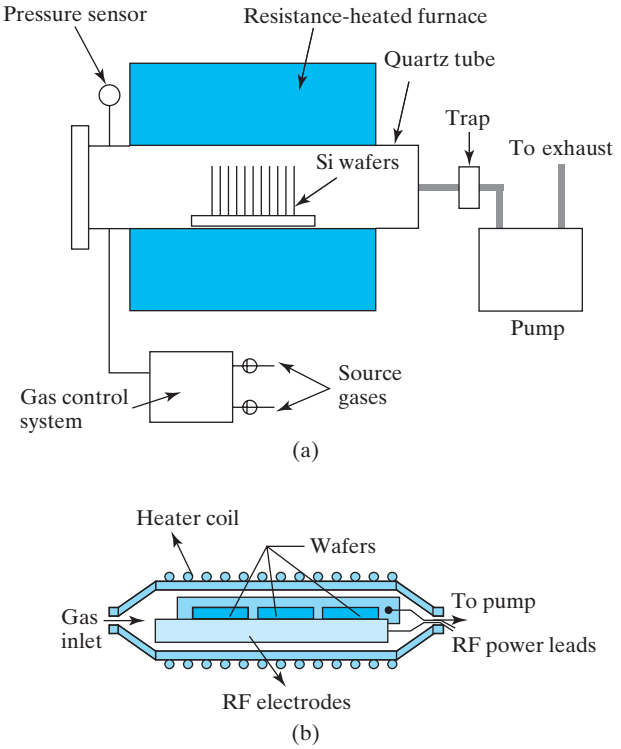


FIGURE 3–19 Schematic illustration of (a) an LPCVD system (after [1]) and (b) a PECVD reactor chamber with plasma generated radio-frequency power.

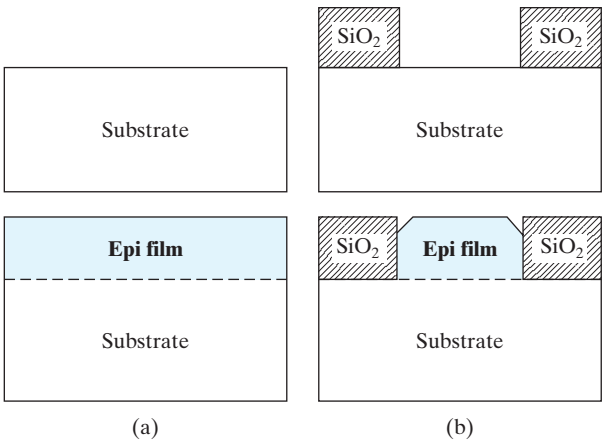


FIGURE 3–20 (a) Epitaxial and (b) selective epitaxial deposition of single crystalline film.

a variation of the basic epitaxy technology and has interesting device applications. In selective epitaxy deposition, an etching gas is introduced to simultaneously etch away the material. The net deposition rate is positive, i.e., atoms are deposited, only

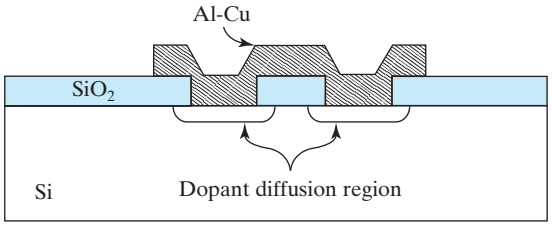
over the single crystal substrate. There is no net deposition over the oxide mask because the deposition rate over the oxide is lower than the etching rate.

Epitaxy is useful when we want a lightly doped layer of crystal Si over a heavily doped substrate (see Fig. 8–22). Also, a different material may be epitaxially deposited over the substrate material as long as the film and the substrate have closely matched lattice constants (see Section 1.1). Epitaxially grown dissimilar materials are widely used in light-emitting diodes (see Fig. 4–30) and diode lasers (see Fig. 4–33). The interface between two different semiconductors is called a **hetero-junction**. An application example of selective hetero-junction epitaxial growth (of SiGe over Si) may be found in Fig. 7–1.

