Chapter

11

Doping

Overview

One of the unique properties of semiconductor materials is that their conductivity and the type of conductivity (N or P) can be created and controlled. In this chapter, the formation of specific "pockets" of conductive regions and N-P in and on the wafer surface is described. The principles and practice of the two doping techniques, diffusion and ion-implantation, are explained.

Objectives

Upon the completion of this chapter, you should be able to:

1. Define an N-P junction.
2. Draw a flow diagram of a complete diffusion process.
3. List the three most common dopants used in silicon technology.
4. List the three types of deposition sources.
5. Draw a typical concentration-versus-distance curve for a deposition and drive-in.
6. List the major parts of an ion implanter.
7. Describe the principle of an ion implanter.
8. Compare the advantages and disadvantages of diffusion and ion implant processes.

Introduction

The structure that makes transistors and diodes work is an N-P or N-P junction. A *junction* is the separation between a region that is

TSMC-1011
U.S. Patent 8,796,779

rich in negative electrons (N-type region) and a region that is rich in holes (P-type region). The exact location of a junction is where the concentration of electrons equals the concentration of holes. This concept is explained later in the section, "Formation of a Doped Region and Junction by Diffusion."

The usual way to form junctions in the surface of semiconductor wafer is by thermal *diffusion* or by *ion implantation*. With thermal diffusion, dopant materials are introduced into the exposed top surface of the wafer, typically through a hole in the top silicon dioxide layer. With heating, they spread down into the bulk of the wafer. The amount and depth of the spread is governed by a set of rules, as explained below. These rules arise from a set of chemical rules that govern any movement of dopants in the wafer whenever the wafer is heated to a threshold temperature. In ion implantation, the dopant materials are literally shot into the wafer surface, with most of dopant atoms coming to rest below the surface. Additional movements of the implanted atoms are also governed by the rules of diffusion (Fig. 11.1). Thus, this chapter starts with a discussion of semiconductor junctions, goes on to the rules and techniques of diffusion, and ends with a description of the ion implant processes.

Formation of a Doped Region by Diffusion

A major advance in semiconductor production was the development of diffusion doping. Diffusion, the movement of one material through another, is a natural chemical process with many examples in everyday life. Two conditions are necessary for a diffusion to take place. First, one of the materials must be at a higher concentration than the other. Second, there must be sufficient energy in the system for the higher-concentration material to move into or through the other. An example of a gas-state diffusion is the action of a common pressurized spray can (Fig. 11.2) such as a room deodorant. When the nozzle is depressed, the pressurized material leaves the can and moves into the surrounding air. Thereafter, movement of the gas into the room proceeds by the process of diffusion. The movement takes place while the nozzle is depressed and continues after it is closed. The diffusion will continue as long as the advancing spray is at a concentration higher than that in the air. As the material moves away from the can, the concentration of the material becomes progressively less. This is a characteristic of a diffusion process. Diffusion will continue until the concentration is even throughout the room.

Another example of a liquid-state diffusion is represented when a drop of ink is dropped into a glass of water. The ink is more concentrated than the surrounding water and immediately starts diffusing

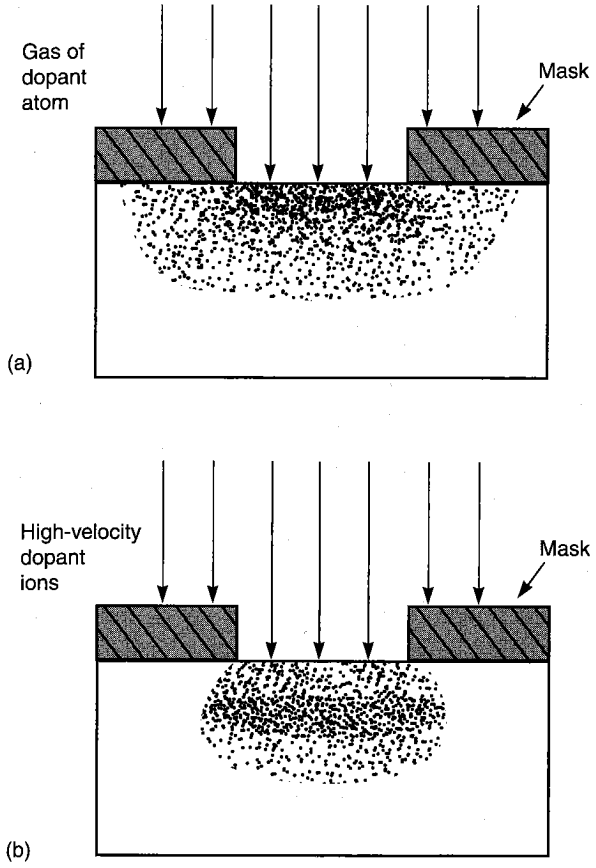


Figure 11.1 (*a*) Dopant concentration from diffusion and (*b*) dopant concentration from ion implantation.

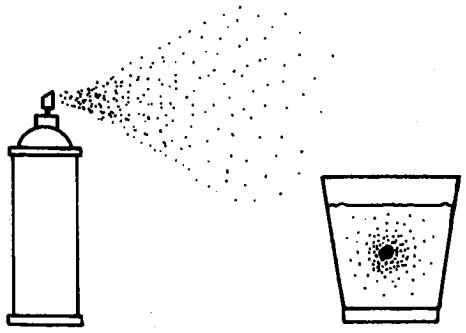


Figure 11.2 Examples of diffusion.

into the glass of water. The diffusion will continue until the whole glass of water is the same color. This example can be used to illustrate the influence of energy on the diffusion process. If the water in the glass is heated (giving the water more energy), the ink will spread into the water faster.

The same diffusion phenomenon takes place when a doped wafer is exposed to a concentration of atoms higher than the concentration in the wafer. This is called *solid state* diffusion.

Formation of a Doped Region and Junction by Diffusion

The formation of a doped region and junction is illustrated by examining the doping of a wafer in a diffusion process. The starting condition is depicted in Fig. 11.3. The wafer illustrated is from a P-type crystal. The + symbols in the diagram represent the P-type dopants that were incorporated into the crystal during the crystal-growing process. They are uniformly distributed throughout the wafer.

The wafer receives a thermal oxidation and a patterning process that leaves a hole in the oxide layer. In a diffusion tube, the wafer is exposed to a concentration of N-type dopants at a high temperature (the + symbols in Fig. 11.4). The N-type dopants diffuse through the hole in the oxide layer.

The effect in the wafer is illustrated by examining what happens at different levels in the wafer. The conditions in the diffusion tube are set such that the number of N-type dopant atoms that diffuse into the wafer surface are greater than the number of P-type atoms in layer no. 1. In the illustration, there are seven more N-type atoms than P-type atoms, and that level is electrically an N-type layer. This process has converted the top layer from P-type to N-type.

The diffusion process proceeds with N-type atoms diffusing from the first level down to the second level (Fig. 11.5). At the second level, there are again more N-type dopants than P-type dopants, converting level 2 to N-type. In the table (Fig. 11.6) is an accounting of the number of N-type and P-type atoms at each level. This process goes deeper into the wafer.

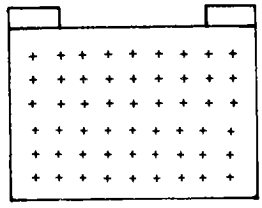


Figure 11.3 P-type wafer ready for diffusion.

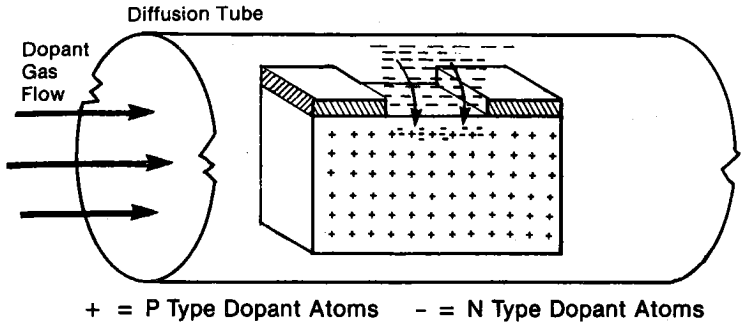


Figure 11.4 Start of a diffusion process.

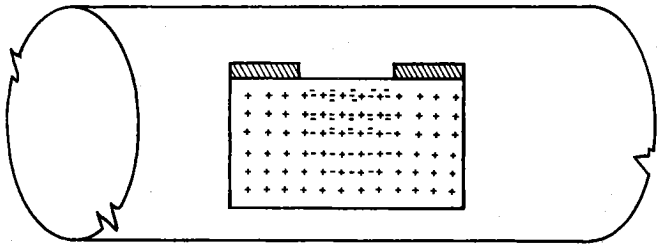


Figure 11.5 Cross section of wafer at conclusion of diffusion.

Layer	# N's (-)	# P's (+)	Net (N - P)	Layer
1	12	5	7	N
2	10	5	5	N
3	8	5	3	N
4	5	5	0	Jct
5	3	5	-1	P
6	0	5	-5	P

Figure 11.6 Dopant amounts and level conductivity type.

The N-P junction

At level 4, there are exactly the same number of N-type and P-type atoms. This level is the location of an N-P junction. The definition of an N-P junction is the location in the wafer where the number of N-type and P-type dopant atoms are equal. Note that, below the junction at level 5, there are three N-type atoms, which are not enough to convert that layer to N-type.

The term *N-P junction* indicates that there is a higher concentration of N-type dopants on one side of the junction. A P-N junction would indicate a higher P-type on one side of the junction. The behavior of elec-

trical currents across semiconductor junctions gives rise to the particular performance of individual semiconductor devices and is the subject of Chapter 14. In this chapter, the emphasis is on the formation and character of doped regions in a wafer.

Goals of solid-state diffusion

The goals of a diffusion process, thermal diffusion or ion-implantation, are threefold:

1. The creation of a specific number (concentration) of dopant atoms in the wafer surface
2. To create an N-P (or P-N) junction at a specific distance below the wafer surface
3. To create a specific distribution and concentration of dopant atoms in the wafer surface

Lateral diffusion

The diffusion doping process depicted in Fig. 11.5 shows the incoming dopant atoms traveling straight down into the wafer. In reality, the dopant atoms move in all directions. An accurate cross section (Fig. 11.7) would show that some of the atoms have moved in a lateral direction, forming a junction under the oxide barrier. This movement is also called *lateral* or *side diffusion*. The amount of the lateral or side diffusion is approximately 85 percent of the vertical junction depth. Lateral diffusion takes place regardless of whether the introduction was through diffusion or ion implantation. The effects of side diffusion on circuit density are discussed in the introduction to ion implantation.

Same-type doping

Some devices call for a doping with a dopant type the same as the host region. In other words, an N-type dopant will be put into an N-type

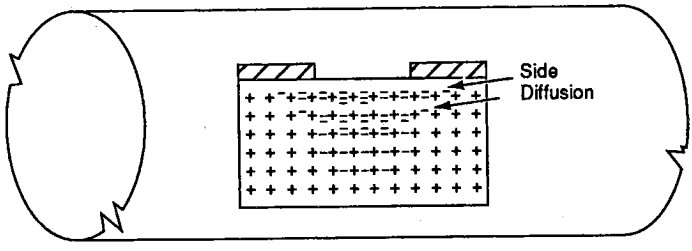


Figure 11.7 Side-diffused N-type dopants.

wafer, or a P-type dopant will be put into a P-type wafer (Fig. 11.8). When this situation happens, the added dopant atoms simply increase the concentration of the dopant atoms in the localized region. No junction is formed.

Graphical representation of junctions

In a cross section (Fig. 11.9) of a semiconductor device, the N-P junctions are indicated simply as regions in the device. There is no graphical convention to represent N- or P-type areas. The drawings just show the relative location of the doped region and the junction. This type of graphical representation gives little information about the concentration of the dopant atoms and only approximates the actual dimensions of the regions. A drawing of a 2-μm-deep junction in a 20-mm-thick wafer, scaled to an 8-ft wafer, would have a junction depth only 0.4 in thick!

Concentration versus depth graphs

Another two-dimensional graphical representation of a doped region is the concentration-versus-depth graph. The concentration is represented on the vertical axis and the depth into the wafer on the horizontal axis. An example of such a graph is illustrated in Fig. 11.10. The illustration uses the data from the doping example shown in Fig. 11.6. First, the P-type dopant concentration is plotted. In the example, there are exactly five P-type dopant atoms at every level, resulting in a straight horizontal line on the graph (Fig. 11.10*b*). Next, the number of N-type dopant atoms is plotted. Since the number of atoms decreases deeper into the wafer, the plotted line slopes down and to the right. At level 4, the number of N- and P-type dopants is

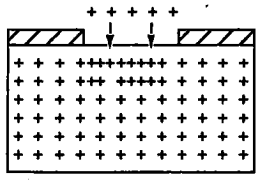


Figure 11.8 P-type doping of P-type wafer.



Figure 11.9 Cross-sectional representation of diffuse junctions.

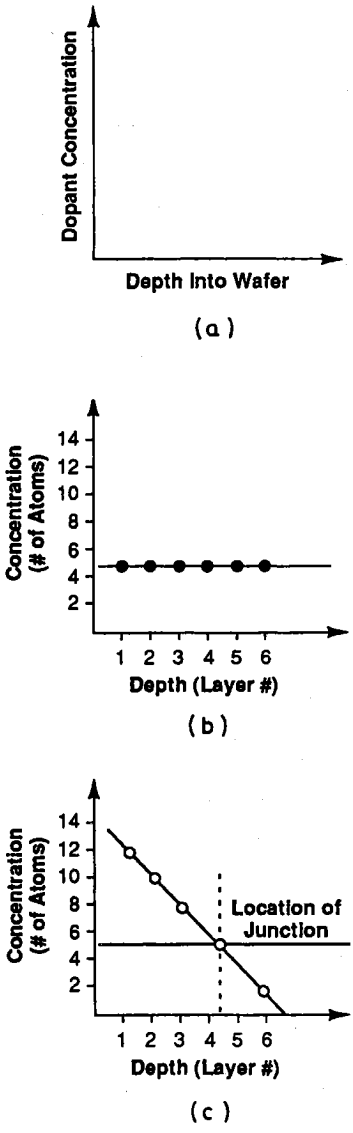


Figure 11.10 Construction of concentration-versus-depth curve. (*a*) Axes, (*b*) P-type dopant, and (*c*) N-type and P- type dopant.

equal, and the lines cross. This is a graphical representation of the location of the junction.

The graph of an incoming dopant concentration versus depth profile for an actual process is not a straight line. They are curved lines. The shape of the curve is determined by the physics of the dopant technique. The actual shapes are discussed in the deposition and drive-in sections.

Diffusion Process Steps

The use of *solid-state thermal diffusion* to create junctions in semiconductor wafers requires two steps. Step one is called *deposition,* and step number two is called *drive-in oxidation.* Both steps take place in a horizontal or vertical tube furnace. The equipment is the same described in Chapter 7 for oxidation.

Deposition

The first step of a diffusion process is called *deposition*; it is also called *predeposition*, *dep*, or *predep*. It takes place in a tube furnace, with the wafers placed in the flat zone of the tube. A source of dopant atoms is located in the source cabinet and their vapors are transferred into the tube at a required concentration (Fig. 11.11).

In the tube, the dopant atoms diffuse into the exposed wafer. Within the wafer, the dopant atoms move by two different mechanisms: vacancy and interstitial movement. In the vacancy model (Fig. 11.12*a*) the dopant atoms move by filling empty crystal positions, called *vacancies*. The second model (Fig. 11.12*b*) relies on interstitial movement of the dopant.[1] In this model, the dopant atom moves through the spaces between the crystal sites; that is, intersite.

A deposition process is controlled or limited by several factors. One is the *diffusivity* of the particular dopant. Diffusivity is the rate (speed) of movement of the dopant through the particular wafer material. The higher the diffusivity, the faster the dopant moves through the wafer. Diffusivity increases with temperature.

Another factor is the *maximum solid solubility* of the dopant in the wafer material. It is the maximum concentration of a specific dopant that can be put into the wafer. A familiar analogy is the maximum liquid solubility of sugar in coffee. The coffee can dissolve only a certain amount of sugar before collecting in the bottom of the cup as a solid.

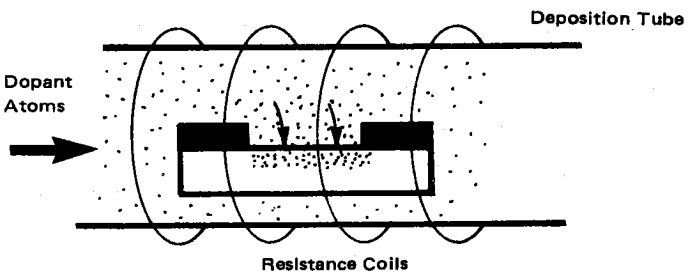


Figure 11.11 Deposition.

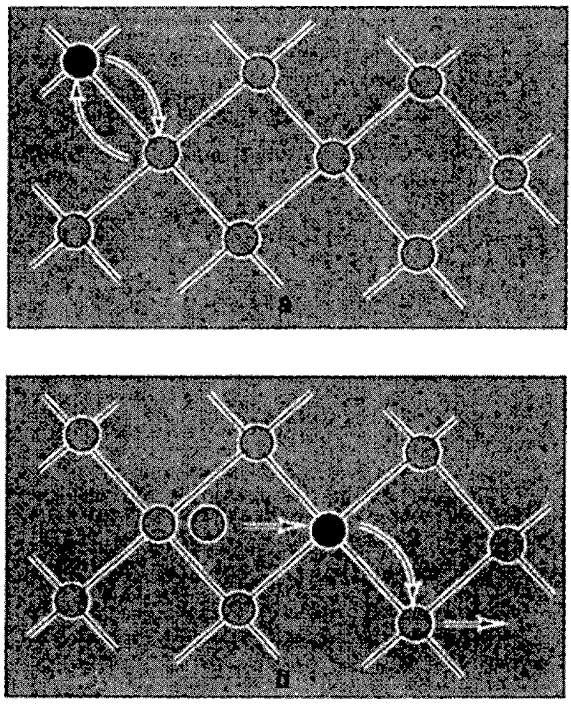


Figure 11.12 Diffusion models. (*a*) Vacancy model and (*b*) interstitial model.

The maximum solid solubility limit increases with increasing temperature.

In a semiconductor deposition step, the concentration of the dopant is purposely set higher than the maximum solid solubility of the dopant in the wafer material. This situation ensures that the maximum amount of the dopant will be accepted by the wafer.

The amount of dopant entering the wafer surface is a function of the temperature only, and the deposition is said to take place at solid solubility. Maximum solubility levels for various dopants in silicon are shown in Fig. 11.13.

The concentration of dopant atoms at each level in the wafer is an important factor in the performance of junction diodes and transistors. A dopant concentration versus depth curve for a deposition is shown in Fig. 11.14. The shape of the curve follows a mathematical formula specifically known as an *error function*. An important factor in device performance is the concentration of the dopant at the wafer surface. This is called the *surface concentration* and is the quantity indicated where the error function curve intersects the vertical axis. Another

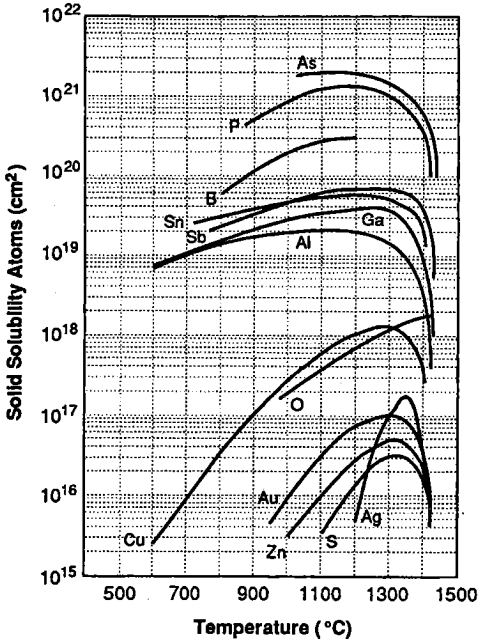


Figure 11.13 The solid solubility of impurities in silicon.

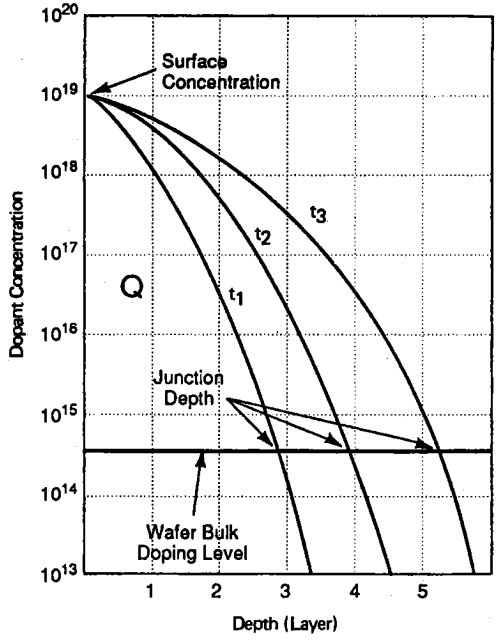


Figure 11.14 Typical deposition (error function) dopant profile for three different deposition times.

deposition parameter is the total quantity of atoms diffused into the wafer. This amount increases with the time of the deposition. Mathematically, the quantity of atoms (Q) is represented by the area under the error function curve.