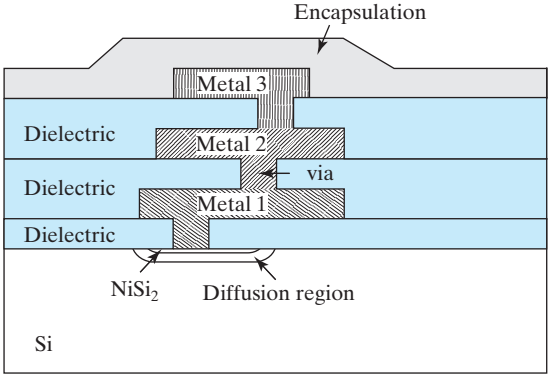
3.8 • INTERCONNECT—THE BACK-END PROCESS •

To build an IC, the individual devices must be interconnected by metal lines. This procedure is sometimes called **metallization**.

A basic **interconnect** is illustrated in Fig. 3–21a. First, the SiO_2 is removed from areas where a contact is to be made with the silicon. Then a layer of metal is deposited over the surface, typically by sputtering. The metal, perhaps aluminum, is then removed from areas where it is not desired (by lithography and dry etching). The metal interconnect in Fig. 3–21a performs the function of connecting the two diffusion regions.



(a)



(b)

FIGURE 3–21 Schematic drawing of device interconnections: (a) a basic metallization example and (b) a multilevel metallization structure.

To build complex and dense circuits, the **multilevel metallization** structure shown in Fig. 3–21b is routinely employed. Up to about ten metal layers may be used. The metal thickness ranges from a small fraction of a micron to several microns. The thinner interconnects route signals while the thicker layers serve as power lines. The adjacent layers of metal are separated by **intermetal dielectric** layers. Electrical connection between the adjacent metal layers is made through a **via**. To reduce the contact resistance (see Section 4.21) between the metal and the N^+ or P^+ diffusion region, a silicide such as $NiSi_2$ is added. An interconnect structure with all the dielectric etched away is shown in Fig. 3–22.

From the first ICs, the interconnect metal has been aluminum, Al. Al interconnects suffer a potential reliability problem called **electromigration**. Electron flow in the metal line, over time, can cause the metal atoms to migrate along crystal grain boundaries or the metal/dielectric interfaces in a quasi-random manner. Voids may develop in the metal lines as a result and cause the line resistance to increase or even become open-circuited. Copper has replaced Al as the interconnect material in advanced ICs. Cu has excellent electromigration reliability and 40% lower resistance than Al. Copper may be deposited by plating or CVD. Because dry etching of Cu is difficult, copper patterns are commonly defined by a **damascene** process, which is illustrated in Fig. 3–23.

Because Cu diffuses rapidly in dielectrics, a barrier material such as TiN is deposited as a liner before Cu is deposited in Fig. 3–23c. Excess copper is removed by **chemical-mechanical polishing** or **CMP**. In CMP, a polishing pad and slurry are used to polish away material and leave a very flat surface.