Deposition steps

A deposition process requires four steps. They are:

1. Preclean and etch
2. Tube deposition
3. Deglaze
4. Evaluation

Preclean and etch. Wafers coming to the deposition station first receive a preclean to remove particulates and stains. The chemicals and processes are the same as in the preoxidation process cleaning. After the preclean, the wafers will be chemically etched in an HF or HF and water solution to remove any oxide that may have grown on the exposed silicon surface. An oxide can form on the silicon from exposure to the air and from the chemical preclean process. Removal of the oxide is essential to allow the dopants to diffuse unimpeded into the wafer surface. The etch time and concentration must be balanced to prevent the surface-blocking oxide from being removed or thinned too much.

Deposition. The deposition process, like oxidation, requires a minimum of three cycles. The first cycle is the loading cycle, which takes place in a nitrogen atmosphere. The second cycle is the actual doping cycle. The third cycle is the exit cycle, which also takes place in a nitrogen atmosphere.

Wafers are positioned on the boats at right angles (Fig. 11.15) or parallel to the tube axis. Right-angle placement is the highest packing density, but it can cause uniformity problems, because the wafers act as baffles to the gas flow. For uniform doping, the gas must mix uniformly between all of the wafers. Parallel placement offers the advantage of more uniform doping, since the doping gas proceeds unimpeded through the wafer boat. The disadvantage is a lower loading density. In both placement systems, dummy wafers are placed on the outsides and/or the front and back positions of the boat to create uniform doping of the interior device wafers.

Deglaze. During the deposition cycle, a thin oxide can form on the exposed wafer surface. This oxide is doped and can act as an unwanted

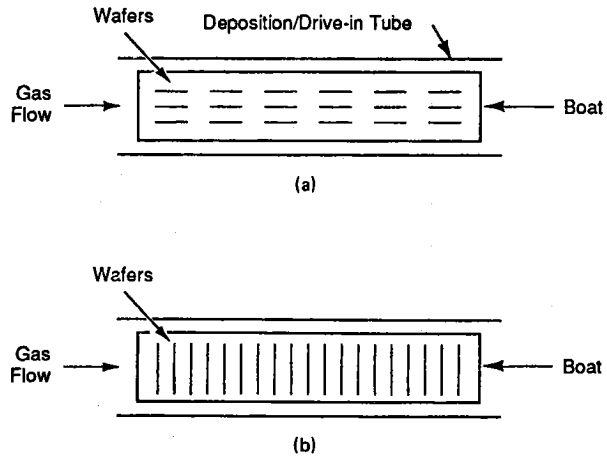


Figure 11.15 Boat loading patterns. (*a*) Parallel and (*b*) perpendicular.

source of dopant during the drive-in-oxidation step. Also, a deposition-created oxide can be difficult to etch, causing incomplete etch in a subsequent masking process. The oxide is removed from the surface by immersion in a diluted HF solution, followed by a water rinse and a drying step.

Evaluation. Test wafers are added to the boat of device. A deposition test wafer has no patterns and has a conductivity type opposite that of the dopant. They are placed on the deposition boat in locations to sample the deposition distribution throughout the wafer batch. After deglazing, the test wafers are evaluated. The primary in-line test is of the sheet resistance with a four-point probe or contactless apparatus. The concept of sheet resistance and the measurement techniques are explained in Chapter 14. The junction depth after deposition is very thin and is generally not measured at this point.

Certification of process and tube cleanliness is determined by oxidizing test wafers and performing a capacitance-voltage measurement for mobile ionic contaminants (Chapter 14). Device wafers are 100 percent or sample-inspected for surface contamination and/or stains with high-intensity ultraviolet lamps or microscopes.

Dopant sources

A deposition depends on the presence of a concentration of dopant atom vapors in the tube. The vapors are created from a dopant source located in the source cabinet of the tube furnace and passed into the

tube with a carrier gas. Dopant sources are either in liquid, gaseous, or solid states. Several dopant elements are available in more than one state (Fig. 11.16).[2]

Liquid sources. Liquid sources of dopants are chlorinated or brominated compounds of the desired element. Thus, a boron liquid source is boron tribromide (BBr_3) and a phosphorus liquid source is phosphorus oxychloride ($POCl_3$). The liquid sources are held in temperature-controlled quartz flasks (Fig. 11.17). An inert gas, such as nitrogen, is bubbled through the heated liquid, and the gas becomes saturated with the dopant vapors. The flask is connected to a gas manifold whose valves are controlled by a microprocessor. The nitrogen carries the dopant vapors into the tube and maintains a gas volume sufficient to create a laminar flow condition in the tube. Laminar flow is required to prevent the gas from spiraling in the tube, creating nonuniform doping. Another device used to create laminar flow is a baffle

Type	Element	Compound Name	Formula	State	Diffusion Reactions*
N	Antimony	Antimony Trioxide	Sb_2O_3	Solid	
	Arsenic	Arsenic Trioxide	As_2O_3	Solid	$2AsH_3 + 3O_2 \longrightarrow As_2O_3 + 3H_2O$
		Arsine	AsH_3	Gas	
	Phosphorus	Phosphorus Oxychloride	$POCl_3$	Liquid	$4POCl_3 + 3O_2 \longrightarrow 2P_2O_5 + 6Cl_2$
		Phosphorus Pentoxide	P_2O_5	Solid	$2PH_3 + 4O_2 \longrightarrow P_2O_5 + 3H_2O$
		Phosphine	PH_3	Gas	
P	Boron	Boron Tribromide	BBr_3	Liquid	$4BBr_3 + 3O_2 \longrightarrow 2B_2O_3 + 6Br_2$
		Boron Trioxide	B_2O_3	Solid	$B_2H_6 + 3O_2 \longrightarrow B_2O_3 + 3H_2O$
		Diborane	B_2H_6	Gas	
		Boron Trichloride	BCl_3	Gas	$BCl_3 + 3H_2 \longrightarrow 2B + 6HCl$
		Boron Nitride	BN	Solid	
	Gold	Gold	Au	Solid (Evap.)	
	Iron		Fe		
	Copper		Cu		
	Lithium		Li	Undesirable impurities from contamination	
	Zinc		Zn		
	Manganese		Mn		
	Nickel		Ni		
	Sodium		Na		

*Note: Only selected diffusion reactions are listed

Figure 11.16 Deposition source table.

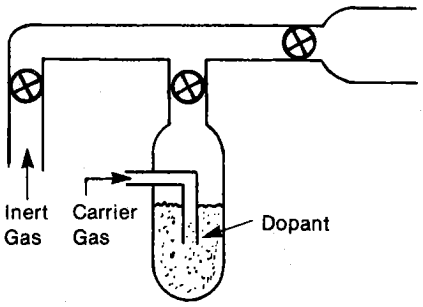


Figure 11.17 Liquid dopant source.

(Fig. 11.18) at the tube inlet. Baffles break up the incoming gas stream into a laminar form.

Also connected to the manifold is a reaction gas that is required to create the elemental dopant form in the tube. For BBr_3, the reaction gas is oxygen, which creates the boron oxide (B_2O_3), as shown in Fig. 11.19. At the wafer surface, a boron trioxide layer deposits on the silicon, and the boron diffuses from the oxide into the surface. Liquid sources offer the advantages of low to moderate cost and consistent doping. The disadvantages are uniformity problems (especially for larger-diameter wafers), safety considerations, and the potential of contamination associated with the opening of the flasks for recharging. Several vendors supply liquid dopants in pluggable sealed ampules that minimize contamination and safety problems.

Gas sources. Many wafer manufacturers prefer gas dopant sources. These are hydrated forms of the dopant atom such as arsine (AsH_3 and diborane (B_2H_6). The gases are mixed in different dilutions in pressurized containers and connected directly to the gas manifold

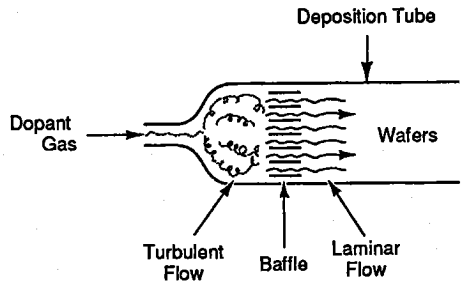


Figure 11.18 Use of baffle to create laminar flow.

$$4BBr_3 + 3O_2 \rightarrow 2B_2O_3 + 6Br_3$$

Figure 11.19 Reaction of liquid source BBR3 and oxygen in a deposition tube.

(Fig. 11.20). Gas sources offer the advantage of precise control through pressure regulators and are favored for deposition on larger-diameter wafers. The processes are in general cleaner, since the pressurized sources last longer than liquid sources. On the downside, unwanted chemical reactions in the manifold can create silica dust that can contaminate the tube and wafers.[2]

Solid sources. The original deposition sources were solid. An oxide powder of the desired dopant was placed on a quartz holder, called a *spoon*, and placed in a source tube furnace attached to the main deposition tube (Fig. 11.21*a*). In the source furnace, the oxide gives off source vapors that are carried into the deposition tube where the diffusion takes place. This setup is called a *remote solid source*. Remote solid sources are economical but nonuniform. They are used primarily in the doping of discrete devices where less precision is required.

A more popular solid source is the planar source wafer (Fig. 11.21*b*). These are wafer-size "slugs" that contain the desired dopant. Boron slugs are a compound of boron and a nitride (BN). Slugs are also available for arsenic and phosphorus diffusions.

The slugs are stacked on the deposition boat, with one slug between every two device wafers. This arrangement is called a *solid neighbor source*. In the tube, the dopant diffuses out of the slug, crosses the short distance to the wafer, and diffuses into the surface. This system provides good uniformity for larger-diameter wafers, since the slug is the same size as the wafer. Slugs are safer to use, having no toxic vapors at room temperature. Drawbacks to the slug process include breakage of the slugs, lower productivity (the slugs take up space in

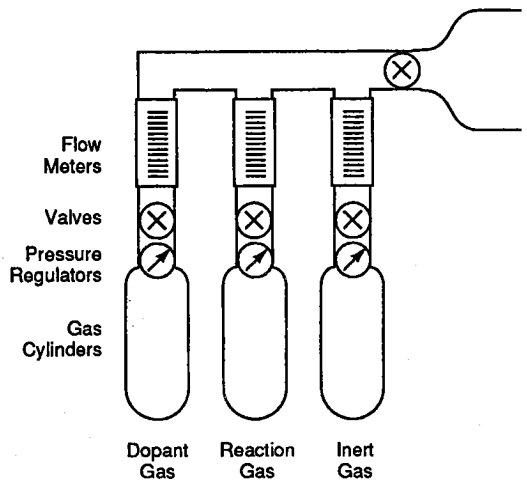


Figure 11.20 Gas source manifold.

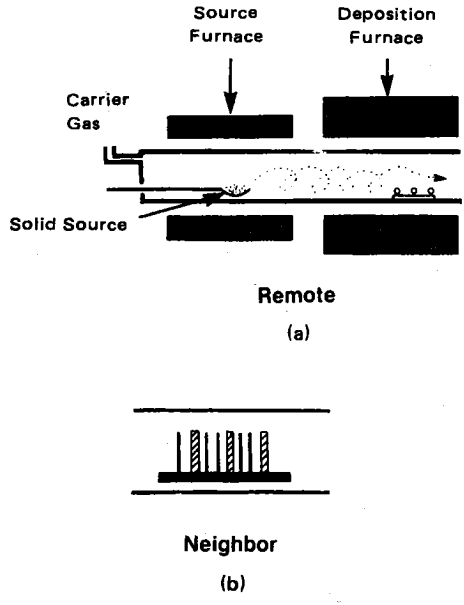


Figure 11.21 Solid source. (*a*) Remote furnace and (*b*) neighbor source.

the tube), and the necessity of cleaning the slugs. Some slug compounds require bake steps to maintain the dopant activity.

The third solid dopant source is a conformal layer spun directly on the wafer surface. The sources are powdered oxides (same as remote sources) mixed in solvents. They are spun onto the wafer surface using photoresist-type spinners and baked to evaporate the solvents. Left on the surface is a layer of doped oxide that conforms to the wafer surface. The wafers are placed on a boat and positioned in the deposition tube, where the heat drives the dopant out of the oxide and into the wafer.

Spin-on dopants have the potential of high uniformity and offer high productivity, and, like the slugs, are safe to handle. Problems with the system are the distribution of the dopant in the oxide layer; thickness variations, especially over surface steps; and the costs of the additional spin and bake steps.

Drive-in Oxidation

The second major part of the diffusion process is the drive-in-oxidation step. It is also variously known as *drive-in, diffusion, reoxidation,* and *reox.* The purpose of this step is twofold: redistribution of the dopant in the wafer and growth of a new oxide on the exposed silicon surface.

1. The first step is redistribution of the dopant deeper into the wafer. During the deposition, a high-concentration layer and a shallow

layer of dopant are diffused into the surface. In the drive-in, there is no dopant source. The heat alone drives the dopant atoms deeper and wider into the wafer just as material from a spray can will continue to spread into the room after the nozzle is released. During this step, the total amount of atoms (Q) from the deposition step remains constant. The surface concentration is reduced, and the distribution of atoms takes a new shape. The distribution after the drive-in is described by mathematicians as a gaussian distribution (Fig. 11.22). The junction depth increases. Generally, the drive-in-oxidation step takes place at a higher temperature than the deposition step.

2. The second purpose of drive-in oxidation is the oxidation of the exposed silicon surfaces. The atmosphere in the tube is oxygen or water vapor, which performs the oxidation simultaneously as the dopants are being driven deeper into the wafer.

The setup, process steps, and equipment for the drive-in-oxidation step are the same as an oxidation process. After the completion of the drive-in, the wafers are again evaluated. The test wafers (from the deposition step) are again measured for surface concentration

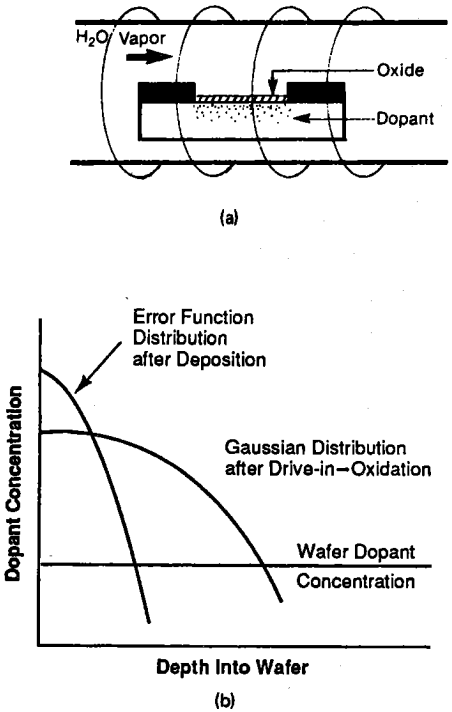


Figure 11.22 Drive-in oxidation. (*a*) Cross section of wafer and (*b*) dopant concentration in wafer.

with a four-point probe, and the device wafers are checked for cleanliness. The test wafers are measured for junction depth and perhaps for mobile ionic contamination. After some diffusion steps, a test structure in the engineering die may be electrically probed for junction parameters.

Oxidation effects

The oxidation of the silicon surface affects the final distribution of the dopants.[3] The effects are related to the relocation of the top-level dopants after the oxidation. Recall that the silicon in the silicon dioxide film is consumed from the wafer surface. The question to ask is, "What happened to the dopants that were in the top level?" The answer to that question depends on the conductivity type of the dopant.

If the dopant is an N-type, an effect called *pile-up* (Fig. 11.23*a*) occurs. As the oxide-silicon interface advances into the surface, the N-type dopant atoms segregate into the silicon rather than the oxide. The effect is to increase the number of these dopants in the new top layer of the silicon. In other words, the N-type dopants pile up in the wafer surface, and the surface concentration of the dopant is increased. Pile-up changes the electrical performance of the devices.

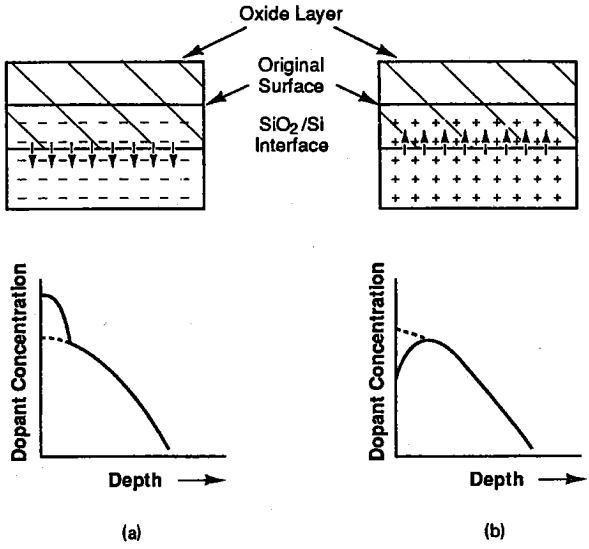


Figure 11.23 Pile-up and depletion of dopants during oxidation. (*a*) Pile-up of N-type dopants and (*b*) depletion of P-type dopants.

If the dopant is the P-type boron, an opposite effect occurs. The boron atoms are more soluble in the oxide and are drawn up into it (Fig. 11.23*b*). The effect on the wafer surface is a lowering of the concentration of boron atoms, which lowers the surface concentration and also affects the electrical performance of the devices. A summary of the deposition and drive-in-oxidation steps is provided in Fig. 11.24 The effect of the deposition and drive-in steps on the diffusion electrical parameters is shown in Fig. 11.25.

	Deposition	**Drive-In**
Goals	Introduction of Dopant	1. Redistribution of Dopant 2. Reoxidation
Variables		1. Surface Composition 2. Junction Depth 3. Time 4. Diffusivity 5. Temperature 6. Quantity of Atoms
Source Conditions	Continuous Source	No Source
Temperature Range	900 - 1100°C	1050 - 1200°C
Oxidation	No	Yes

Figure 11.24 Summary of deposition and drive-in steps.

Change	**Parameter**	**Deposition**	**Drive-in/ Oxidation**
Increase Temperature (same time)	Dopant concentration (Dc)	Increases	Same
	Surface concentration (Co)	Increases	N type-Lower P type- Higher
	Junction depth (Xj)	Deeper	Deeper
	Sheet Resistance	(Rs1) Lower	(Rs2) N type-Lower P type- Higher
Increase Time (same temp.)	Dopant concentration (Dc)	Increases	Same
	Surface concentration (Co)	Same	N type-Lower P type- Higher
	Junction depth (Xj)	Deeper	Deeper
	Sheet Resistance	(Rs1) Lower	(Rs2) N type-Lower P type- Higher

Figure 11.25 Doping process parameter changes.

Introduction to Ion Implantation

The advent of high-density circuits requires smaller feature sizes and closer spacing of the circuit components. Thermal diffusion places a limit on the production of advanced circuits. Five challenges are lateral diffusion, ultra thin junctions, poor doping control, surface contamination interference, and dislocation generation. Lateral diffusion occurs during deposition and drive-in but also continues every time the wafer is heated into a range where diffusion movement can take place (Fig. 11.26). The circuit designer must leave enough room between adjacent regions to prevent the laterally diffused regions from touching and shorting. The accumulative effect for a dense circuit can be a largely increased die area. Another problem with high temperatures is crystal damage. Every time a wafer is heated and cooled, crystal damage from dislocations occurs. A high concentration of these dislocations can cause device failure from leakage currents. One goal of an advanced process sequence is a reduced *thermal budget* to reduce these two problems.

The advent of MOS transistors created two new doping requirements: low dopant concentration control and ultra-thin junctions. Gate regions with dopant concentrations below 10^{15} atoms/cm^2 are required for efficient MOS transistors. However, this level is difficult to achieve consistently with a diffusion process. Scaling MOS transistors smaller to achieve higher packing densities also requires thin junction depths in the source/drain areas.[4] At the 0.18-μm design rule level, junctions will be in the 40 nm range; at the sub-0.10 μm range, they will be in the 20-nm range.[5]

A fourth problem is imposed by the physics/mathematics of a diffused region. As illustrated in Figs. 11.14 and 11.22, the majority of the dopant atoms are located near the wafer surface. This puts the majority of the current traveling near the surface where the dopant atoms are located. Unfortunately, this is the same location as contaminants (in and on the wafer surface) that interfere or degrade the current flow. A requirement of advanced devices is special wells in the surface (Fig. 11.27) with specific dopant gradients that cannot be achieved with diffusion technology. These wells allow high-performance transistors (see Chapter 16).

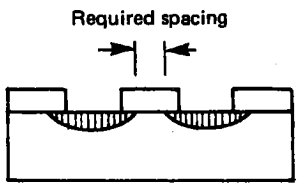


Figure 11.26 Side diffusion.



Figure 11.27 Ion implantation analogy.

Ion implantation overcomes these limits of diffusion and also adds additional benefits. During the ion implant process, there is no side diffusion, the process takes place at close to room temperature, the dopant atoms are placed below the wafer surface, and a wide range of doping concentrations are possible. With ion implantation, there is greater control of the location and number of dopants put in the wafer. Also, photoresist and thin metal layers can be used as doping barriers along with the usual silicon dioxide layers. Given the benefits, it is not surprising that the majority of doping steps for advanced circuits are done by ion implantation. Diffusion still finds use in the doping of less-critical layers and will continue to be used for lower-integration-level circuits.

Concept of Ion Implantation

Diffusion is a chemical process. Ion implantation is a physical process; that is, the act of implanting does not rely on chemical interaction between the dopant and the wafer material. An analogy that demonstrates the concept of ion implantation is a cannon firing balls into a wall (Fig. 11.27). Given enough momentum from the powder in the cannon, the balls will penetrate the wall and come to rest below the surface of the wall. The same events take place in an ion implantation machine. Instead of cannonballs, dopant atoms are ionized and isolated, accelerated (gain momentum), formed into a beam, and swept across the wafer. The dopant atoms physically bombard the wafer, enter the surface, and come to rest below the surface (Fig. 11.28*b*).

Ion Implantation System

An ion implanter is a collection of very sophisticated subsystems (Fig. 11.28*a*), each performing a specific action on the ions. Ion implanters come in a variety of designs used for advanced research and/or high-volume production. All of the machines have the same major subsystems as described below.

Implant sources

The same dopant elements used in diffusion processes are used in ion implant processes. In diffusion processes, the dopants originate in liq-

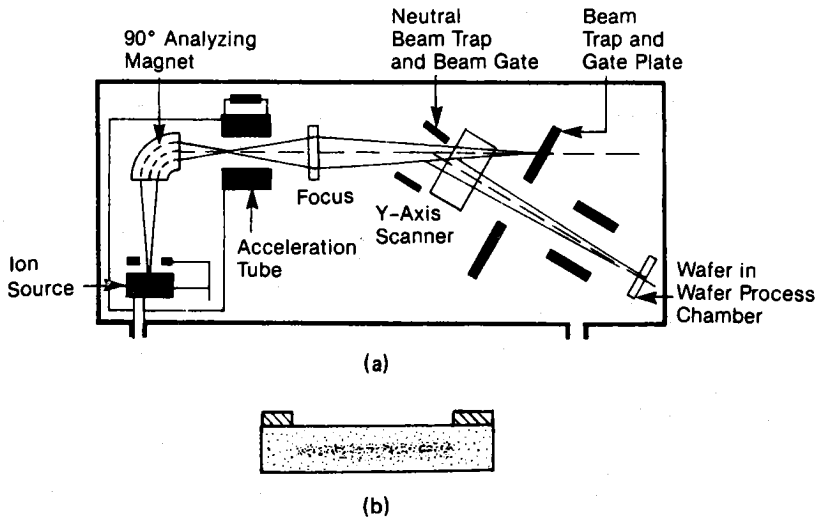


Figure 11.28 Ion implantation. (*a*) Block diagram of ion implanter and (*b*) distribution of implanted atoms in wafers.

uid, gas, or solid sources. For ion implantation, only gas and solid sources are used.

Gases are favored for ion implantation because of their ease of use and higher control. The gases most used are AsH_3, PH_3, and BF_3. The gas cylinders are connected to the ion source subsystem through mass flow meters, which offer more control of the gas flow than normal flow meters. A newer system, called the *safe delivery system* (SDS) has the dopant, such as arsine, at low pressure and absorbed onto a solid surface.[6]

Ionization chamber

The name "ion implant" implies that ions are a part of the process. Recall that ions are atoms or molecules with a negative or positive charge. The ions implanted are ionized atoms of the dopants. The ionization occurs in a chamber that is fed by the source vapors. The chamber is maintained at a low pressure (vacuum) of about 10^{-3} torr. Inside the chamber is a filament that is heated to the point where electrons are created from the filament surface. The negatively charged electrons are attracted to an oppositely charged anode in the chamber. During the travel from the filament to the anode, the electrons collide with the dopant source molecules and create a host of positively charged ions from the elements in the molecule. The results of the ionization of the source BF_3 are shown in Fig. 11.29.

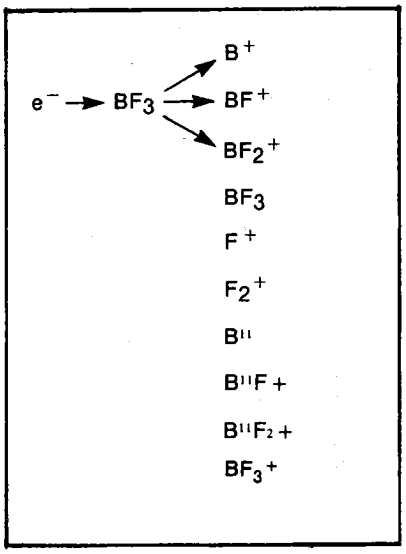


Figure 11.29 Ion species of BF_3.

Another ionization method uses a cold-cathode technique to generate the electrons, A high-voltage electric field is created between a cathode and anode, which creates the electrons in a self-sustaining process.

Mass analyzing/ion selection

At the top of the list in Fig. 11.29 is a lone boron ion. This is the atom that is desired in the wafer surface. The other species resulting from the ionization of the boron trifluoride are not wanted in the wafer. The boron ion must be *selected* from the group of positive ions. This process is called *analyzing*, *mass analyzing*, *selection*, or *ion separation*.

Selection is accomplished in a mass analyzer. This subsystem was first developed during the Manhattan Project for the atomic bomb. The analyzer (Fig. 11.30) creates a magnetic field. The species leave the ionization subsystem with voltages of 15 to 40 keV (thousand electron volts). In other words, they are traveling at a relatively high speed.

In the field, each of the positively charged species is bent into an arc with a specific radius. The radius of the arc is dictated by the mass of the individual species, its speed, and the strength of the magnetic field. At the exit end of the analyzer is a slit that will allow only one species to exit. The magnetic field is adjusted to match the path of the boron ion to the exit slit position. Thus, only the boron ion leaves the analyzing subsystem.

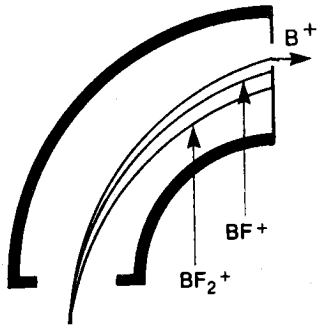


Figure 11.30 Analyzing magnet.

In some systems, the analysis also takes place after acceleration (Fig. 11.30). After acceleration, analysis is necessary if the implant species is a molecule that might separate in the acceleration process and/or to ensure an uncontaminated beam.

Acceleration tube

On leaving the analyzing section, the boron ion moves into an acceleration tube. The purpose is to accelerate the ion to a high enough velocity, thus gaining sufficient momentum to penetrate the wafer surface. *Momentum* is defined as the product of the mass of the atom multiplied by its velocity. This section is kept at a high vacuum (low pressure) to minimize contaminants from entering into the beam. Turbo vacuum pumps (Chapter 13) are typically used for this purpose.

The required velocity is achieved by taking advantage of the fact that negative and positive charges attract each other. The tube is a linear design with annular anodes along its axis. Each of the anodes has a negative charge. The charge amount increases down the tube.

As the positively charged ion enters the tube, it immediately starts to accelerate down the tube. The voltage value is selected based on the mass of the ion and the momentum required at the wafer end of the implanter. The higher the voltage, the higher the momentum, and the faster and deeper the dopant ion can be implanted. Voltages range from 5 to 10 keV for low-energy implanters to 0.2 to 2.5 MeV (million electron volts)[6] for high-energy implanters.

Ion implanters are classified into the categories of medium- and high-current machines, high-energy devices, and oxygen ion implanters. The stream of positive ions exiting the tube is actually an electric current. The beam current level translates into the number of ions implanted per minute. The higher the current, the more atoms are implanted. The amount of atoms implanted is called the *dose*. Medium-current machines produce currents in the 0.5 to 1.7 mA (milliampere)

range at energies from 30 to 200 keV (thousand electron volts). High-current machines generate beam currents of about 10 mA at energies up to 200 keV.[7] High-energy implanters are finding use in several CMOS doping applications, including retrograde wells, channel stops, and deep buried layers (see Chapter 16).

Wafer charging

High-current implants create an unacceptable degree of electrical charging (*wafer charging*) of the wafer surface. The high-current beam carries excess positive charges that charge the wafer surface. The positive charge draws neutralizing electrons from the surface, the bulk, and from the beam. High charge levels can degrade or destroy surface dielectric layers. Wafer charging is a particular problem on thinner MOS gate dielectrics.[8] Methods used to neutralize or reduce the charge are *flood guns* specifically designed to provide electrons, a plasma bridge method of providing low-energy electrons,[9] and control of the electron path with magnetic fields.[10]

The relationships of current and energy for production-level implanters is shown in Fig. 11.31. High-energy machines accelerate ions in the 10 keV to 3.0 MeV range, with beam currents up to 1.0 mA. Oxygen implant machines are high-current implanters used to implant oxygen in SOI applications (see Chapter 16).

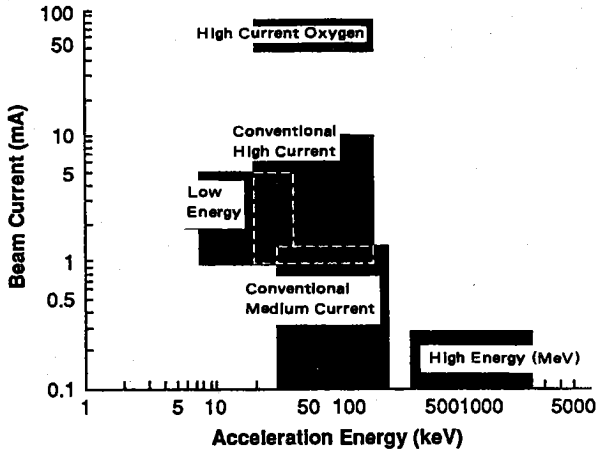


Figure 11.31 Ion implanters. The conventional descriptions of ion implanters were based on applications. However, some of today's more advanced implanters cannot be classified easily; several systems are capable of a broader process window than the conventional descriptions indicate. *(From Ref. 4.)*

A successful ion implant relies on the implantation of only the desired dopant atom. Single-dopant implantation requires that the system be maintained at a low pressure, greater than 10^6 torr. The danger is that any residue molecules in the system (such as air) can become accelerated and end up in the wafer surface. Either oil-diffusion or cryogenic high-vacuum pumps are employed to reduce the pressure. The operation of these systems is described in Chapter 12.

Beam focus

On exiting the acceleration tube, the beam separates due to repulsion of like charges. The separation (or defocus) causes uneven ion density and nonuniform layers in the wafer. For successful implantation, the beam must be focused. Electrostatic or magnetic lenses are used to focus the ions into a small diameter beam or a band of parallel beams.[11]

Neutral beam trap

Despite the vacuum removal of the majority of the air in the system, there are still some residual gas molecules in the vicinity of the ion beam. Collisions between the ions and the residual gas atoms result in a neutralization of the dopant ion.

$$P^+ + N_2 \rightarrow P^0 \text{ (neutral)} + N_2^+$$

In the wafer, these "neutrals" cause nonuniform doping and, because they cannot be "counted" by evaluation equipment, they result in incorrect counting of the amount of dopants in the wafer. Suppression of the neutral beam is accomplished by bending the ion beam (Fig. 11.32) with electrostatic plates, leaving the neutral beam to travel straight ahead away from the wafers.

Beam scanning

Ion beams have a smaller diameter than the wafer (~1 cm). Covering the entire wafer with a uniform doping requires scanning the beam

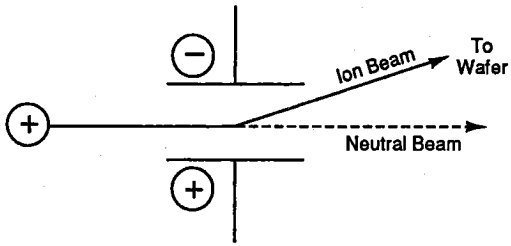


Figure 11.32 Deflection of ion beam from neutral beam.

across the wafer. Three methods are used: beam scanning, mechanical scanning, and shuttering, alone or in combination.

A beam-scanning system (Fig. 11.33) has the beam pass between a number of electrostatic plates. Negative and positive charges can be controllably changed on the plates to attract and repel the ionized beam. By manipulating the charges in two dimensions, the beam can be swept across the entire wafer surface in a raster scan pattern.

Beam sweeping is used primarily in medium-current machines for single-wafer implanting. The procedure is fast and uniform. A drawback is the requirement that the beam be moved completely off the wafer to make the turns. For a large-diameter wafer, this procedure can lengthen the implant time by 30 percent or more. Another problem occurs with high-current machines where the high density of ions causes discharges (called *space charge forces*) that destroy the electrostatic plates. Wide beams are swept across the wafer. In some systems, the wafer is rotated 90° after each sweep of the beam to ensure uniformity.[12]

Mechanical scanning approaches the scanning problem by holding the beam in one position and moving the wafer(s) in front of it. Mechanical scanning is used primarily on high-current machines. One advantage is that there is no wasted time to turn the beam, and the beam speed is constant. If the wafer is at an angle to the beam, nonuniform implant depths can result. But, in some cases, the wafer will be oriented at an angle to the beam. Beam shuttering employs either an electronic field or a mechanical shutter to turn the beam on when it

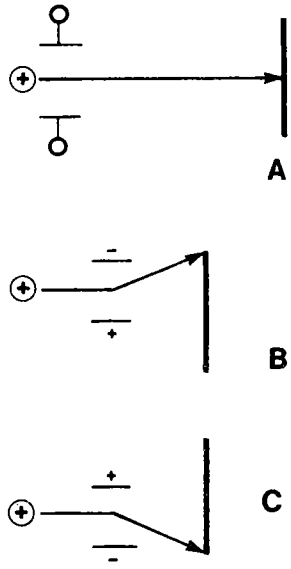


Figure 11.33 Electrostatic beam scanning.

is on the wafer and off when it is not. Most systems use a combination of beam sweeping and mechanical movement.

End station and target chamber

The actual implantation takes place in the target chamber of the end station. It includes the scanning system and the wafer load and unload mechanisms. There are several tough requirements for end stations. Wafers must be loaded into the chamber; a vacuum pulled; wafers individually placed on the holder; the implant completed; and the wafers dismounted, loaded into a cassette, and removed from the chamber.

Both batch and single-wafer designs are used to present the wafer surface to the beam (Fig. 11.34). A batch process is more efficient, but there are higher maintenance and alignment chores. For batch processing, the wafers are placed on a disk that rotates them in front of the beam as it is being scanned. The multiple motions increase the dose uniformity. Single-wafer designs require more time to process a group of wafers due to the additional time to load, pull the vacuum, implant, and unload.

Cryogenic pumps are favored for evacuating the end station. Contamination produced during the process consist of nitrogen outgassing from the wafers and hydrogen from the photoresist masking layer. Cryogenic pumps (see Chapter 13) are a capture type and hold the potentially dangerous hydrogen frozen in the pump.

The mechanical motions can take longer than the implant itself. Improvements include load locks to allow loading without breaking the

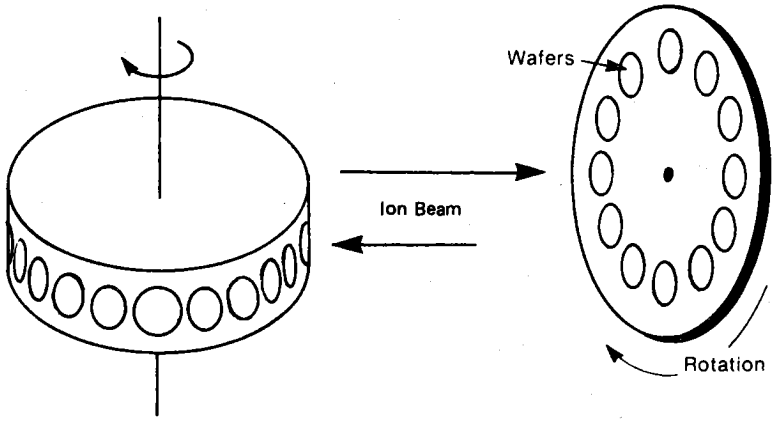


Figure 11.34 Mechanical scanning.

chamber vacuum. A big challenge is to keep particulate generation low during all of the mechanical movements in the chamber.[13] Installation of antistatic devices in the chamber is critical. Electrostatic handlers (no mechanical clamps) are an option.[14]

Wafer breakage can cause contamination from wafer chips and dust, which in turn requires time-consuming cleaning. Contamination on the wafers causes shadowing that blocks the ion beam. Production speed must be maintained with a system that quickly evacuates the chamber for implanting and returns it to room pressure for exit and reloading. The target chamber may house a detector (called a *Faraday cup*) to "count" the number of ions impacting the surface. These detectors can automate the process by allowing beam contact with the wafer until the correct dose is achieved.

High-current implantation can cause the wafer to heat up, and these machines often have cooling mechanisms on the wafer holders. These machines may also have an *electron flood gun* (Fig. 11.35) designed to minimize a buildup of charge on the wafer surface that can electrostatically attract contamination.

Ion implant masks

A major advantage of ion implantation is the variety of masks that are effective blocks to the ion beam. In diffusion doping, the only effective mask is silicon dioxide. Most films employed in the semiconductor process can be used to block the beam, including photoresist, silicon dioxide, silicon nitride, aluminum, and other thin metal films.

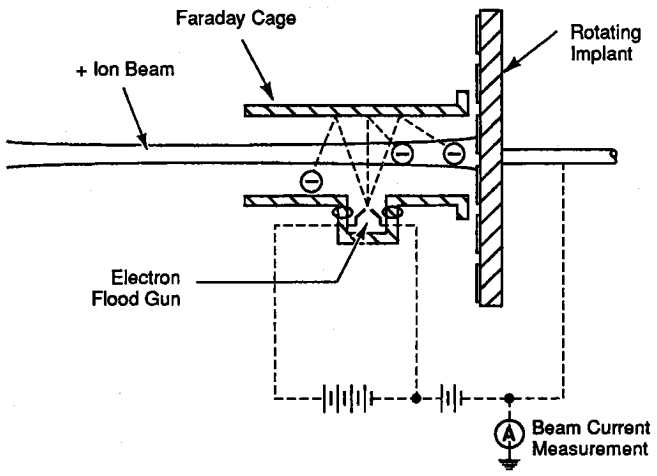


Figure 11.35 Electron flood gun.

Figure 11.36 compares the thicknesses required to block a 200-keV implant for various dopants.

The use of resist films as a beam block rather than an etched opening in an oxide layer offers the same dimension control advantage as the lift-off process; the etch step and its variability are eliminated. Use of resist layers also is more productive. Options to the use of silicon dioxide increase overall yield by minimizing the number of heating steps the wafers undergo.

Dopant Concentration in Implanted Regions

The distribution of the ions in the wafer surface is different from the distribution after a diffusion process. The number and location of the dopant atoms in a diffusion process is determined by the diffusion laws and time and temperature. In an ion implant process, the number of atoms (dose) implanted is determined by the beam current density (ions per square centimeter) and the implant time.

The location of the ions in the wafer is a function of the incoming energy of the ions, the orientation of the wafer, and the stopping mechanism of the ion. The first two factors are physical ones. The heavier the incoming ion and/or the higher its energy, the deeper into the wafer it will move. The wafer orientation influences the stopping position because different crystal planes have different atom densities and the incoming ions are stopped by the wafer atoms.