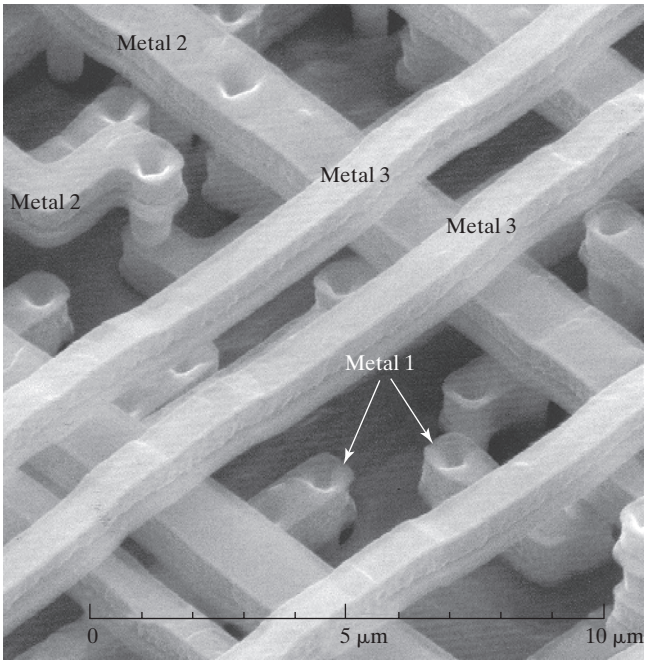


FIGURE 3–22 An example of a metal interconnect system. (Courtesy of Analytical Laboratory Services, Inc.)

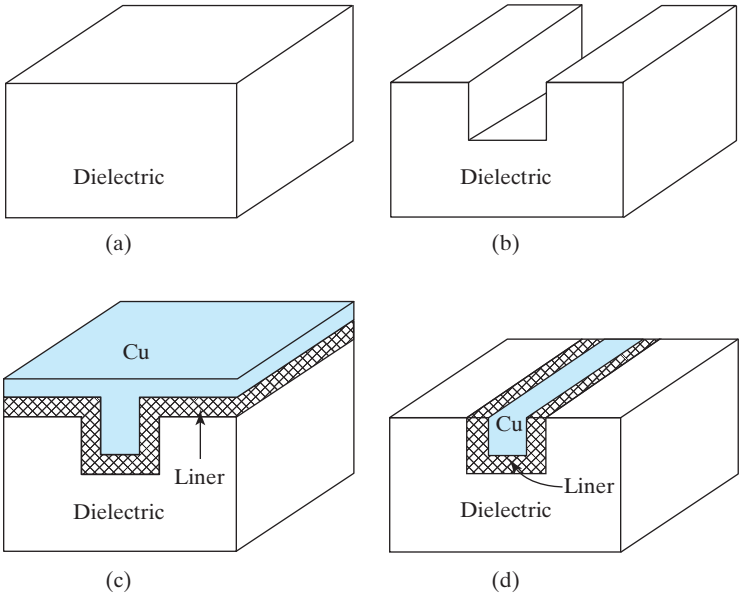


FIGURE 3–23 Basic steps of forming a copper interconnect line using the damascene process: (a) cover the wafer with a dielectric such as SiO_2; (b) etch a trench in the dielectric; (c) deposit a liner film and then deposit Cu; and (d) polish away the excess metal by CMP.

The dielectric material between the metal layers used to be SiO_2. It has been supplemented with **low-*k* dielectrics**, which often contain carbon or fluorine, and are designed to have much lower dielectric constants (k) than SiO_2. Lower k leads to lower capacitances between the interconnects. This is highly desirable because capacitance in a circuit slows down the circuit speed, raises power consumption (see Sections 6.7.2 and 6.7.3), and introduces **cross talk** between neighboring interconnect lines.

Since a large number of metal layers and process steps are involved, making the interconnects *consumes a large part of the IC fabrication budget*. This part of the fabrication process is called the **back-end process**. In contrast, the steps used to produce the transistors are called the **front-end process**.

• Planarization •

A flat surface is highly desirable in IC processing because it greatly improves subsequent optical lithography (the whole surface is in focus) and etching. For this reason, CMP **planarization** may also be performed in the front-end process, for example, in the formation of the shallow trench isolation (see Fig. 6–1). Although there are several ways to perform planarization, CMP provides the best flatness.

3.9 • TESTING, ASSEMBLY, AND QUALIFICATION •

After the wafer fabrication process is completed, individual ICs are electrically probed on the wafer to determine which IC chips are functional. The rest are marked and will not be packaged.