Within the wafer, the ions are slowed and stopped by two mechanisms. The positive ions are slowed by electronic interactions with the negatively charged electrons in the crystal. The other interaction is the physical collision with the nucleus of the wafer atoms. All of the stopping factors are variable; the ions have a distribution of energies, the crystal is not perfect, and the electronic interactions and collisions vary. The net result is that the ions come to rest over an area in the

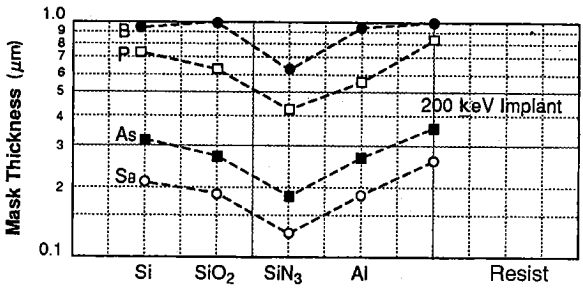


Figure 11.36 Barrier thickness required to block ion beam.

wafer (Fig. 11.37). They are centered about a depth called the *projected range* and fall off in density on each side of it. Additional implants create similar distribution patterns. Projected ranges for different dopants are shown in Fig. 11.38. The mathematical shape of

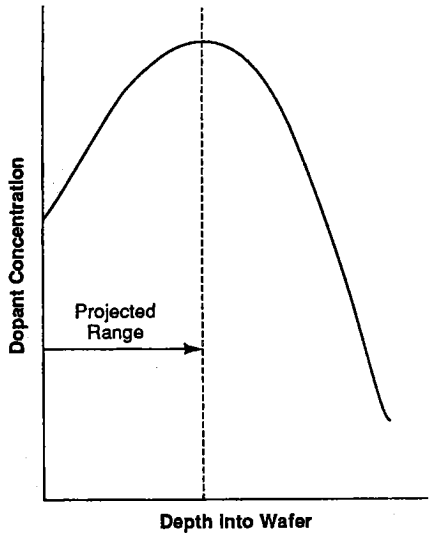


Figure 11.37 Dopant concentration profile after an ion implant.

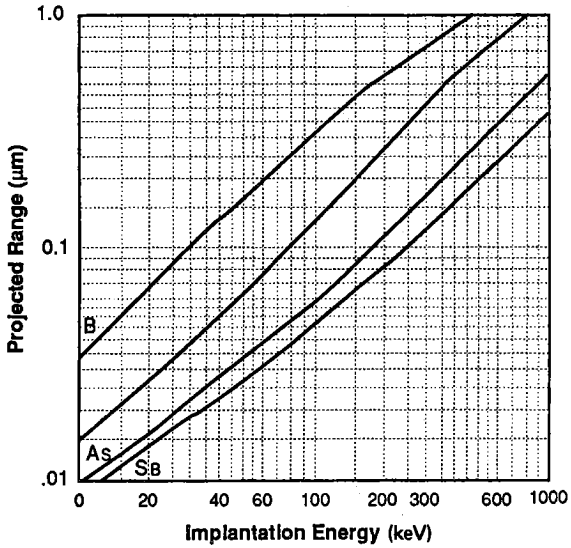


Figure 11.38 Projected range of various dopants in silicon. *(After Blanchard, Trapp, and Shepard.)*

the ion distribution is a gaussian curve. A junction between the implanted ions and the bulk doping in the wafer takes place where the ion concentration equals the bulk doping concentration.

Crystal damage

During the process of implantation, the wafer crystal structure is damaged by the colliding ions. There are three types of damage: lattice damage, damage cluster, and vacancy-interstitial.[15]

Lattice damage occurs when the ions collide with host atoms and displace them from their lattice site. A damage cluster occurs when displaced atoms in turn displace other substrate atoms, creating a cluster of displaced atoms. The most common implant-produced defect is a vacancy-insterstitial. This defect comes about when an incoming ion knocks a substrate atom from a lattice site and the displaced atom comes to rest in a nonlattice position (Fig. 11.39).

Light atoms, such as boron, produce a small number of displaced atoms. The heavier atoms, phosphorus and arsenic, generate a large number of displaced atoms. With prolonged bombardment, the regions of dense disorder may change to an amorphous (noncrystal) structure. In addition to the structural damage to the wafer from ion implantation, there is an electrical effect. The damaged regions do not have the required electrical characteristics, because the implanted atoms do not occupy lattice sites.

Annealing and dopant activation

Restoration of the crystal damage and electrical activation of the dopants can be achieved by a thermal heating step. The temperature of the anneal is below the diffusion temperature of the dopant to prevent lateral diffusion. A typical anneal in a tube furnace will take place between 600 and 1000°C in a hydrogen atmosphere.

RTP techniques are also used for post-implant annealing. RTP offers fast surface heating that restores the damage without the substrate temperature rising to the diffusion level. Additionally, the anneal can take place in seconds, whereas a tube process takes 15 to 30 min. If the wafer has a heavy subsurface amorphous layer, the

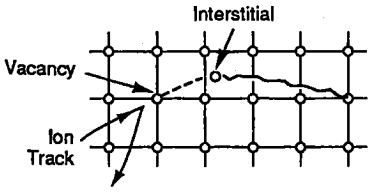


Figure 11.39 Vacancy-interstitial damage mechanism.

damage may be restored with a second implant of a light atom, such as oxygen or neon.

Channeling

The crystalline structure of the wafer presents a problem during the ion implantation process. The problem comes about when the major axis of the crystal wafer is presented to the ion beam. Ions can travel down the channels, reaching a depth as much as 10 times the calculated depth. An ion concentration profile of a channeled cross section (Fig. 11.40) shows an significant amount of additional dopants. Channeling is minimized by several techniques: a blocking amorphous surface layer, misorientation of the wafer, and creating a damage layer in the wafer surface.

The usual blocking amorphous layer is simply a thin layer of grown silicon dioxide (Fig. 11.41). The layer randomizes the direction of the ion beam so that the ions enter the wafer at different angles and not directly down the crystal channels. Misorientation of the wafer 3 to 7° off the major plane also has the effect of preventing the ions from entering the channels (Fig. 11.42). Predamaging the wafer surface with a heavy silicon or germanium implant creates a randomizing layer in

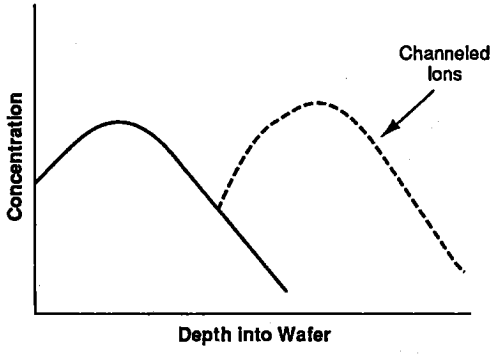


Figure 11.40 Effect of channeled ions on total dose.

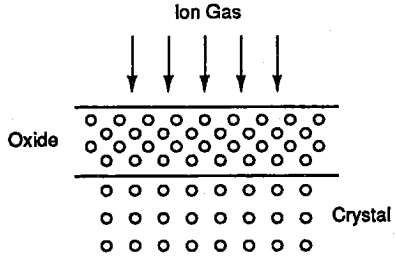


Figure 11.41 Implant through an amorphous oxide layer.

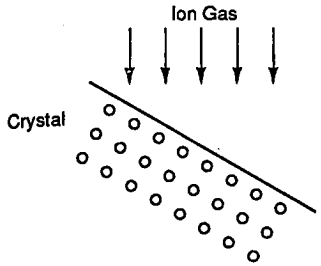


Figure 11.42 Misorient the beam direction to all crystal axes.

the wafer surface (Fig. 11.43). The method increases the use of the expensive ion-implant machine. Channeling is more of a problem with low-energy implants and heavy ions.[16]

Evaluation of Implanted Layers

The evaluation of implanted wafers is essentially the same as for diffused layers. Four-point probes are used to determine the sheet resistance of the layer. Spreading resistance techniques and capacitance-voltage techniques determine the concentration profile, dose, and junction depth. Junction depths can also be determined by bevel and decoration methods. These procedures are explained in Chapter 14.

For implanted layers, a special structure, called the *van der Pauw structure*, is sometimes used in place of a four-point probe (Fig. 11.44). The structure allows a determination of sheet resistance without the contact resistance problems of a four-point probe. Variation in the implanted wafers can come from many sources: the beam uniformity,

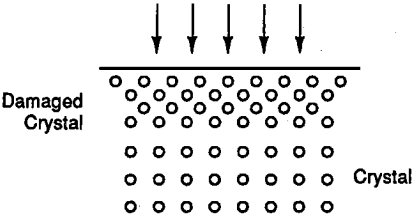


Figure 11.43 Predamage on the crystal surface.

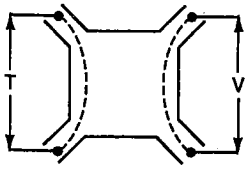


Figure 11.44 Van der Pauw text pattern.

variations in voltage, scanning variations, and problems in the mechanical systems. These potential problems have the possibility of causing a wider sheet resistance surface variation than a diffusion process. To detect and control the sheet resistance across the wafer surface, mapping techniques are popular and, for critical implants, required. Wafer surface maps (Fig. 11.45) are drawn from four-point probe measurements that are computer corrected for proximity and edge effects.

A measurement technique unique to ion implantation is optical dosimetry. The technique requires spinning a glass disk with a layer of photoresist. Before being placed in the ion implanter, the resist film is scanned with a dosimeter that measures the absorption of the film. The information is stored in a computer. The wafer and film receive the same implant as the device wafers. The resist absorbs the ion dose and darkens. After the implant, the film is again scanned. The computer subtracts the before-implant value for each location and prints out a contour map of the surface. The spacing of the contour lines is an indication of the uniformity of the dopants in the surface (Fig. 11.46).

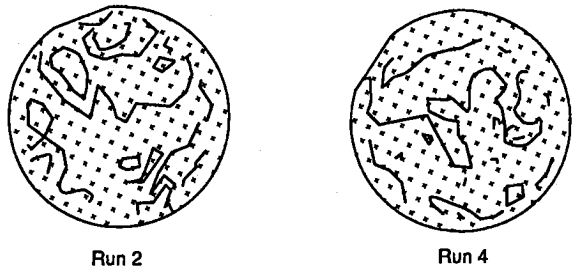


Figure 11.45 Four-point probe surface measurement pattern. *(Courtesy of Prometrics.)*

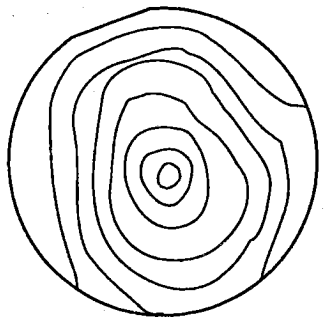


Figure 11.46 Contour map of ion dose.

Uses of Ion Implantation

An ion implantation can be substituted for any diffusion deposition process. The greater control and lack of side diffusion make it the preferred doping technique for dense and small-feature-size circuits. A predeposition application in CMOS devices is the creation of the deep P-type wells (see Chapter 16), called *retrograde wells,* using high-energy implants.

A particular challenge is ultra-shallow junctions. These are junctions in the sub-125-nm range. As devices are continually scaled to smaller dimensions, junction areas also become reduced. This in turn leads to lower energy ion implant processes to minimize surface damage and channeling. This leads to implanting pure boron instead of BF_3, which contains corrosive fluorine. All of these requirements are being met in a new generation of implanters that can deliver high-dose beams at acceptably low energies.

One of the most important uses of ion implantation is for MOS gate threshold adjustment (Fig. 11.47). A MOS transistor consists of three parts: a source, a drain, and a gate. During operation, a voltage is applied between the source and drain regions. However, no current can flow between the regions until the gate becomes conductive. The gate becomes conductive when a voltage applied to it causes a conductive channel to form in the surface and connects the source and drain. The amount of voltage required to first form the connecting channel is called the *threshold voltage* of the device. The threshold voltage is very sensitive to the dopant concentration in the wafer surface under the gate. Ion implantation is used to create the required dopant concentration in the gate region. Also, in MOS technology, ion implantation is used to alter the field dopant concentration. However, in this use, the purpose is to set a concentration level that prevents current flow between adjacent devices. In this application, the implanted layer is part of a device *isolation* scheme.

In bipolar technology, ion implantation is used to create all of the various transistor parts. The "custom" dopant profiles available from

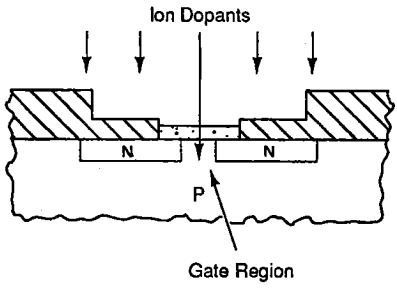


Figure 11.47 Ion doping of MOS gate region.

implantation can improve device performance. One particular application is arsenic buried layers. When the buried layer is diffused, the high concentration of arsenic atoms affects the quality of a subsequent epitaxial layer deposited on the surface. By using an arsenic ion implant, a high concentration of arsenic is possible, and an anneal process restores the damage, allowing a higher-quality epitaxial layer to be deposited.

Resistors for both MOS and bipolar circuits are good candidates for ion implantation. Diffused resistors vary in uniformity from 5 to 10 percent, whereas ion-implanted resistors have variations of less than 1 percent or better. Figure 11.48 is a table of typical applications of ion-implantation doping.

The Future of Doping

Ion implantation also has its drawbacks. The equipment is expensive and complex. Training and maintenance take longer and have more stringent requirements. The machines present new dangers in the form of high voltages and more toxic gas use. From the process perspective, the biggest worry is over the ability of the annealing processes to completely eliminate the implant-induced damage. However, despite the drawbacks, ion implantation is the preferred doping process for advanced circuits. In addition, many new structures are possible only with the unique advantages offered by ion implantation.

Ion implantation may run out of steam beyond the shallow junction range of 40 nm.[17] A new technique under development is *plasma ion immersion* (PII). In this technique the analyzing magnet is removed from the system. Dopants leave the source section and a plasma field increases their energy.[19] This technique places the wafer in a plasma field (similar to ion milling or sputtering) in the presence of dopant atoms. With proper charges on the dopant atoms and the wafer, the dopant atoms accelerate to the wafer surface and penetrate much like an ion implantation. The difference is the lower energy of the plasma

Typical Uses of Ion Implantation
• Gate threshold voltage adjustment
• Ultra-shallow junctions
• Buried layers (retrograde wells)
• Predeposition layer
• Insulation layer of SOI

Figure 11.48 Ion implantation uses.

field yielding less wafer charging and giving more control of the shallow junctions.[18]

Ion implantation, in one form or another, is the doping technology that will carry the industry into the nanometer era. The benefits include:

- Precise dose control from 10^{10} to 10^{16} atoms/cm^2
- Uniform topping of large areas
- Dopant profile control through energy selection
- Relative ease of implanting all dopant elements
- Minimal side diffusion
- Implanting of nondopant atoms
- Can dope through surface layers
- Choice of dopant barriers for selective doping
- Tailored doping profiles in deep (retrograde) wells

Review Questions

1. Explain the difference between a deposition and a drive-inoxidation.
2. Name the three types of sources used in a doping processes.
3. List two evaluation measurements made after a doping process.
4. Is the junction depth deeper after deposition or the drive-in oxidation?
5. What type of source is BCl_3?
6. Name the subsystems in an ion implantation.
7. In what subsystem is the dopant ion selected?
8. What is the difference between a diffusion and an ion-implanted dopant profile? Sketch an example of each.
9. Name two materials used as an ion implant doping barrier.
10. Why is an anneal required after ion implantation?

References

1. P. B. Griffin and J. D. Plummer, "Advanced Diffusion Models for VLSI," *Solid State Technology*, May 1988, p. 171.
2. K. T. Robinson, "A Guide to Impurity Doping," *Micromanufacturing and Test*, April 1986, p. 52.
3. P. Guise and R. Blanchard, *Modern Semiconductor Fabrication,* Reston Books, Reston, VA, 1986, p. 46.

4. S. Felch, et al., "A Comparison of Three Techniques for Profiling Ultrashallow p^+-n Junctions," *Solid State Technology*, PennWell Publishing Company, January 1993, p. 45.
5. N. Cheung, "Ion Implantation," *Semiconductor International*, Cahners Publishing, January 1994, p. 411.
6. P. Lessard, Vacuum Issues in the Semiconductor Industry, Ion Implantation, Vacuum Technology and Coating, June 2003, p. 35.
7. P. Burggraaf, "Ion Implanters: Major Trends," *Semiconductor International*, April 1986, p. 78.
8. R. Iscoff, "Are Ion Implanters the Newest Clean Machines?," *Semiconductor International*, Cahners Publishing, October 1994, p. 65.
9. N. Cheung, "Ion Implantation," *Semiconductor International*, Cahners Publishing, January 1993, p. 35.
10. J. England, et al., Charge Neutralization During High-Current Ion Implantation," *Solid State Technology*, PennWell Publishing, July 1994, p. 115.
11. Japan Report, *Semiconductor International*, Cahners Publishing, November 1994, p. 32.
12. Eaton Corp., Product Video, The NV8200P, 1993.
13. *Ibid.*
14. R. Iscoff, "Are Ion Implanters the Newest Clean Machines?" *Semiconductor International*, Cahners Publishing, October 1994, p. 65.
15. "Wafer Handler for Ion Implanters, Varian Semiconductor Equipment," *Solid State Technology*, PennWell Publishing, July 1994, p. 131.
16. J. Hayes and P. Van Zant, *Doping Today Seminar Manual*, *Semiconductor Services*, San Jose, CA, 1985.
17. D. Zrudsky, "Channeling control in ion implantation," *Solid State Technology*, July 1988, p. 73.
18. N. Cheung, "Ion implantation," *Semiconductor International*, Cahners Publishing, January 1993, p. 35.
19. A. Braun, Ion Implantation Goes Beyond Traditional Parameters, *Semiconductor International*, March 2002, p. 48.
20. P. Singer, "Plasma Doping: An Implant Alternative?," *Semiconductor International*, Cahners Publishing, May 1994, p. 34.

Chapter

12

Layer Deposition

Overview

Doped regions and N-P junctions are the electronic hearts of the active components in a circuit. However it takes various other layers of semiconductors, dielectrics, and conductors to complete the components and facilitate the integration of the components into the circuit. These layers are added to the wafer surface by a number of techniques. The principle ones are chemical vapor deposition (CVD), physical vapor deposition (PCD), electroplating, spin-on, and evaporation. This chapter describes the most commonly used CVD techniques and dielectric and semiconductor materials deposited on the wafer surface. PVD, electroplating, spin-on, and evaporation processes are described in Chapter 13.

Objectives

Upon completion of this chapter, you should be able to:

1. Name the parts of a CVD reactor.
2. Describe the principle of chemical vapor deposition.
3. List the conductor, semiconductor, and insulator materials deposited by CVD techniques.
4. Know the difference between atmospheric CVD, LPCVD, hot-wall, and cold-wall systems.
5. Explain the difference between epitaxial and polysilicon layers.

Introduction

Advances in photomasking technology have allowed the fabrication of ever-smaller dimensioned ULSI circuits. But as the circuits have

shrunk, they also have grown in the vertical direction through increased numbers of deposited layers. In the 1960s, bipolar devices had two layers deposited by chemical vapor deposition (CVD), an epitaxial layer, and a top-side passivation layer of silicon dioxide (Fig. 12.1), while early MOS devices had just a passivation layer (Fig. 12.2). By the 1990s, advanced devices featured four levels of metal interconnects, requiring numerous deposited layers (see Fig. 13.2, Chapter 13). The added layers take a variety of roles in the device/circuit structures. They include the following:

- Deposited doped silicon layers called *epitaxial layers* (see related section in this chapter)
- Intermetal dielectrics (IMDs)
- Vertical (trench) capacitors
- Intermetal conducting plugs
- Metal conducting layers
- Final passivation layers

There are two primary techniques for layer deposition: chemical vapor deposition (CVD) and physical vapor deposition (PVD). The metallization deposition techniques of evaporation and sputtering are explained in Chapter 13. The uses of the particular films, while indicated in this chapter, are detailed in Chapters 16 and 17. Chemical vapor deposition (CVD), the subject of this chapter, is practiced in a number of atmospheric and low-pressure techniques.

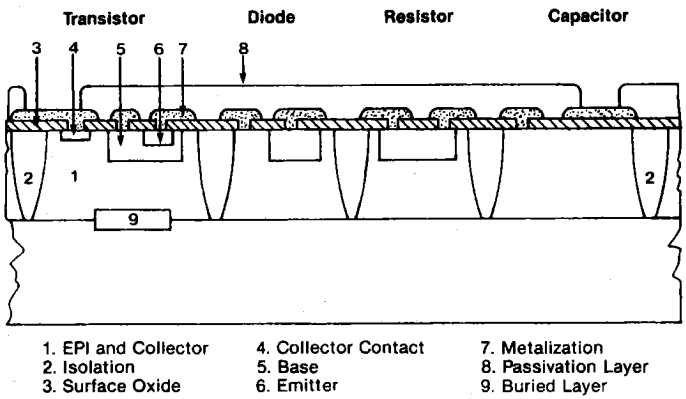


Figure 12.1 Cross section of bipolar circuit showing epitaxial layer and isolation.

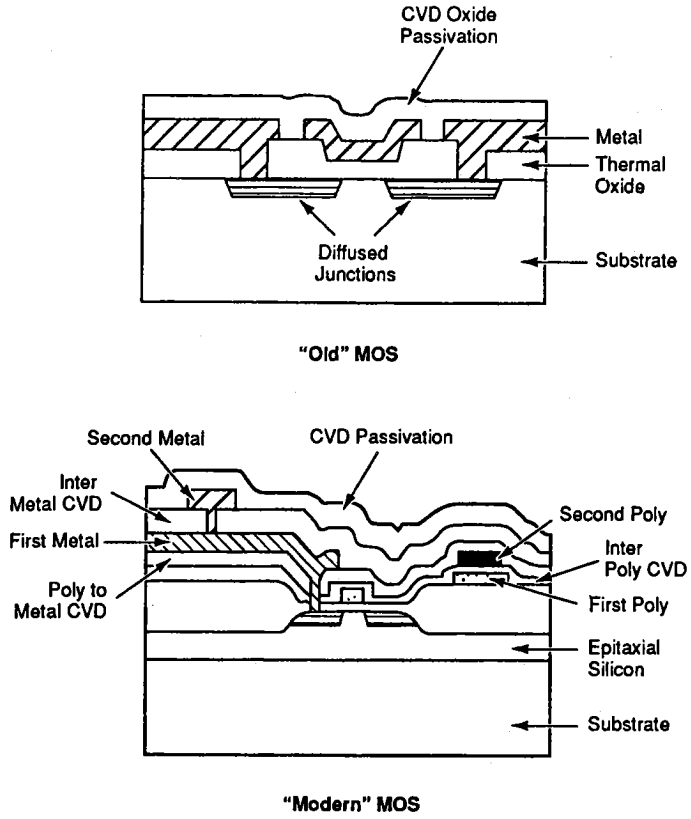


Figure 12.2 Evolution of MOS layers.

Film parameters

Device layers must meet general and specific parameters. The specific parameters are noted in the sections on individual layer materials. General criteria that all films must meet for semiconductor use include:

- Thickness/uniformity
- Surface flatness/roughness
- Composition/grain size
- Stress free
- Purity
- Integrity

Uniform thickness is required of films to meet both electrical and mechanical specifications. Deposited films must be continuous and free of pinholes to prevent the passage of contamination and to prevent shorting of sandwiched layers. This is of great importance for thin films. Epitaxial films have shrunk from 5-micron levels to submicron thicknesses. Recall that the thickness of a layer is one of the factors contributing to its resistance. Also, thinner layers tend to have more pinholes and less mechanical strength. Of particular concern is the maintenance of thickness over steps (Fig. 12.3). Excessive thinning at a step can cause electrical shorts and/or unwanted induced charges in the device. The problem becomes very acute in deep and narrow holes and trenches, called *high-aspect-ratio patterns*. The ratio is calculated by dividing the depth by the width (Fig. 12.3). One problem is a thinning of the deposited film at the lip of the trench. Another is thinning in the bottom of the trench. Filling high-aspect trenches is a major issue in the execution of multimetal structures.

Surface flatness is as important as the thickness. In Chapter 10, the effect of steps and surface roughness on image formation was detailed. Deposited films must be flat and as smooth as the material and deposition method will allow to minimize steps, cracking, and subsurface reflections.

Deposited films must be of the desired uniform composition. Some of the reactions are complex, and it is possible that films will be deposited with something other than the intended composition. *Stoichiometry* is the methodology by which the quantities of reactants and products in chemical reactions are determined. In addition to chemical composition, grain size is important. During deposition, the film materials tend to collect or grow into grains. Varying grain size within films of the same composition and thickness will yield different electrical and mechanical properties. This is because electrical current flow is affected as it passes through the grain interfaces. Mechanical properties also change with the size of grain interface area.

Stress-free films are another requirement. A film deposited with excess stress will relieve itself by forming cracks. Cracked films cause

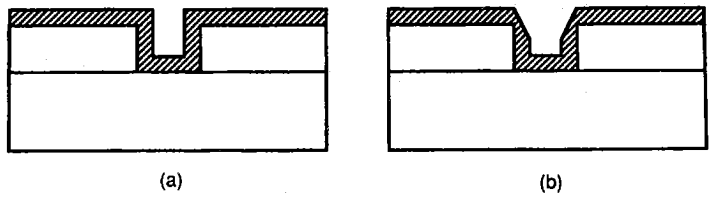


Figure 12.3 Thinning of deposited layer (*b*) at step.

surface roughness and can allow contamination to pass through to the wafer. In the extreme, they cause electrical shorts.

Purity (i.e., a lack of unwanted chemical elements or molecules in the film) is required for the film to carry out its intended function. For example, oxygen contamination of an epitaxial film will change its electrical properties. Purity also includes the exclusion of mobile ionic contaminants and particulates.

An electrical parameter of importance to deposited films is capacitance (see Chapter 2). Semiconductor metal conduction systems need high conductivity and, therefore, low-resistance and low-capacitance materials. These are referred to as *low-k dielectrics*. Dielectric layers used as insulators between conducting layers need high capacitances or *high-k dielectrics*.

Chemical Vapor Deposition Basics

Not surprisingly, the growth in the number and kinds of deposited films has resulted in a number of deposition techniques. Where the process engineer of the 1960s had a choice of only atmospheric chemical vapor deposition (CVD), today's engineer has many more options (Fig. 12.4). These techniques are described in the following sections.

Thus far, the terms *deposition* and *CVD* have been used without explanation. In semiconductor processing, deposition refers to any process in which a material is physically deposited on the wafer surface. Grown films are those, such as silicon dioxide, that formed from the material in the wafer surface. The majority of films are deposited by a CVD technique. In concept, the process is simple (Fig. 12.5). Chemicals (C) containing the atoms or molecules required in the final film are mixed and reacted in a deposition chamber to form a vapor (V). The atoms or molecules deposit (D) on the wafer surface and build up to form a film. Figure 12.5 illustrates the reaction of silicon tetrachloride ($SiCl_4$) with hydrogen to form a deposited layer of silicon on the wafer. Generally, CVD reactions require energy to take place.

Atmospheric Pressure (AP)	Low Pressure (LP) and Ultra High Vacuum (UHV)
Cold wall	Hot wall
• Horizontal	Plasma enhanced
• Vertical	Vertical isothermal
• Barrel	Molecular beam epitaxy (MBE)
• Vapor phase epitaxy	
Metalorganic CVD	

Figure 12.4 Overview of deposition systems.

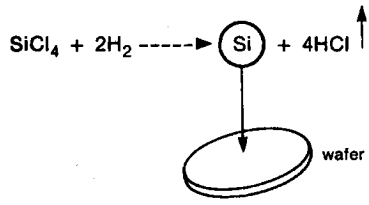


Figure 12.5 Chemical vapor deposition of silicon from silicon tetrachloride.

The chemical reactions that take place fall into the four categories of pyrolysis, reduction, oxidation, and nitridation (Fig. 12.6). Pyrolysis is the process of chemical reaction driven by heat alone. Reduction causes a chemical reaction by reacting a molecule with hydrogen. Oxidation is the chemical reaction of an atom or molecule with oxygen. Nitridation is the chemical process of forming silicon nitride by exposing a silicon wafer to nitrogen at a high temperature.

Deposited films grow in several distinct stages (Fig. 12.7). The first stage, *nucleation*, is very important and critically dependent on substrate quality. Nucleation occurs as the first few atoms or molecules deposit on the surface. These first atoms or molecules form islands that grow into larger islands. In the third stage, the islands spread, finally coalescing into a continuous film. This is the transition stage of the film growth with a typical thickness of several hundred angstroms. The transition region film has chemical and physical properties much different from those of the final, thicker "bulk" film.[1]

Pyrolysis	$SiH_4 \rightarrow Si + 2H_2$
Reduction	$SiCl_4 + 2H_2 \rightarrow SI + 4HCL$
Oxidation	$SiH_4 + O_2 \rightarrow SiO_2 + 2H_2$
Nitridation	$3SiH_2Cl_2 + 4NH_3 \rightarrow Si_3N_4 + pH + 6H_2$

Figure 12.6 Examples of CVD reactions.

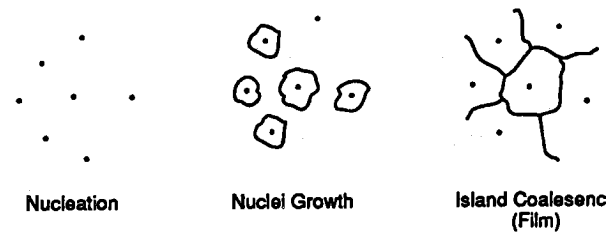


Figure 12.7 CVD film growth steps.

After the transition film is formed, the bulk growth begins. Processes are design to produce three different structures: amorphous, polycrystalline, and single crystal (Fig. 12.8). These terms have been defined previously. A poorly defined or controlled process can result in a film with the wrong structure. For example, attempting to grow a single-crystal epitaxial film on a wafer with islands of unremoved oxide will result in regions of polysilicon in the bulk film.

Basic CVD system components

CVD systems come in a wide variety of designs and options. Understanding the many variations is helped by an examination of the basic subsystems common to most CVD systems (Fig. 12.9). In most respects, a CVD system has the same basic parts as a tube furnace (described in Chapter 7): source cabinet, reaction chamber, energy source,

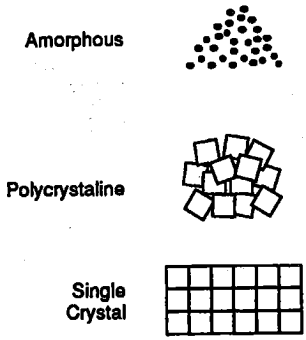


Figure 12.8 Types of film structure.

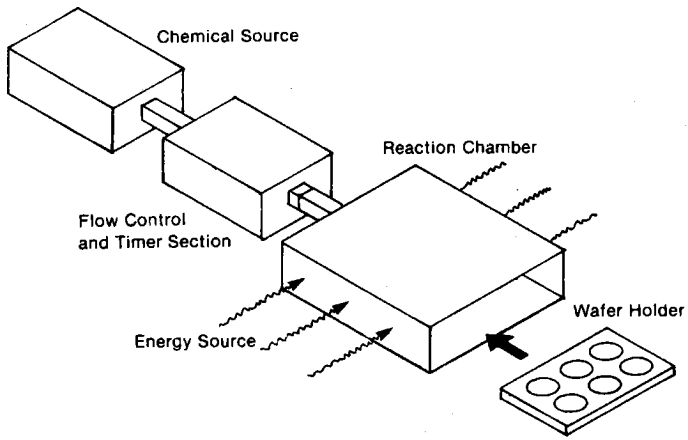


Figure 12.9 Basic CVD subsystems.

wafer holder (boat), and loading and unloading mechanisms. In some cases, the CVD system *is* a tube furnace identical to those used for oxidation and diffusion. The source chemicals are housed in a source section. Vapors are generated from pressurized gas cylinders or liquid source bubblers. Gas flow control is maintained by pressure regulators, mass flow meters, and timers.[2]

The actual deposition takes place on the wafers in a reaction chamber. Energy sources can be heat conduction, convection, induction RF, radiant, plasma, or ultraviolet. Energy sources are explained in the sections on particular systems. Temperatures range from room temperature to 1250°, depending on the reaction, film thickness required, and the growth parameters.

The fourth basic part of the system is the wafer holder. Different chamber configurations and heat sources dictate the style and material of the holders. Most production-level systems for ULSI circuits are automated from wafer load to unload. A full production system will have an associated cleaning section or station and a loading area.

CVD Process Steps

A CVD process follows the same steps as an oxidation or diffusion process. For review, the steps are preclean (and etch, if required), deposition, and evaluation. Cleaning processes are those already described to remove particulates and mobile ionic contaminants. Chemical vapor deposition, like an oxidation, takes place in cycles. First, the wafers are loaded into the chamber, usually with an inert atmosphere. Next, the wafers are brought to temperature. Chemical vapors are introduced for as long as required to deposit the film. Finally, the chemical source vapors are flushed out and the wafers removed. Evaluation of the films are for thickness, step coverage, purity, cleanliness, and composition. Evaluation techniques are explained in Chapter 14.

CVD System Types

CVD systems (Fig. 12.10) are divided into two primary types: atmospheric-pressure (AP) and low-pressure (LP). There are a number of atmospheric pressure CVD systems (APCVD). However, most films for advanced circuits are deposited in systems where the pressure has been lowered. These are called low-pressure CVD or LPCVD.

Another differentiation is cold wall versus hot wall. Cold-wall systems directly heat the wafer holder or wafers, with induction or radiant heating. The walls of the chamber remain cold (or cooler). Hot-wall systems heat the wafers, the wafer holder, and the chamber walls. The advantage of cold-wall CVD is that the reaction occurs only

Atmospheric systems	Low-pressure systems
Hot wall	Hot Wall
Cold wall	Plasma enhanced
• Epitaxial	• Hot-wall tube
• Barrel	• Cold-wall planar—batch
• Pancake	• Cold-wall planar—single
• Single-wafer	
	High-density plasma

Figure 12.10 CVD reactor designs.

at the heated wafer holder. In a hot-wall system, the reaction occurs throughout the chamber, leaving reaction products on the inside chamber walls. The reaction products build up, necessitating rigorous and frequent cleaning to avoid contaminating the wafers.

CVD systems are operated with two principal energy sources: thermal and plasma. Thermal sources are tube furnaces, hot plates, and RF induction. Plasma enhanced chemical vapor deposition (PECVD) in combination with lower pressure offers the unique advantage of lowered temperatures and good film composition and coverage.

A specialty CVD used to deposit compound films, such as GaAs, is vapor phase epitaxy (VPE). A newer technique used to deposit metals is a metalorganic (MOCVD) source in a VPE system. The last deposition method described is the non-CVD molecular beam epitaxy (MBE) used for low-temperature deposition of thin films in a very controlled process.

Atmospheric-Pressure CVD Systems

As the name implies, atmospheric CVD system reactions and deposition take place at atmospheric pressure. There are a number of system designs that fall under this category (Fig. 12.10).

Horizontal-tube induction-heated APCVD

The first widespread use of CVD was for the deposition of silicon epitaxial films for bipolar devices. The basic system design is still used

(Fig. 12.11). It is essentially a horizontal tube furnace, but with some significant differences. First, the tube has a square cross section. The major difference, however, is in the heating method and wafer holder.

Wafers are arranged on a flat graphite slab and positioned in the tube. Surrounding the tube are copper coils that are connected to an RF generator. The RF waves traveling in the coils pass through the quartz tube and the flowing gas in the tube without heating them. This is the cold-wall aspect of the system. When the radiant waves reach the graphite wafer holder, they "couple" with the molecules of the holder, causing the graphite to heat up. This heating method is called *induction.*

The heat of the holder is passed to the wafers by conduction. The film deposition takes place at the wafer surface (and at the holder surface). One problem with this type of system is downstream depletion of the reactants in the laminar gas flow. Laminar flow is required to minimize turbulence. But if the wafers are laid flat in the chamber, the layer of gas closest to the wafers becomes depleted. This results in successively thinner films along the wafer holder. A wafer holder tilted in the tube corrects the problem (Fig. 12.12).

Barrel radiant-induction-heated APCVD

Larger-diameter wafers laid horizontally on a holder in a horizontal system have a low packing density. And larger wafer holders strain the uniformity capabilities of the system.

The development of the barrel radiant-heated system (Fig. 12.13) solved these problems. The reaction chamber of the system is a cylindrical stainless steel barrel with high-intensity quartz heaters placed about the inside surface. The wafers are placed on a graphite holder that rotates in the center of the barrel. The rotation of the wa-

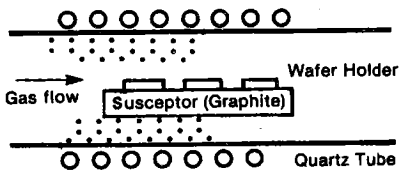


Figure 12.11 Cold-wall induction APCVD with horizontal susceptor.

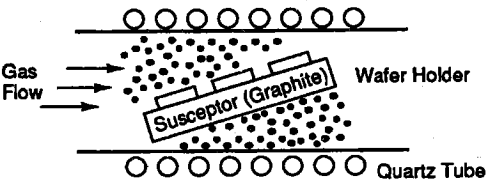


Figure 12.12 Cold-wall induction APCVD with tilted susceptor.

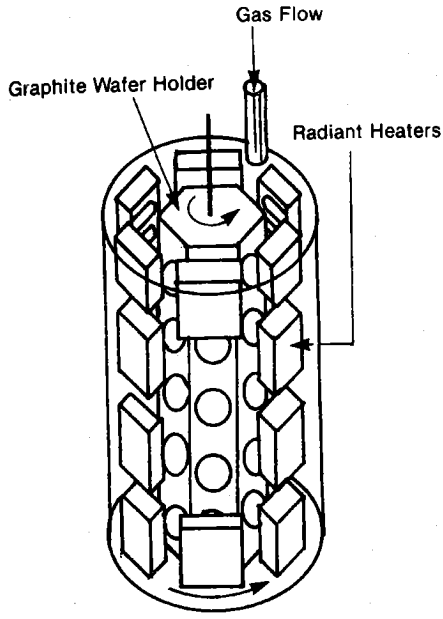


Figure 12.13 Cylindrical or barrel system.

fers produces a more uniform film thickness as compared to horizontal systems.

Radiant heat from the lamps heats the wafer surface, where the deposition takes place. While some heating of the chamber walls occurs, the system is close to a cold-wall deposition. Direct radiant heating produces a very controlled and even film growth. In an induction-heated system, the wafers are heated from the bottom, and, as the film grows, there is some small but measurable drop in temperature at the film surface. In the barrel system, the wafer surface is always facing the lamps and receives a more uniform temperature and film growth rate.

In 1987, Applied Material introduced a jumbo barrel system for large-diameter wafers featuring an induction heating system.[3] A principal advantage of the barrel reactor is an increased productivity based on the increased number of wafers per cycle. This system configuration is popular for deposition of epitaxial silicon in the 900 to 1250°C range.

Pancake induction-heated APCVD

The pancake or vertical-flow APCVD system has been a favorite for small fabrication lines and R&D labs (Fig. 12.14). The wafers are held on a rotating holder of graphite and heated by induction-conduction

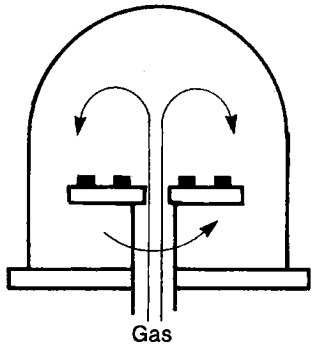


Figure 12.14 Rotating pancake APCVD.

from an RF coil below the holder. The reaction gases are fed through a tube exiting above the wafers. Vertical gas flow offers the advantage of a continuous supply of fresh reactants to the wafers, thus minimizing downstream depletion. The combination of the rotation and vertical flow of the gases produces good film uniformity. Productivity in smaller systems is restricted, as in the horizontal tube system, by the number of wafers that the pancake can accommodate.

A production-level variation of the pancake design is produced by Gemini Research. The reactor features radiant-resistance heating and a large-capacity holder with robot autoloading.[4]

Continuous conduction-heated APCVD

Two horizontal conduction-heated APCVD systems feature mixing the gases outside the chamber and showering them onto the wafers. In one design, a heated hot-plate wafer holder moves back and forth (Fig. 12.15) under a series of nozzles that dispense a vapor of the desired material. Another version of the system (Fig. 12.16) has wafers moving on a belt under a plenum that dispenses the reacted gases.

Horizontal conduction-heated APCVD

One of the original CVD designs is the horizontal conduction-heated APCVD system (Fig. 12.17) used to deposit the silicon dioxide passiva-

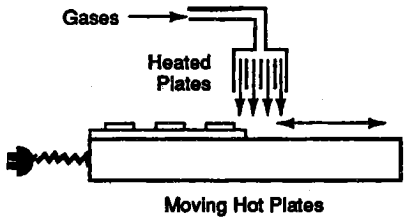


Figure 12.15 Moving hot-plate APCVD.

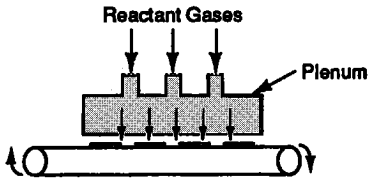


Figure 12.16 Continuous hot-plate APCVD.

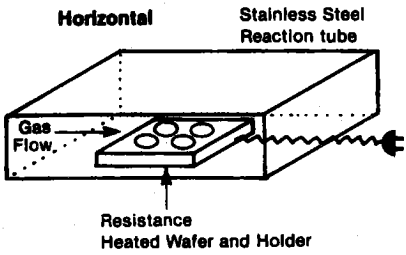


Figure 12.17 Hot-plate APCVD.

tion layers. Wafers are mounted on a removable hot plate and placed in a stainless steel chamber. The hot plate heats the wafers and chamber walls (hot wall system). Reaction gases are fed into the chamber.