After this preliminary functional testing, the wafer is diced into individual circuits or chips by sawing or laser cutting. Functional chips may be encased in plastic or ceramic packages or directly attached to circuit boards. Multiple chips may be put in one package to make **multi-chip modules**. The electrical connections between the chip and the package are made by automated wire bonding or through **solder bumps**. In the solder bump process, the metal pads on the IC chip are aligned with the matching pads on the ceramic package. All connections are simultaneously made by melting preformed solder bumps on the IC pads in what is called the **flip-chip** bonding process. Finally, the package is sealed with a ceramic or metal cover before it undergoes final at-speed testing. As the complexity of ICs increases, testing becomes more and more difficult and expensive. Ease of testing is an important consideration in circuit design.

The quality of manufacturing and the reliability of the technology are verified with a **qualification** routine performed on hundreds to thousands of product samples including an **operating life test** that lasts over one thousand hours. This process is long and onerous but the alternative, shipping unreliable parts, is unthinkable. To ensure a very high level of reliability, every chip may be subjected to **burn-in** at higher-than-normal voltage and temperature. The purpose is to accelerate failures in order to weed out the unreliable chips.

3.10 • CHAPTER SUMMARY—A DEVICE FABRICATION EXAMPLE •

Figure 3–24 illustrates how the individual fabrication steps are combined and sequenced to fabricate a simple PN diode. A typical IC fabrication process involves over one hundred steps.

The starting point is a flat, P-type single-crystal Si wafer. A preclean removes all particulates, organic film, and adsorbed metal from the semiconductor surface. Then a thermal oxide is grown. Step 2 is a lithography process performed to open a hole in the oxide that will eventually become the position of the PN junction.

The wafer is implanted with an appropriate dose of As at an appropriate energy (step 4). After annealing and diffusion, the junction is formed in step 5. Note that the junction edge is protected by the oxide. Some oxide may be formed in the diffusion process. This must be cleaned off before step 6, metallization.

Sputtering of Al deposits a thin metal film over the entire surface of the wafer as pictured in step 6. A lithography process (step 7) is then performed to pattern the metal. A low-temperature (≤450 °C) anneal is performed to produce a low-resistance contact between the metal and Si. In step 9, SiO_2 and Si_3N_4 films are deposited for encapsulation to protect the device from moisture and other contaminants. In step 10, an opening is made to access the Al for wire bonding. If electrical contact is to be made to the P-type substrate, the oxide grown on the back of the wafer in step 2 must be removed while the front of the wafer is protected with photoresist as shown in step 11. Gold (Au) is deposited at the back of the wafer for electrical contact in step 12. Finally, the wafer is diced into individual diode chips, and each chip is soldered to a package; a bond wire connects the Al to a second electrical lead. For a slide show of the device fabrication steps, see http://jas.eng.buffalo.edu/education/fab/pn/diodeframe.html.