Low-Pressure Chemical Vapor Deposition (LPCVD)

Uniformity and process control within atmospheric-pressure CVD systems rely on temperature control and the flow dynamics in the system. A factor influencing film uniformity and step coverage is the mean free path of the molecules in the reaction chamber. The mean free path is the average distance a molecule will travel before colliding with an object in the chamber, be it another molecule, the wafers, or wafer holder. Collisions change the direction of the particles. The longer the mean free path, the higher the uniformity of the film deposition. A major determiner of the length of the mean free path is the pressure in the system. Lowering the pressure in the chamber increases the mean free path and the film uniformity. Decreasing the pressure also allows a lowering of the deposition temperature.

These benefits became available to the industry in 1974 when Unicorp, under license to Motorola Inc., introduced an LPCVD system that operated at a few hundred millitorr.[5] The list of LPCVD system advantages includes:

- Lower chemical reaction temperature
- Good step coverage and uniformity
- Vertical loading of wafers for increased productivity and lower exposure to particles

- Less dependence on gas flow dynamics
- Less time for gas phase reaction particles to form
- Can be performed in standard tube furnaces

A vacuum pump must be added to the system to reduce the pressure in the chamber. A discussion of the types of pumps used with LPCVD systems appears in Chapter 13.

Horizontal conduction-convection-heated LPCVD

One production-level LPCVD system uses a horizontal tube furnace (Fig. 12.18), with three major exceptions. The tube is connected to a vacuum pump that pulls the system down to a pressure range of 0.25 to 2.0 torr.[6] A second change is a ramping of the temperature in the center zone to offset reaction depletion down the tube. The third change may be special injectors at the gas inlet end to improve gas mixing and deposition uniformity. In some systems, the injectors are positioned directly over the wafers. Disadvantages of this system design are particles formed on the inside wall surface (hot-wall reactions), uniformity along the tube axis, the use of cages around the wafers to minimize particle contamination, and the higher downtime required for frequent cleaning.

These systems are most often used for polysilicon, silicon dioxide, and silicon nitride films with typical thickness uniformities of ±5 percent. The primary deposition variables are temperature, pressure, gas flow, gas partial pressure, and wafer spacing. These variables are carefully balanced for each deposition process. The deposition rates are somewhat lower (100 to 500 Å/min) than AP systems, but productivity is enhanced by the vertical wafer-loading densities that can approach 200 wafers per deposition.

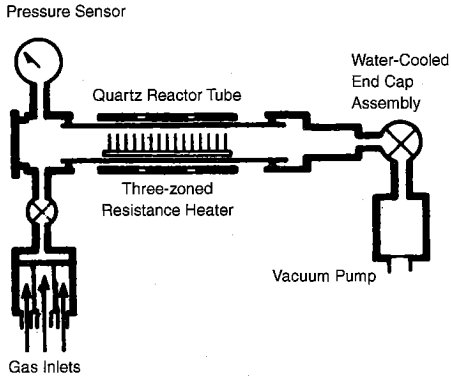


Figure 12.18 Horizontal hot-wall LPCVD system.

Ultrahigh vacuum CVD (UHV/CVD)

Low-temperature deposition is desirable to minimize crystal damage and lower the thermal budget that in turn minimizes the lateral diffusion of doped regions. One approach is the CVD of silicon and silicon/germanium (SiGe) in ultra-low vacuum conditions. Lowering the pressure allows keeping the deposition temperature low. UHV/CVD takes place in a tube furnace where the pressure is initially reduced to 1 to 5 $\times 10^{-9}$ millibar (mbar). Deposition pressure is in the 10^{-3} mbar range.[7]

Plasma-enhanced CVD (PECVD)

Replacement of silicon dioxide passivation layers with silicon nitride led to the development of plasma enhanced (PECV) techniques. A thermal silicon dioxide deposition temperature of approximately 660°C causes unacceptable alloying of aluminum interconnects into the silicon surface (see Chapter 13). A solution to this problem was a plasma enhancement of the deposition energy. The increased energy allows a temperature under the 450°C maximum level for deposition over aluminum layers. Plasma-enhanced systems are physically similar to plasma-etch systems. They feature a parallel plate chamber operated at a low pressure. A radio-frequency-induced glow discharge or other plasma source (see Chapter 9) is used to induce a plasma field in the deposition gas. The combination of low pressure and lower temperatures provides good film uniformity and throughput.

PECVD reactors have the capability of also using the plasma for etching and cleaning the wafer prior to the deposition step. This step is the same as the dry-etch processes described in Chapter 9. This *in situ* cleaning prepares the deposition surface, eliminating the problem of added contamination picked up during the loading step.

Horizontal vertical-flow PECVD. This system follows the design of a bottom-heated pancake vertical-flow CVD (Fig. 12.19). The plasma region is created in the top of the chamber by a radio frequency (RF) feed to an electrode or other plasma source.Wafer heating comes from radiant heaters mounted below the wafer holder, creating a cold-wall deposition system. With PECVD systems, there are several additional critical parameters to control in addition to those in a standard LPCVD reactor. They are the RF power density, the RF frequency, and the duty cycle. Film deposition speed is generally increased, but it must be controlled to prevent film stress and/or cracking.

Another design (developed by Novellus) has the wafers sitting on a series of resistance heated holders. The wafers are indexed around the chamber with the film built up in increments.

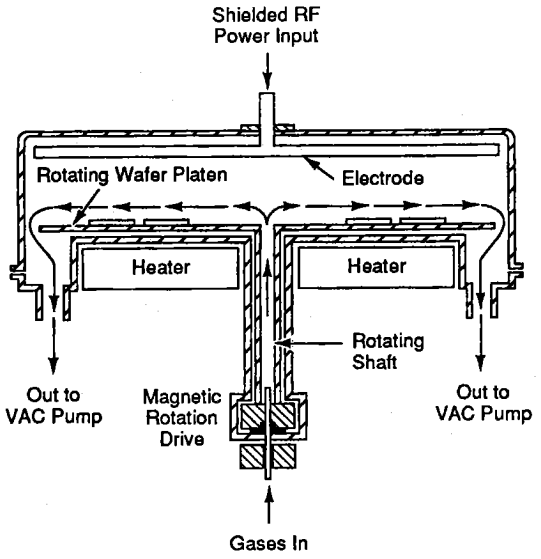


Figure 12.19 Vertical-flow pancake PECVD.

A single-wafer chamber PECVD system (Fig. 12.20), where the chamber is small and each successive wafer is exposed to identical conditions, addresses most of the control needs. Single-wafer systems are generally slower than batch processes. The productivity trade-off with the larger-chamber batch machines lies in fast mechanisms to feed wafers in and out of the chamber and how quickly the vacuum can be established and released. Load-lock systems enhance the productivity by moving the wafers into an antechamber, pumping it down to the required pressure, and moving the wafers into the deposition chamber as it becomes available. System designs include fixed and rotating wafer holders.

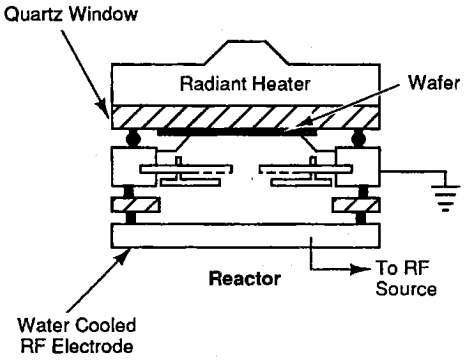


Figure 12.20 Single-chamber planar PECVD.

Barrel radiant-heated PECVD. This system is a standard barrel radiant heated system with low-pressure and plasma capabilities. It is favored for the deposition of tungsten silicide.

High-density plasma CVD (HDPCVD)

Intermetal dielectric (IDL) layers are essential for multimetal structures. They also present a challenge of filling high-aspect-ratio (greater than 3:1) holes. One approach is a deposition and an *in situ* etch sequence. The first deposition exhibits the usual thinning in the bottom. Etching away the shoulder and redeposition creates an uniform layer thickness and a more planar surface.

A system to accomplish this process is *high-density plasma* CVD.[8] A plasma field is created inside a CVD chamber that contains oxygen and silane for the deposition of silicon dioxide. Also included is argon that becomes energized by the plasma and is directed to the wafer surface. This is a sputtering (see "Dry Etching," Chapter 13) action that removes material from the surface and trench. HDPCVD has the potential of depositing a variety of materials for uses as IMD layers, etch stops, and final passivation layers.

CVD methods are summarized in Fig. 12.21.

Atomic Layer Deposition (ALD)

At some point on the roadmap, CVD and its many variations will run out of steam. A candidate for the next generation is *atomic layer deposition* (ALD). Based on a basic CVD approach, ALD uses a pulsing technique. The film is grown in stages with each reactant type separated by a purge gas. Unlike CVD, each of the reactant types chemically reacts with the surface. This makes each individual "deposition" self-limiting as the film grows, much like silicon dioxide growth. Since each film stage is growing at a monolayer rate, the control is very precise. Also, the slow rate facilitates high levels of conformity to the wafer surface and dense film composition.

Level	Temperature Range	Methods
High Temp.:	600-1250°C	R.F. Induction (Cold Wall) Radiant Heat (Cold Wall) Resistance Coils (Hot Wall)
Mid Temp.:	200-600°C	Hot Plates Plasma Enhanced LPCVD
Low Range:	22-200°C	Hot Plates P.E. CVD Photochemical

Figure 12.21 Summary table of CVD methods.

Candidate uses include very thin silicon dioxide gates, filling deep trenches with materials like aluminum oxide, and created barrier metal layers for copper metallization processes.[9]

Vapor Phase Epitaxy (VPE)

VPE differs from the CVD systems described in its ability to deposit compound materials, such as gallium arsenide. A VPE system[10] is a combination of a standard liquid source tube furnace and a two-zone diffusion furnace (Fig. 12.22). An example is the particular arrangement in Fig. 12.22 used to deposit epitaxial gallium arsenide. The creation of the GaAs layer on the wafer in the main chamber proceeds in two stages. $AsCl_3$ is bubbled into the first section of the tube where it reacts with a solid source of gallium that is sitting in a boat. The arsenic trichloride reacts with the hydrogen in the first section to form arsenic by the reaction

$$4AsCl_3 + 6H_2 \rightarrow 12HCl + As = 4$$

The arsenic deposits on the gallium, forming a crust. The hydrogen passing over the crust reacts in the first section to form three gases that pass into the wafer section.

$$\underset{\text{(solid)}}{GaAs} + \underset{\text{(gases)}}{HCl} \leftrightarrow \underset{\text{(gas)}}{GaCl} + \underset{\text{(gas)}}{1/2H_2} + \underset{\text{(gas)}}{1/4As_4}$$

This section is at a somewhat lower temperature, and the reaction proceeds in reverse, depositing GaAs on the wafers. The technique offers the advantages of clean films, since the gallium and arsenic trichloride are available in very pure forms and have higher production rates than the MBE technique. On the downside, the film structures produced are not the quality of MBE films.

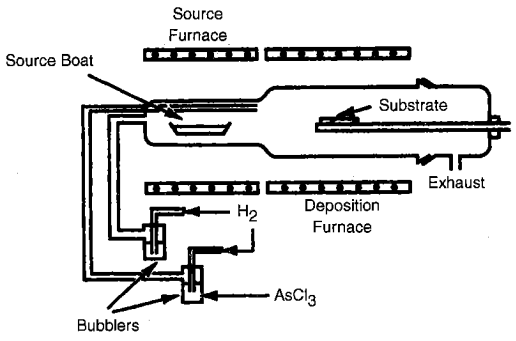


Figure 12.22 Diagram of gallium arsenide VPE deposition system.

Molecular Beam Epitaxy (MBE)

Deposition rate control, low deposition temperature, and controlled film stoichiometry are always goals in film-deposition systems. Molecular beam epitaxy (MBE) has emerged from the laboratory to claim production status as these issues have become more important. MBE is an evaporation rather than a CVD process. The system consists of a deposition chamber (Fig. 12.23) that is maintained at a low pressure to 10^{-10} torr. Within the chamber can be one or more cells (called *effusion cells*) that contain a very pure sample of the target material desired on the wafer. Shutters on the cells allow exposure of the wafer to the source material(s). An electron beam[10] is directed into the center of the target material, which it heats to the liquid state. In this state, atoms evaporate out of the material, exit the cell through an opening, and deposit on the wafers. If the material source is a gas, the technique is called *gas source MBE* or *GSMBE*. For most applications, the wafer in the chamber is heated to give additional energy to the arriving atoms. The additional energy fosters epitaxial growth and good film quality.

If the wafer surface is exposed, the depositing atoms will assume the orientation of the wafer and grow an epitaxial layer. MBE offers the intriguing option of *in situ* doping by the inclusion of dopant sources in the chamber. The usual silicon dopant sources are not usable in MBE systems. Solid gallium is used for P-type doping and antimony for N-type doping. Phosphorus deposition is virtually not possible with MBE.[12]

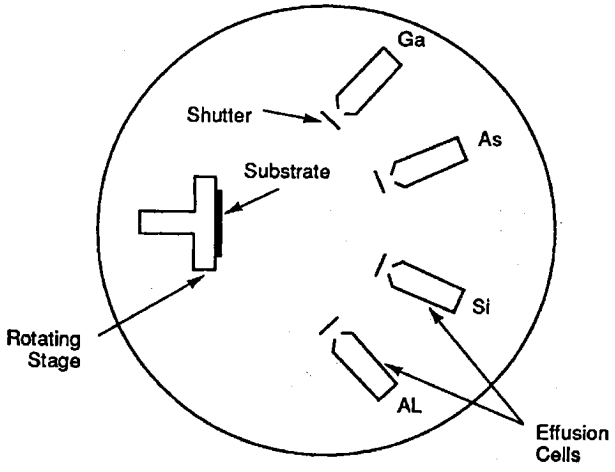


Figure 12.23 Diagram of MBE deposition system.

The primary advantage of MBE for silicon technology is the low temperature (400 to 800°C), which minimizes autodoping and outdiffusion. Perhaps the biggest advantage of MBE is the ability to form multiple layers on the wafer surface during one process step (one pump down). This option requires the mounting of several effusion cells in the chamber and shutter arrangements to direct the evaporant beams to the wafer in the right order and for the correct time.

An advantage and disadvantage of MBE is the low film growth rate of 60 to 600 Å/min.[12] On the plus side, the films produced are very controllable. Films can be grown (and mixed) in one monolayer increments. However, most semiconductor layers do not need this level of control and quality, making the low productivity and expense of the system an expensive luxury.

A bonus possibility with MBE is the incorporation of film growth and quality-analyzing instruments in the chamber. With these instruments, the process can become very controlled and produce uniform films from wafer to wafer. MBE has found production use in the fabrication of special microwave devices and for compound semiconductors such as gallium arsenide.[14]

Metalorganic CVD (MOCVD)

MOCVD is one of the latest options for CVD of compound materials. Whereas VPE refers to a compound material deposition system, MOCVD refers to the sources used in VPE systems (Fig. 12.24). Two chemistries are used, halides and metalorganic. The reactions described for the VPE deposition of gallium arsenide above constitute a

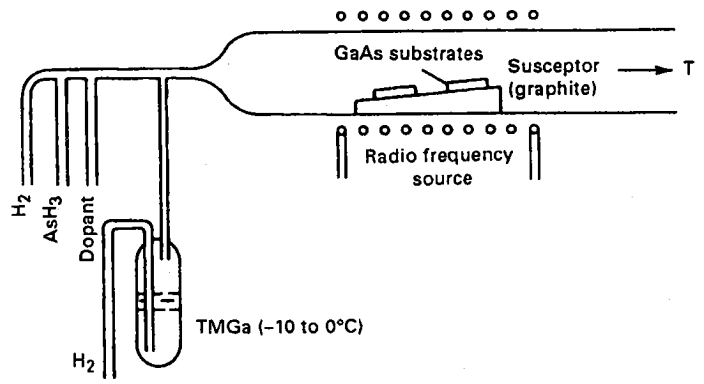


Figure 12.24 MOCVD system. *(Source:* VLSI Fabrication Principles, *S. K. Ghandhi, Wiley-Interscience, 1994.)*

halide process. A group III halide (gallium) is formed in the hot zone, and the III–IV compound is deposited in the cold zone. In the metalorganic process[15] for gallium arsenide, trimethylgallium is metered into the reaction chamber along with arsine to form gallium arsenide by the reaction

$$(CH_3)_3Ga + AsH_3 \rightarrow GaAs + 3CH_4$$

Where MBE processes are slow, MOCVD processes can meet volume production requirements and accommodate larger substrates.[16] MOCVD also has the capability of producing multiple layers with very abrupt changes in composition. This characteristic is critical to compound semiconductor structures. Also MOCVD, unlike MBE, can deposit phosphorus as in InGaAsP devices. Common devices made by MOCVD processes are photocathodes, high-poser LEDs, long-wavelength lasers, visible laser, and orange LEDs (see Chapter 16).

MOCVD is a broad term referring to the metalorganic chemical vapor deposit of semiconductor films. When metalorganic sources are used in a vapor phase epitaxial system to grow epitaxial layers, the method is called MOVPE.[17]

Deposited Films

The types of films deposited by CVD techniques are divided into their electrical classifications of semiconductors, dielectrics, and conductors. In the following sections, the deposition of the various films is examined. For each film, the principal use(s) in semiconductor devices is presented along with the particular film properties for the described use. The uses are treated in general. For a more detailed explanation of film roles in particular devices, see Chapter 16. Deposition methods for conductive metal films are discussed in Chapter 13.

Deposited Semiconductors

So far in this text, we have discussed the formation of wafers as the base of semiconductor devices and circuits. However, there are several drawbacks to using bulk wafers for high-quality devices and circuits. Crystal quality, doping ranges, and doping control all limit bulk wafer use. These factors placed a limit on the fabrication of high-performance bipolar transistors. A solution was found by the development of a deposited silicon layer, called an *epitaxial layer.* It is one of the major advances in the industry. Epitaxial layers were a part of the technology as early as 1950.[17] Since then, deposited silicon layers have found additional uses in advanced bipolar device design, a high-quality base

for CMOS circuits, and silicon epitaxial layers deposited on sapphire and other substrates (Chapter 14). Gallium arsenide and other III–IV and II–VI films are also deposited in epitaxial films. Epitaxial films that are the same material as the substrate (silicon on silicon) are *homoepitaxial.* When the deposited material is different from that of the substrate (GaAs on silicon), the film is called *heteroepitaxial.*

Epitaxial Silicon

The term *epitaxial* comes from the Greek word meaning "arranged upon." In semiconductor technology, it refers to the single crystalline structure of the film. The structure comes about when silicon atoms are deposited on a bare silicon wafer in a CVD reactor (Fig. 12.25). When the chemical reactants are controlled and the system parameters set correctly, the depositing atoms arrive at the wafer surface with sufficient energy to move around on the surface and orient themselves to the crystal arrangement of the wafer atoms. Thus, an epitaxial film deposited on a ⟨111⟩-oriented wafer will take on a ⟨111⟩ orientation.

If, on the other hand, the wafer surface has a thin layer of silicon dioxide, an amorphous surface layer, or contamination, the depositing atoms have no structure to which they can align. The resulting film structure is polysilicon. This condition is useful for some applications, such as MOS gates, and is unwanted if the goal is to grow a single-crystal film structure.

Silicon tetrachloride source chemistry. A number of different sources are used for the deposition of epitaxial silicon (Fig. 12.26). Deposition

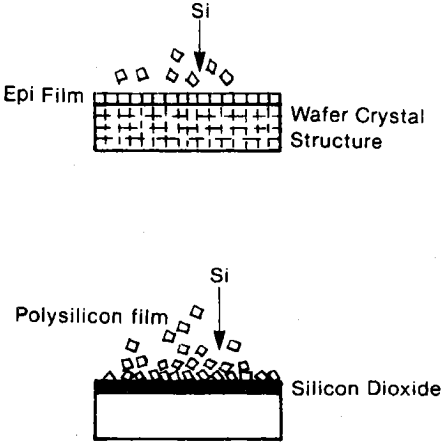


Figure 12.25 Epitaxial and polysilicon film growth.

Silicon tetrachloride	$SiCl_4 + 2H_2 \leftrightarrow Si + 4HCl$
Silane	$SiH_4 + \text{heat} \rightarrow Si + 2H_2$
Dichlorosilane	$SiH_2Cl_2 \leftrightarrow Si + 2HCl$

Figure 12.26 Epitaxial silicon chemical sources.

temperature, film quality, growth rate, and compatibility with a particular system are factors in choosing a silicon source. An important process parameter is the deposition temperature. The higher the temperature, the faster the growth rate. Faster growth rates create more crystal defects and film cracking and stress. Higher temperatures also cause higher levels of autodoping and out-diffusion. (These effects are described in the following text.)

Silicon tetrachloride ($SiCl_4$) is the favored source of silicon for deposition of silicon. It allows a high formation temperature (growth rate) and has a reversible chemical reaction. In Fig. 12.26, there is a double-headed arrow, which indicates that the reaction creates silicon atoms in one direction and removes (etches) silicon in the other direction. Within the reactor, these two reactions compete with each other.

Initially, the silicon surface is etched preparing it for the deposition reaction. In the second stage, the deposition of silicon is faster than the etch, with the net result of a deposited film.

The graph in Fig. 12.27 shows the effect of the two reactions. With an increasing percentage of $SiCl_4$ molecules in the gas stream, the deposition rate first increases. At the 0.1 ratio, the etching reaction starts to dominate and slows down the growth rate. This latter reaction is actually one of the first events in the reactor. Hydrogen chloride (HCl) gas is metered into the chamber, where it etches away a thin layer of the silicon surface, preparing it for the silicon deposition.

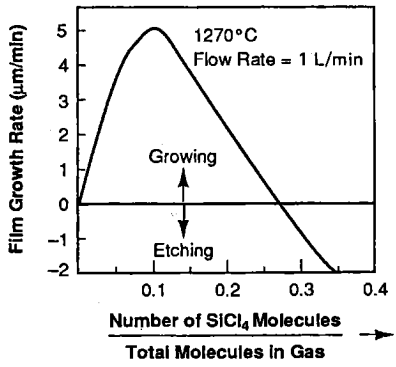


Figure 12.27 Growth-etch characteristics of $SiCl_4$ epitaxial deposition.

Silane source chemistry. The second-most-used silicon source chemistry is silane (SiH_4). Silane offers the advantage of not requiring a second reaction gas. It forms silicon atoms by decomposing when heated. The reaction takes place several hundred degrees lower than a silicon tetrachloride deposition, which is attractive from an autodoping and wafer warping perspective. Also, silane does not produce pattern shift (see "Epitaxial film quality," below). Unfortunately, the reaction occurs at all locations in the system, creating a powdery film inside the reactor which, in turn, contaminates the wafers. Silane finds more use as a source for polysilicon and silicon dioxide depositions.

Dichlorosilane source chemistry. Dichlorosilane (SiH_2Cl_2) is also a lower-temperature silicon source that is used for thin epitaxial films. The lower temperature reduces autodoping and solid-state diffusion from previously diffused buried layers and provides a more uniform crystal structure.

Epitaxial film doping. One of the advantages of an epitaxial film is the precise doping and doping range available by the process. Silicon wafers are manufactured in a concentration range of approximately 10^{13} to 10^{19} atoms/cm^3. Epitaxial films can be grown from 10^{12} to 10^{20} atoms/cm^3. The upper limit is close to the solid solubility of phosphorus in silicon.

Doping in the film is achieved by the addition of a dopant gas stream to the deposition reactants. The sources of the dopant gases are exactly the same chemistries and delivery systems used in deposition doping furnaces. In effect, the CVD deposition chamber is turned into a doping system. In the chamber, the dopants become incorporated into the growing film, where they establish the required resistivity. Both N- and P-type films can be grown on either N- or P-type wafers. The classic epitaxial film in bipolar technology is an N-type epitaxial film on a P-type wafer.

Epitaxial film quality. Epitaxial film quality is a prime concern of the process. In addition to the usual considerations over contamination, there are a number of faults specifically associated with epitaxial growth. Contaminated systems can cause a problem called *haze*.[19] Haze is a surface problem that varies from a microscopic disruption to severe cases that are observable as a dull matte finish. Haze comes about from residual oxygen in the reactant gases or from leaks in the system.

Contaminants on the surface at the start of the deposition result in an accelerated growth known as *spikes* (Fig. 12.28). The spikes can be

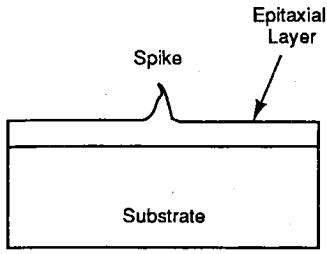


Figure 12.28 Epitaxial growth spike.

as high as the film thickness. They cause holes and disruption in photoresist layers and other deposited films.

During the growth, a number of crystal problems can occur. One is *stacking faults*. A stacking fault is due to the inclusion of an extra atomic plane with a corresponding "dislocation" of the atoms around the plane. A stacking fault begins at the surface and "grows" to the surface of the film. The shape of the stacking fault depends on the orientation of the film and wafer. Faults in ⟨111⟩-oriented films have a pyramidal shape (Fig. 12.29), whereas ⟨100⟩-oriented wafers form rectangular- shaped stacking faults. The faults are detected by either X-ray or etching techniques.

A growth problem associated with ⟨111⟩ wafers is *pattern shift*. This problem occurs when the deposition rate is too high and the film planes grow at an angle to the surface. Pattern shift is a problem when alignment to a subsurface pattern relies on locating it from a film surface step (Fig. 12.30). Another major growth problem is *slip*. This condition comes about from poor control of the deposition parameters and results in a "slippage" of the crystal along plane interfaces (Fig. 12.31).

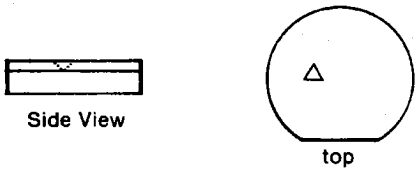


Figure 12.29 Stacking fault on ⟨111⟩ Si.

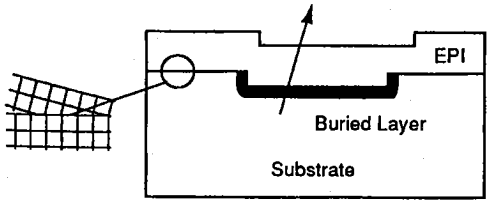


Figure 12.30 Epitaxial pattern shift.

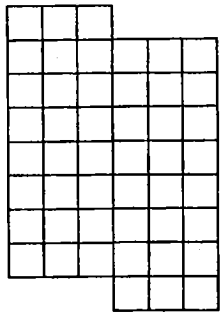


Figure 12.31 Crystal slip.

There are two issues associated with the temperature of the deposition: autodoping and out-diffusion. Autodoping of the growing film occurs when dopant atoms from the back of the wafer diffuse out from the wafer (Fig. 12.32), mix in the gas stream, and become incorporated into the growing film. In the film, they change the resistivity and the conductivity level. *Autodoping* in a P-type film, grown over an N-type wafer, will be less P-type than intended and have a lower P-type concentration as the autodoped atoms neutralize a number of the P-type atoms in the film.

Out-diffusion causes the same effect but at the epitaxial layer-wafer interface. The source of the out-diffused atoms is doped regions diffused into the wafer before the epitaxial deposition. In bipolar devices,

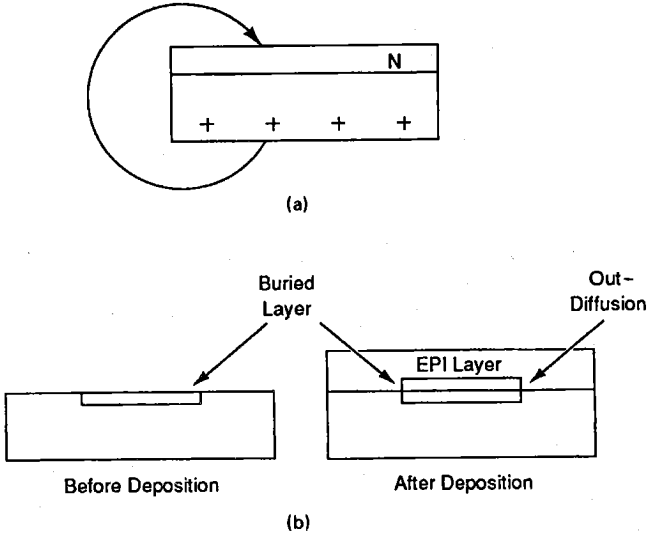


Figure 12.32 (*a*) Epitaxial autodoping and (*b*) out-diffusion.

the regions are called *buried layers* or *subcollectors*. In the usual format, the buried layer is an N-type region in a P-type wafer, over which is grown an N-type epitaxial layer. During the deposition, the N-type atoms diffuse out and become incorporated into the bottom of the epitaxial film, changing the concentration. In the extreme, the buried layer can out-diffuse up into the bipolar device structure causing electrical malfunctions.

CMOS epitaxy. Until the late 1970s, the dominant use of epitaxial films was as the collector region of bipolar transistors. The technique provided a quality substrate for device operation and a clever means of isolating adjacent devices (see Chapter 16). A newer and perhaps more dominant use of silicon epitaxial films is for CMOS circuit wafers. The need for an epitaxial layer was driven by a CMOS circuit problem called *latch-up* (see Chapter 16).

Epitaxial process. A typical epitaxial process starts with a complete and rigorous cleaning of the wafer surface prior to loading the reactor. Within the deposition chamber, a number of steps take place to correctly deposit the film. A typical $SiCl_4$ epitaxial process is shown in Fig. 12.33. The first several steps are a gas-phase cleaning of the wafer surface. Deposition follows the cleaning with a cool-down cycle at the end. During all the steps, control of the temperature and gas flows is critical.

Selective epitaxial silicon. Advancements in epitaxial deposition systems have introduced the selective growth of epitaxial films. Whereas the epitaxial films for bipolar and CMOS substrates are deposited on the entire wafer, in selective growth they are grown through holes in either silicon dioxide or silicon nitride films. The wafer is positioned in

Cycle	Temperature	Gas	Purpose
1	Room	N_2	Purge air from system
2	Room	H_2	Reduce any organic contaminants of wafers in system
3	(Heating)	N_2	Bring system to deposition temperature
4	Deposition Temperature	HCl	Etch wafer to prepare surface for epi deposition
5	Deposition Temperature	Source + Dopant + Carrier	Grow epitaxial film
6	(Heat Off)	N_2	Purge system of reactant gases

Figure 12.33 Typical $SiCl_4$ epitaxial deposition process.

the reactor chamber, and the epitaxial film grows directly on the silicon exposed at the bottom of the hole (Fig. 12.34). As the film grows, it takes on the crystal orientation of the underlying wafer. An advantage of such a structure is that devices formed in the surfaces of the epitaxial regions are isolated from each other by the oxide or nitride regions.

If the deposition is allowed to continue onto the isolating surface, the structure of the film switches to a polysilicon structure. Another outcome of extended deposition is that the overlaying deposited layer becomes entirely epitaxial in nature. All of these outcomes add attractive structure options for advanced device design.

Polysilicon and Amorphous Silicon Deposition

Until the advent of silicon-gate MOS devices (Fig. 12.35) in the mid-1970s, polysilicon layers had little or no use in device structures. Silicon-gate device technology drove the need for reliable processes to deposit thin layers of polysilicon. By the mid-1980s, polysilicon seemed to be the workhorse material of advanced devices. In addition to MOS gates, polysilicon finds use as load resistors in SRAM devices, trench fills, multilayer poly in EEPROMs, contact barrier layers, emitters in bipolar devices, and as part of silicide metallization schemes (see Chapters 13 and 16).

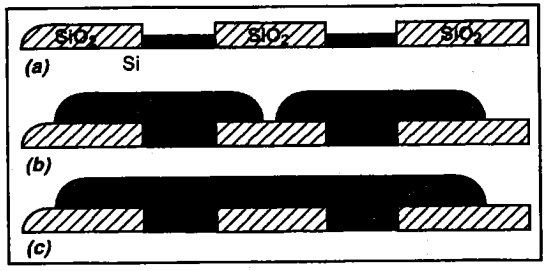


Figure 12.34 Steps in selective epitaxial growth.

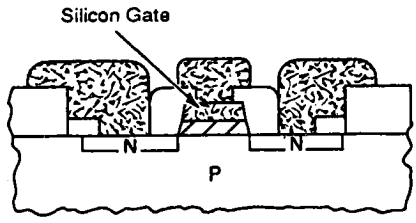


Figure 12.35 Cross section of silicon gate MOS transistor.

Early processes involved simply placing the oxide-covered wafers in a horizontal APCVD system and letting the polysilicon deposit on the oxide. The major difference of early polysilicon depositions from epitaxial depositions was the use of silane sources. While silane is not favored for epitaxial film deposition, it is more than adequate for polysilicon depositions.

Typical polysilicon deposition processes take place in the 600 to 650°C range. The deposition may be from either 100 percent silane or from gas streams containing N_2 or H_2. The structure of polysilicon was previously described as a total nonarrangement of the silicon atoms. In the case of deposited polysilicon, the structure is somewhat different. During the early stages of deposition, at temperatures below 575°C, the structure is amorphous (no structure). The polysilicon structure formed by deposition techniques consists of small pockets (crystallites or grains) of single-crystalline silicon separated by grain boundaries. This structure is called *columnar poly.*

The importance of grain size and grain boundary consistency shows up in the electrical current flow characteristics of the films. Current resistance comes as the current crosses the grain boundaries. The larger the grain boundaries, the higher the resistance. The achievement of consistent current flow from device to device and within a device is dependent on a well controlled polysilicon structure. One of the advantages claimed for the use of H_2 in the gas stream is the reduction of surface impurities and moisture, which in turn results in a reduced grain size. Moisture or oxygen impurities in the system cause the growth of silicon dioxide within the structure. The oxide increases the resistance of the film and its etchability in subsequent masking steps.

All of the system's usual operating parameters (temperature, silane concentration, pump speed, nitrogen flow, and other gas flows[19]) affect the deposition rate and the grain size. Often, the wafers will receive a postdeposition anneal in the 600°C range to further crystallize the film. The process of recrystallization goes on whenever the wafers go through a high-temperature process. The grain size and electrical parameters of the polysilicon film on the finished device or circuit are never the same as the deposited film.

Also influencing the grain size is the presence of dopants in the gas stream. In many devices or circuits, a strip of polysilicon functions as a conductor which requires doping to decrease its resistivity. Doping can be done by diffusion before or implantation after the deposition.

In situ doping takes place by adding gas dopant sources in the source cabinet and metering them into the chamber. When diborane (boron source) is added, there is a large increase in the deposition rate. An opposite effect takes place when phosphine (phosphorus source) or

arsine (arsenic source) is the dopant gas. Undesirable effects of *in situ* doping are a loss of film uniformity, doping uniformity, and control of the deposition rate.

Doped polysilicon film resistivities are less than those of equally doped epitaxial or bulk silicon. The lower resistivities are due to dopants being trapped in the grain boundaries.

Most polysilicon layers are deposited with LPCVD systems that provided good productivity and lower deposition temperatures. LPCVD provides good step coverage (Fig. 12.36), a requirement because polysilicon layers are usually deposited later in the process, and the surface has become varied in its topography. Single-chamber polysilicon LPCVD systems offer the advantage of higher deposition rates without raising the temperature.[20]

SOS and SOI

These two acronyms stand for *silicon on sapphire* and *silicon on insulator*. Both refer to the deposition of silicon on a nonsemiconductor surface. The need for such structures came about from the limits placed on some MOS devices by the presence of a semiconducting substrate under the active device. These problems are resolved by forming a silicon layer on an insulating substrate. The first substrate used for this purpose was sapphire (SOS). As different substrates were investigated, the term was expanded to the more general silicon on insulator (SOI).

One technique is a direct deposition on the substrate followed by a recrystallization process (laser heating, strip heaters, oxygen implan-

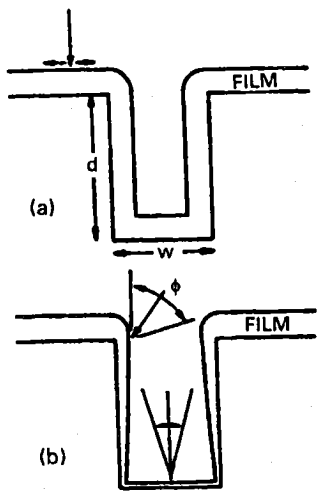


Figure 12.36 Step coverage. (*a*) Good step coverage and (*b*) nonconformal coverage.

tation) to create a usable film.[22] Another approach is a selective deposition through holes in a surface oxide with an overgrowth to form the continuous film.

Another SOI method is *SIMOX*. In this process, the top layer of a wafer is converted to oxide with a heavy oxygen implant. An epitaxial layer is grown on top of the oxide. There is some exploration of *bonded wafers*. This approach has two wafers bonded together followed by the thinning (grinding and polishing) of one to device layer thickness.[22]

Gallium Arsenide on Silicon

Gallium arsenide is a great III-V semiconductor material. However it is fragile and limited to wafer sizes of 4 inch (102 mm) diameter. Attempts to grow GaAs films on silicon wafers have been thwarted by the mismatch of the lattice sizes of each. During the growing process, the mismatch cause dislocations that degrade device performance. A new approach is to first grow a thin layer of strontium titanate on the silicon wafer. It reacts with the silicon to form a amorphous silicon dioxide layer. When a film of GaAs is deposited, the silicon dioxide layer acts as a cushion that absorbs the mismatch, thus allowing a single crystalline layer.[24]

Insulators and Dielectrics

CVD is the favored method of depositing films that will function in the device or circuit as insulators or dielectrics. The two films in widespread use are silicon dioxide and silicon nitride. In general, the two films find a multiplicity of uses in device and circuit designs. While they have processing and quality differences, they must meet the same general requirements as other deposited films.

Silicon dioxide. Deposited silicon dioxide films are best known from their long-term use as a final passivation layer covering the completed wafer. In this role, they provide physical and chemical protection to the underlying circuit devices and components. Deposited silicon dioxide films used as a protective top layer are known by the proprietary terms *Vapox*, *Pyrox*, and *Silox*®. Vapox (vapor-deposited oxide) is a term coined by Fairchild engineers. Pyrox stands for pyrolitic oxide. Silox is a registered trademark of Applied Materials, Inc. Sometimes, the layer is simply called a *glass*. This protective role has expanded, and deposited silicon oxide layers are used as interdielectric layers in multimetallization schemes, as insulation between polysilicon and metallization layers, as doping barriers, as diffusion sources, as isola-

tion regions. Silicon dioxide has become a major part of silicon gate structures.

There are gate stacks, consisting of thermal oxide/silicon dioxide or oxynitride/silicon dioxide (TEOS deposited), and various silicon dioxide fillers for plugs in multimetal designs.[25]

CVD-deposited silicon dioxide films vary in structure and stoichiometry from thermally grown oxides. Depending on the deposition temperature, deposited oxides will have a lower density and different mechanical properties, such as index of refraction, resistance to cracking, dielectric strength, and etch rate. These factors are highly affected by the addition of dopants to the film. In many processes, the deposited film will receive a high-temperature anneal, a process called *densification*. After the densification, the deposited silicon dioxide film is close to the structure and properties of a thermal oxide.

The need for a low-temperature-deposited SiO_2 was dictated by the unacceptable alloying of aluminum and silicon at temperatures above 450°C. The early deposition process used was a horizontal conduction heated APCVD system from silane and oxygen by the reaction

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

This process produced films that were of unacceptable quality for use in advanced device designs and on larger wafers due to the poorer film quality produced by the 450°C deposition temperature.

The development of LPCVD systems made possible higher-quality films, especially for the factors of step coverage and lower stress. LPCVD processes are the preferred deposition techniques from both quality and productivity considerations. High-temperature (900°C) LPCVD of silicon dioxide is performed with a dichlorosilane reaction with nitrous oxide.

$$SiCl_2 + 2NO_2 \rightarrow SiO_2 + 2N_2 + HCl$$

Tetraethyl orthosilicate (TEOS). By far, the majority of silicon dioxide are deposited from $Si(OC_2H_5)$ sources. The source is known as tetraethyl orthosilicate, or TEOS. TEOS history goes back to the 1960s.

Early systems relied on the simple pyrolysis of the TEOS in the 750°C range. Current depositions are based on the hot-wall LPCVD systems established in the 1970s, with temperatures the 400°C+ range. TEOS sources used with plasma-assistance (PECVD or PE-TEOS) allowed deposition temperatures in the sub-400°C range.[24] This process faces limits on conformal coverage of high aspect ratio patterns in 0.5-μm devices. Step coverage is improved by the addition of ozone (O_3) to the gas stream.[26]

Another option is the reaction of silane with nitrous oxide in an argon plasma.

$$SiH_4 + 4N_2O \rightarrow SiO_2 + 4N_2 + 2H_2O$$

Doped silicon dioxide. Silicon dioxide layers are doped to improve their protective characteristics and flow properties, or for use as dopant sources. The earliest dopant used with deposited oxides was phosphorus. The phosphorus source is phosphine (PH_3) gas added to the deposition gas stream. The resultant glass is called *phosphorus silicate glass* or PSG. Within the glass, the phosphorus is in the form of phosphorus pentoxide (P_2O_5), making the glass a dual compound or, more correctly, a binary glass.