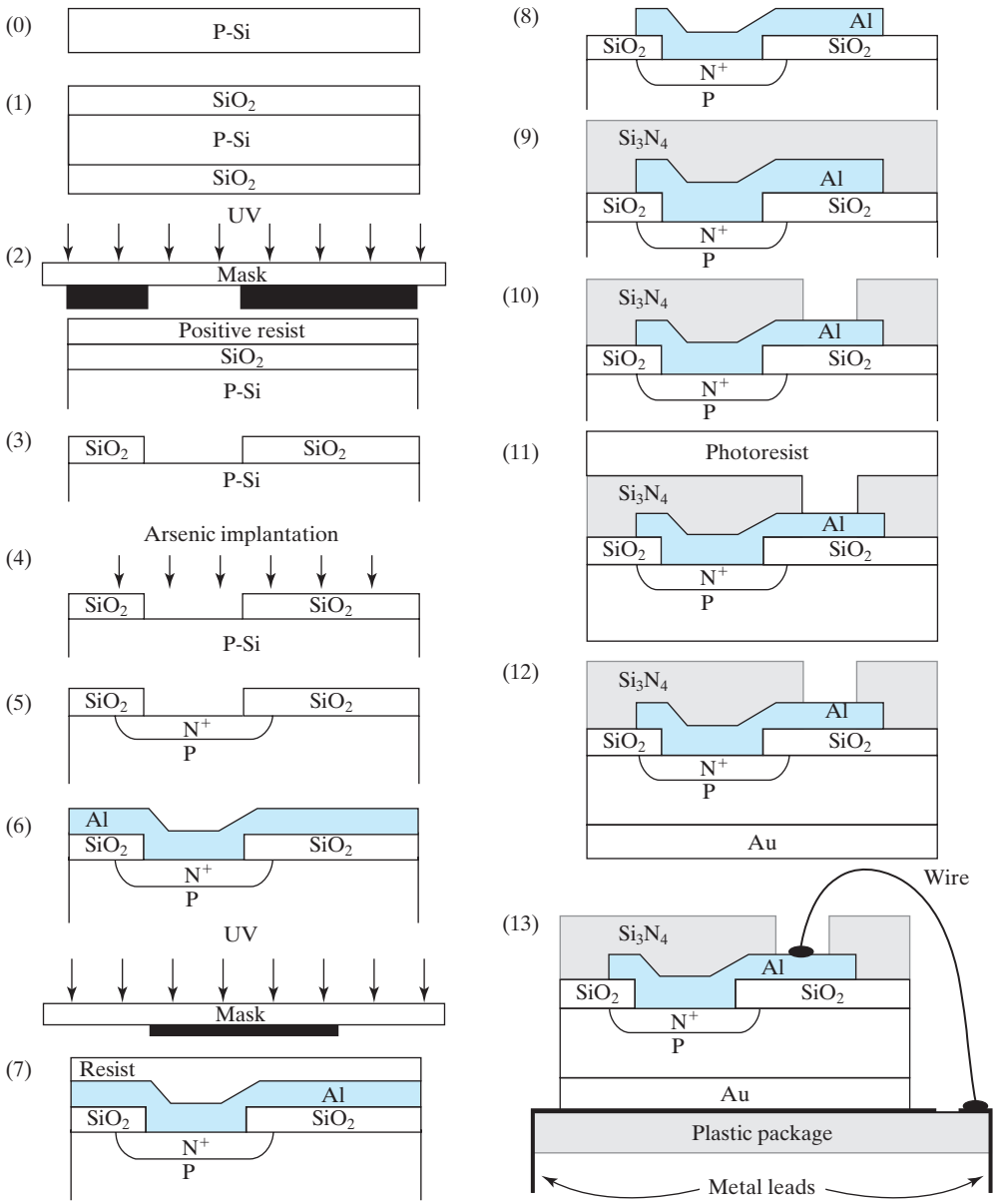


FIGURE 3–24 Graphical summary of the major processing steps in the formation of a PN junction diode. (0) Start; (1) oxidation; (2) lithography; (3) oxide etching; (4) As implantation; (5) annealing and diffusion; (6) sputtering Al; (7) lithography; (8) metal etching; (9) CVD nitride deposition; (10) lithography and bonding window etching; (11) removal of oxide from back side of wafer; (12) deposition of Au on back side; and (13) dicing and packaging. (After [6].)

PROBLEMS

Terminology and General Knowledge

3.1 Copy all the bold-faced terms in Chapter 3 Introduction and Sections 3.1–3.5. Give each of them a short definition or explanation (one word to two sentences), preferably in your own words.

3.2 Do Problem 3.1 for all the bold-faced terms in the remaining sections of Chapter 3.

3.3 Answer each of the following questions in one to three sentences.

(a) What is lithography field?

(b) What is misalignment in lithography?

(c) What is selectivity in an etching process?

(d) What is end-point detection in an etching process?

3.4 Answer the following questions.

(a) In an older MOSFET technology, the field oxide is a 1-μ thick thermal oxide. Would you grow it in a dry or wet ambient? Why?

(b) For etching a small feature with faithful replication of the resist pattern, is dry or wet etching technique preferred? Why?

(c) If the junction depth is to be kept as small as possible, which ion species would you use to make a P–N junction (for an ion implantation process on a P-type silicon substrate)? Give reasons to support your answer.

(d) If you want to deposit oxide at the lowest possible temperature, what processing technology would you use?

(e) What processing technology would you use to deposit aluminum? What is the processing technology you would use to etch a fine aluminum line? What chemicals are involved?

(f)

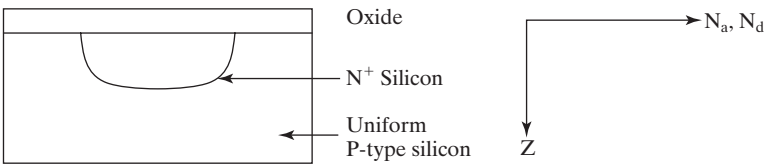


FIGURE 3–25

In the accompanying N_a N_d vs. Z coordinates (Fig. 3–25), quantitatively draw typical N_a and N_d profiles through the P–N junction and indicate the position of the junction. Assume the N^+ dopant peak is at the $Si–SiO_2$ interface.

Oxidation

3.5 Why is wet oxidation faster than dry oxidation? Please speculate. One or two sentences will be sufficient.

3.6 Assume that the oxide thickness is T_{init} at time 0 and that the oxide thickness is given by $T_{ox}^2 + A T_{ox} = B(t + \tau)$, where $\tau \equiv \dfrac{T_{init}^2 + A T_{init}}{B}$. (For example, see 900°C wet oxidation curve in Fig. 3–4.)

(a) Calculate the final thickness of the silicon dioxide on a wafer that initially has 0.2 μm after an additional 3 h of 1,000°C dry oxidation ($A = 0.165$ μm and $B = 0.0117$ μm²/h at 1,000°C dry oxidation).

(b) There are two important limiting cases for this equation. For sufficiently thin oxides, the quadratic term is negligible. On the other hand, if the oxide is sufficiently thick, the linear term can be ignored. How much error is introduced if this question is answered with the linear approximation and the quadratic approximation?

• Deposition •

3.7 Verify that chemical equations in Section 3.7.2 are balanced. If some are not balanced, correct them by providing the right coefficients.

• Diffusion •

3.8 Assume $x_j = C\sqrt{Dt}$, where C may be assumed to be 1.

(a) Show that additional diffusion with an increment $\Delta(Dt)$ would increase the junction depth by $\Delta(Dt)/2x_j$.

(b) If a boron doped junction has a depth of $x_j = 0.1$ μm, by how much will x_j increase at 500 K in 10 years?

3.9 Assume $D = D_0 e^{-E_a/kT}$ is the diffusion coefficient of boron in silicon surface, where $D_0 = 10.5$ cm²/s and $E_a = 3.7$ eV. The substrate is N-type silicon doped to 10^{15} cm^3. $N_0 = 10^{15}$ cm^2 of boron is introduced just below the silicon surface.

(a) What is the junction depth after a 1-h drive-in at 1,100°C?

(b) By how much will the junction depth change after 10^6 h (~100 years) of operation at 100°C?

• Visualization •

3.10 For the following process steps, assume that you use a positive photoresist and that etch selectivity is infinite. A composite plot of four photomasks is given in Fig. 3–26. Assume that mask alignment is perfect. All contact sizes are 0.5 × 0.5 μm. The poly 1 and poly 2 areas are opaque, and the contact 1 and contact 2 areas are clear in the masks. Draw the cross section at the end of each process step along the cut line shown in the figure.

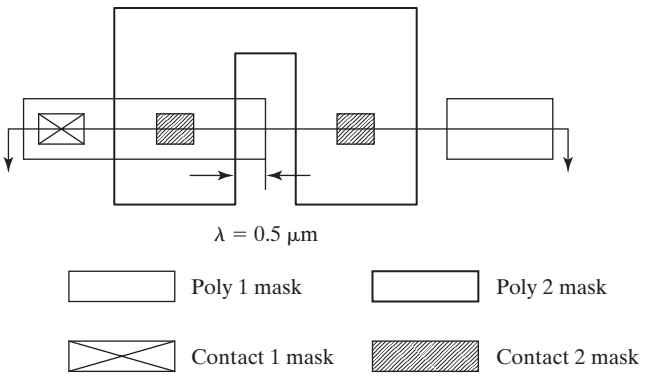


FIGURE 3–26

(a) Grow 1 μm thermal oxide on < 100 > bare Si wafer.

(b) Expose and develop photoresist with contact 1 mask. Assume that the resist thickness is 1 μm.

(c) Etch the 1 μm thermal oxide anisotropically. Assume the final oxide profile is perfectly vertical.

(d) Remove the photoresist with O_2 plasma.

(e) Implant phosphorus and anneal. Assume that the final junction depth is 0.3 μm.

(f) Deposit 1 μm *in situ* doped poly silicon by LPCVD. The thickness on the sidewalls is the same as that on the flat surface.

(g) Expose and develop the photoresist with poly 1 mask.

(h) Etch the 1 μm poly silicon anisotropically.

(i) Remove the photoresist with O_2 plasma.

(j) Deposit 1 μm oxide with PECVD. Again, the thickness on the sidewalls is the same as that on the flat surface.

(k) Expose and develop photoresist with contact 2 mask.

(l) Etch 0.2 μm of the PECVD oxide with HF. Assume the profile is cylindrical as shown in Fig. 3–8a.

(m) Etch the remaining 1.8 μm oxide anisotropically.

(n) Remove the photoresist with O_2 plasma.

(o) Implant phosphorus and anneal. Assume the junction depth is 0.3 μm and there is no additional dopant diffusion.

(p) Deposit 1 μm *in situ* doped poly silicon by LPCVD. The thickness on the sidewalls is the same as that on the flat surface.

(q) Expose and develop photoresist with poly 2 mask.

(r) Etch the 1.0 μm poly silicon anisotropically.

(s) Remove the photoresist with O_2 plasma.

(This is just an exercise. The structure does not have any known usefulness.)

3.11 Assume a negative resist is used instead of a positive resist in Problem 3.10 with the same contact 1 mask. Answer parts (a), (b), (c), and (d) of Problem 3.10. What changes does one have to make in order to obtain the same cross section as Problem 3.10 (d) with a negative resist?

REFERENCES

1. Dance, B. "Europe Prepares Its Future Technology," *Semiconductor International* (1995), 125.

2. Jaeger, R. C. *Introduction to Microelectronic Fabrication*, Vol. 5, in *The Modular Series on Solid State Devices,* 2nd ed., G. W. Neudeck and R. F. Pierret. Reading, MA: Addison-Wesley, 2002, pp. 49, 24.

3. Warren, J. "Leaping into the Unknown with 0.18 mm," *Semiconductor International* (1998), 111.

4. Runyan, W. R., and K. E. Bean. *Semiconductor Integrated Circuit Processing Technology*, Reading, MA: Addison-Wesley, 1990.

5. Sze, S. M. *VLSI Technology*, 2nd ed. New York: McGraw-Hill Book Company, 1988.

6. Pierret, R. F. *Semiconductor Device Fundamentals.* Reading MA: Addison-Wesley Publishing Company Inc., 1996, 158 and 167.

7. Herman, M. A., W. Richter, and H. Sitter. *Epitaxy: Physical Foundation and Technical Implementation*. New York: Springer-Verlag, 2004.

8. Wolf, S., and R. N. Tauker. *Silicon Processing for the VLSI Era*, 2nd ed. Sunset Beach, CA: Lattice Press, 2000.

9. Burggraaf, P. "Stepping to Mix-and-Match I-line Lithography," *Semiconductor International* (1995), 47.

GENERAL REFERENCES

1. Sze, S. M. *Semiconductor Devices: Physics and Technology*, 2nd ed. New York: John Wiley & Sons, 2002.

2. Wolf, S., and R. N. Tauker. *Silicon Processing for the VLSI Era*, 2nd ed. Sunset Beach, CA: Lattice Press, 2000.