The role of the phosphorus is threefold. The added dopant increases the moisture-barrier property of the glass. Mobile ionic contaminants become attached to the phosphorus and are prevented from traveling into the wafer surface. This action is called *gettering*. The third result is an increase of the flow characteristics (Fig. 12.37), which aid the planarization of the glass surface after a heating step in the 1000°C range. The phosphorus content is limited to about 8 weight by percent. Above this level, the glass becomes hydroscopic and attracts moisture. The moisture can react with the phosphorus, form phosphoric acid, and attack underlying metal lines.

Boron is often added to the glass from a diborane (B_2H_6) source. The purpose of the boron is to also aid the flow characteristics (Fig. 12.37).

The resultant glass is called a borosilicate glass (BSG). The boron and phosphorus are often used together in the glass. The result is referred to as BPSG (borophosphorus silicate glass).

Silicon nitride. Silicon nitride is a replacement for silicon dioxide uses, especially for top layer protection. Silicon nitride is harder, which provides better scratch protection, is a better moisture and sodium barrier

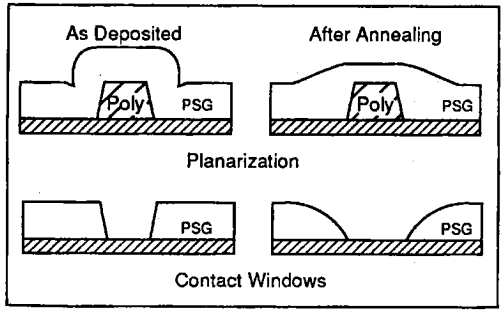


Figure 12.37 Planarization of surface by flowing glass.

(without doping), has a higher dielectric strength, and resists oxidation. The latter property has led to its use in the local oxidation of silicon (LOCOS) for isolation purposes. Figure 12.38 illustrates the process, where patterned islands of silicon nitride prevent oxidation under the islands. After thermal oxidation and removal of the nitride, there are wafer surface regions ready for device formation, separated by isolating regions of oxide. A disadvantage of silicon nitride is that it does not flow as easily as silicon oxide and is more difficult to etch. The etch restriction has been overcome with the development of plasma etch processes.

An early limit on the use of silicon nitride protective films was the lack of a low-temperature deposition process. In APCVD systems, a temperature of 700 to 900°C is required for the deposition of silicon nitride from silane or dichlorosilane (Fig. 12.39). The result is a film with the composition Si_3N_4. The reactions also take place in LPCVD reactors but at a temperature low enough for deposition over an aluminum metallization layer. The advent of PECVD has opened up the use of different source chemistries. One use is silane reacted with ammonia (NH_3) or nitrogen in the presence of an argon plasma.

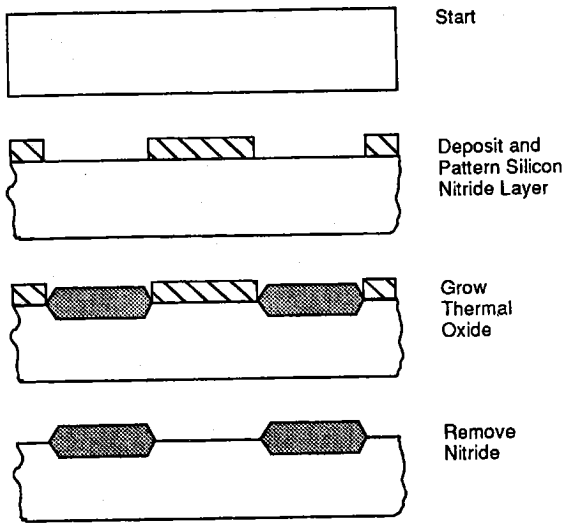


Figure 12.38 LOCOS process.

Silane	$3SiH_4 + 4NH_3 \rightarrow Si_3N_4 + 12H_2$
Dichlorosilane	$3SiCl_2H_2 + 4NHCl + 6H_2 + 6HCl$

Figure 12.39 Silicon nitride deposition reactions.

High-k and low-k dielectrics

In addition to the dielectric films mentioned above, there are numerous other types deposited for special applications. They fall into the broad categories of high-k and low-k. In Chapter 2, the basics of a capacitor were described. For review, the k value of a material is called its *dielectric constant.* It relates to the level of capacitance a capacitor will have. High-k materials produce capacitors with high capacitance that are desirable for charge storage. Uses of high-k dielectrics are discussed in Chapter 16.

Low-k materials are used in metallization systems as barriers between the wafer surface and the primary metal system. In this case, the capacitor function should be low to facilitate conduction. These low-k materials are discussed in Chapter 13.

Conductors

The traditional metal conductors of aluminum and aluminum alloys are deposited by evaporation or sputtering techniques. The advent of silicon gate MOS transistors added doped polysilicon as a device conductor. Plus, the advent of multimetal structures and new conducting materials has thrust CVD and PVD techniques into the conducting metal business. The techniques and use of these deposited metals are explained in the next chapter.

Review Questions

1. Sketch and name the major subsystems of a basic CVD system.
2. Describe the differences between APCVD, LPCVD, and PECVD.
3. Define an epitaxial film.
4. Why is the deposition temperature of a silicon dioxide passivation layer limited to 450°C?
5. List an advantage of a horizontal vertical-flow CVD reactor.
6. Write the reaction equation for the deposition of silicon from silicon tetrachloride.
7. Describe the difference between MBE, VPE, and MOCVD systems.
8. What wafer surface condition is necessary for the deposition of a polysilicon layer?
9. Describe the advantages of plasma-assisted depositions compared to APCVD systems.
10. Why, and with what, are deposited silicon dioxide layers doped?

References

1. J. Hayes and P. Van Zant, *CVD Today Seminar Manual, Semiconductor Services*, San Jose, CA, 1985, p. 9.
2. S. Wolf and R. Tauber, *Silicon Processing for the VLSI Era*, Lattice Press, Sunset Beach, CA, 1986, p. 165.
3. M. L. Hammond, "Epitaxial Silicon Reactor Technology: A Review," *Solid State Technology*, May 1988, p. 160.
4. M. L. Hammond, "Epitaxial Silicon Reactor Technology: A Review," *Solid State Technology*, May 1988, p. 160.
5. J. Hayes and P. Van Zant, *CVD Today Seminar Manual*, Semiconductor Services, San Jose, CA, 1985, p. 13.
6. S. Wolf and R. Tauber, *Silicon Processing for the VLSI Era*, Lattice Press, Sunset Beach, CA, 1986, p. 165.
7. B. Meyerson, H. Kaiser, and S. Schultz, "Extending Silicon's Horizon through UHV/CVD," *Semiconductor International*, Cahners Publishing, March 1994, p. 73.
8. P. Singer, "The Future of Dielectric CVD: High Density Plasma?" *Semiconductor International*, July 1997, p. 127.
9. A. Braum, "ALD Breaks Materials, Conformity Barriers," *Semiconductor International*, October, 2001, p. 52.
10. R. Williams, *Gallium Arsenide Processing Techniques*, Artech House, Dedham, MA, 1984, p. 44.
11. S. Wolf and R. Tauber, *Silicon Processing for the VLSI Era*, Lattice Press, Sunset Beach, CA, 1986, p. 157.
12. P. Burggraaf, "The Status of MOCVD Technology," *Semiconductor International*, Cahners Publishing, July 1993, p. 81.
13. P. Singer, "Molecular beam epitaxy," *Semiconductor International*, Oct. 1986, p. 42.
14. P. Burggraaf, "The Growing Importance of MOCVD," *Semiconductor International*, November 1986, p. 47.
15. P. Burggraaf, "The Growing Importance of MOCVD," *Semiconductor International*, November 1986, p. 48.
16. P. Burggraaf, "The Status of MOCVD Technology," *Semiconductor International*, Cahners Publishing, July 1993, p. 81.
17. A. Thompson, R. Stall, and B. Droll, "Advances in Epitaxial Deposition Technology," *Semiconductor International*, Cahners Publishing, July 1994, p. 173.
18. M. L. Hammond, "Epitaxial Silicon Reactor Technology: A Review," *Solid State Technology*, May 1988, p. 159.
19. R. P. Roberge, et al., "Gaseous Impurity Effects in Silicon Epitaxy," *Semiconductor International*, January 1988, p. 81.
20. S. M. Sze, *VLSI Technology*, McGraw-Hill, New York, 1983, p. 119.
21. M. Venkatesan and I. Beinglass, "Single-wafer Deposition of Polycrystalline Silicon," *Solid State Technology*, PennWell Publishing, March 1993, p. 49.
22. L. Jastrzebski, "Silicon CVD for SOI: Principles and Possible Applications," *Solid State Technology*, September 1984, p. 239.
23. K. Yallup, "SOI Provides Total Dielectric Isolation," *Semiconductor International*, Cahners Publishing, July 1993, p. 134.
24. P. Singer, "GA-As-on-Silicon, Finally!" *Semiconductor International*, October, 2001, p. 36
25. P. Singer, "Directions in Dielectrics in CMOS and DRAMs," *Semiconductor International*, Cahners Publishing, April 1994, p. 57.
26. B. L. Chin and E. P. van de Ven, "Plasma TEOS Process for Interlayer Dielectric Applications," *Solid State Technology*, April 1988, p. 119.
27. K. Maeda and S. Fisher, "CVD TEOS/O3: Development History and Applications," *Solid State Technology*, PennWell Publishing, June 1993, p. 83.

Chapter

13

Metallization

Overview

Fabrication of circuits is divided into two major segments. First the active and passive parts are fabricated in and on the wafer surface. This is called the *front end of the line* (FEOL). In the *back end of the line* (BEOL), the metal systems necessary to connect the devices and different layers are added to the chip. In this chapter, the materials, specifications, and methods used to complete the metallization segment is presented along with other uses of metals in chip manufacturing. Vacuum pumps (used in CVD, evaporation, ion implant, and sputtering systems) are explained at the end of the chapter.

Objectives

Upon completion of this chapter, you should be able to:

1. List the requirements of a material for use as a chip surface conductor.
2. Draw cross sections of single and a two layer metal schemes.
3. Describe the purpose and operation a low-k dielectric layer.
4. Make a list of three materials used in the metallization of semiconductor devices and identify their specific use(s).
5. Describe the principle of sputtering.
6. Draw and identify the parts of a sputtering system.
7. Describe the principle and operation of turbo and cryogenic high-vacuum pumps.

Introduction

The most common use of metal films in semiconductor technology is for surface wiring. The materials, methods, and processes of "wiring" the component parts together is generally referred to as the *metallization process.* Depending on device complexity and performance requirements, the circuit may require a single-metal system, or a multiple-level system. It may use an aluminum alloy or gold as the conducting metal.

Conductors—Single-Level Metal

In the MSI era metallization was relatively straightforward (Fig. 13.1), requiring only a single-level metal process. Small holes, called *contact holes* or *contacts*, are etched through the surface layers, exposing areas on the device/circuit component parts. Following contact masking, a thin layer (10,000 to 15,000 Å) of the conducting metal is deposited by vacuum evaporation, sputtering, or CVD techniques over the entire wafer. The unwanted portions of this layer are removed by a conventional photomasking and etch procedure or by lift-off. This step leaves the surface covered with thin lines of the metal that are called *leads*, *metal lines*, or *interconnects*. Generally, a

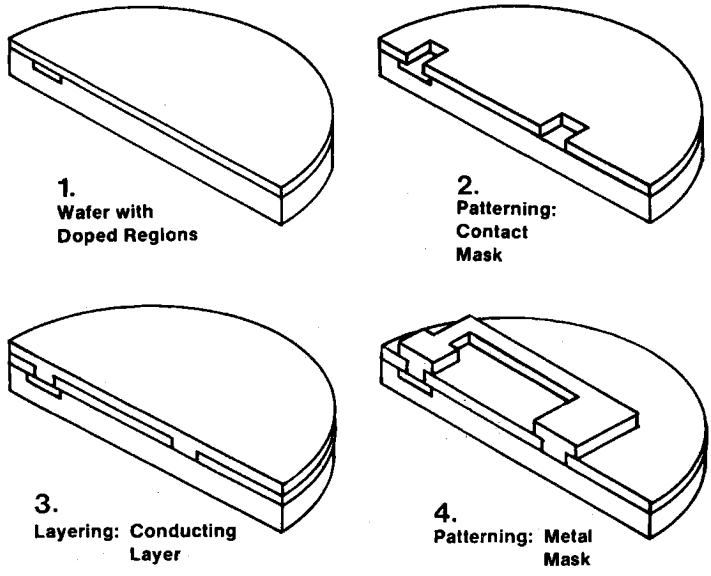


Figure 13.1 Metallization sequence.

heat-treatment step, called *alloying*, is performed after metal patterning to ensure good electrical contact between the metal and the wafer surface.

Regardless of the structure, a metal system must meet the following criteria:

- Good electrical current-carrying capability (current density)
- Good adhesion to the top surface of the wafer (usually SiO_2)
- Ease of patterning
- Good electrical contact with the wafer material
- High purity
- Corrosion resistance
- Long-term stability
- Capable of deposition in uniform void- and hillock-free films
- Uniform grain structure

Conductors—Multilevel Metal Schemes

Increasing chip density has placed more components on the wafer surface, which in turn has *decreased* the area available for surface wiring. The answer to this dilemma has been multilevel metallization schemes with two to four individual metal layers (Fig. 13.2). By 2012, it is expected that chips will carry up to nine wiring levels.[1] A typical two-metal stack is shown in Fig. 13.3. The stack starts with a *barrier layer* formed by silicidation of the silicon surface to produce a lowered electrical resistance between the surface and the next layer. Barrier layers also prevent alloying of aluminum and silicon if aluminum is the conducting material. Next comes a layer of an dielectric material, called an *intermetallic dielectric layer* (IDL or IMD) that provides the electrical isolation between metal layers. This dielectric may be a deposited oxide, silicon nitride, or a polyimide film. This layer receives a masking step that etches new contact holes, called *vias* or *plugs*, down to the first-level metal. Conducting plugs are created by depositing conducting material into the hole. Next, the first-level metal layer is deposited and patterned. The IMD/plug/metal deposition/patterning sequence is repeated for the subsequent layers. A multilevel metal system is more costly, of lower yield, and requires greater attention to planarization of the wafer surface and intermediate layers to create good current-carrying leads.

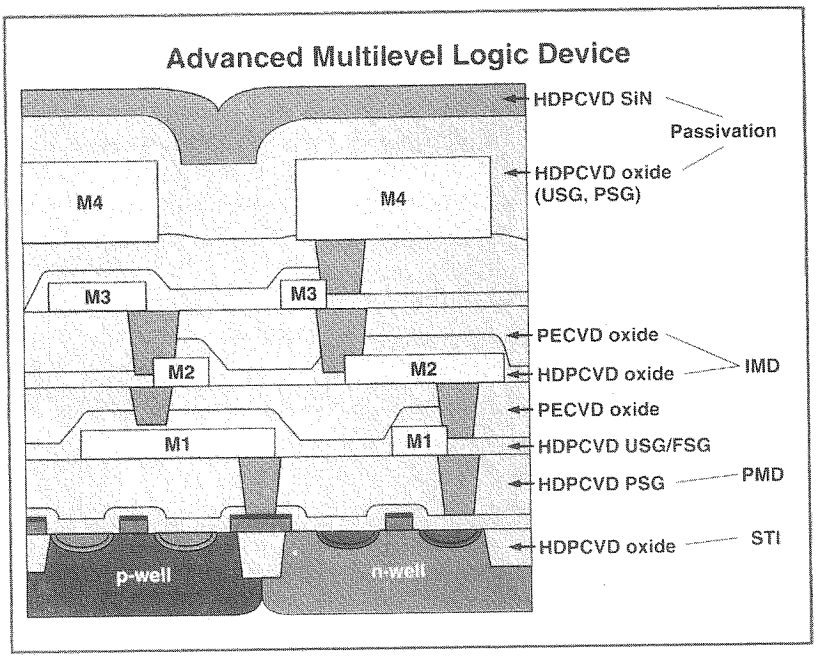


Figure 13.2 Multi-metal level structure. *(Courtesy of* Semiconductor International, *July 1997.)*

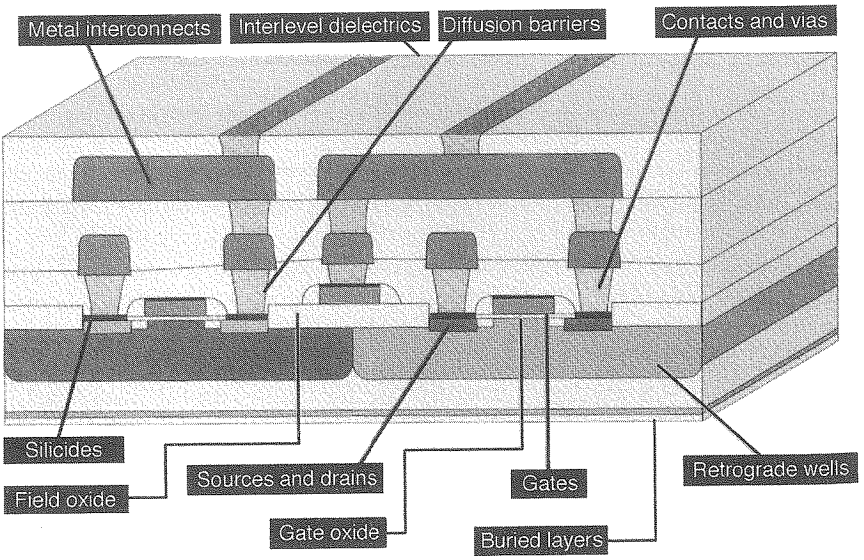


Figure 13.3 Two-metal structure. *(Courtesy of* Semiconductor International, *January 1998.)*

Conductors

Aluminum

This section addresses the three primary materials used for the metal interconnection layers. Prior to the development of VLSI-level circuits, the primary metallization material was pure aluminum. The choice of aluminum and its limitations are instructive to the understanding of metallization systems in general. From an electrical conduction standpoint, aluminum is less conductive than copper and gold. Copper, if used as a direct replacement for aluminum, has a high contact resistance with silicon and raises havoc with device performance if it gets into the device areas. Aluminum emerged as the preferred metal because it avoids the problems just mentioned. It has a low enough resistivity (2.7 $\mu\Omega$-cm),[2] and good current-carrying density. It has superior adhesion to silicon dioxide, is available in high purity, has a naturally low contact resistance with silicon, and is relatively easy to pattern with conventional photolithography processes. Aluminum sources are purified to 5 to 6 "nines" of purity (99.999 to 99.9999 percent).

Aluminum-silicon alloys

Shallow junctions in the wafer surface presented one of the first problems with the use of pure aluminum leads. The problem came with the need to bake aluminum-silicon interfaces to stabilize the electrical contact. This type of contact is called *ohmic* because the voltage-current characteristics behave according to Ohm's law. Unfortunately, aluminum and silicon dissolve into each other and, at 577°, reach a eutectic formation point. A eutectic formation occurs when two materials heated in contact with each other melt at temperatures much lower than their individual melting temperatures. Eutectic formations occur over a temperature range, and the aluminum-silicon eutectic starts to form at about 450°, also the temperature necessary for good electrical contact. The problem is acute with shallow junctions. If the alloy region is deep, it can extend completely through the junction, shorting it out (Fig. 13.4).

Two solutions to this problem are employed. One is a barrier metal layer (see section on barrier metals) that separates the aluminum and silicon and prevents the eutectic alloy from forming. The second is an alloy of aluminum with 1 to 2 percent silicon. During the contact heating step, the aluminum alloys more with the silicon in the alloy and less with the silicon from the wafer. This process is not 100 percent effective, and some alloying between the aluminum and wafer always occurs.

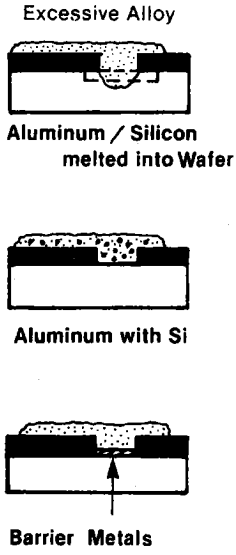


Figure 13.4 Eutectic alloying of aluminum and silicon contacts.

Aluminum-copper alloy

Aluminum suffers a problem called *electromigration*. The problem occurs when long, skinny leads of aluminum are carrying high currents over long distances, as is the situation in VLSI/ULSI circuits. The current sets up an electric field in the lead that decreases from the input side to the output. Also, heat generated by the flowing current sets up a thermal gradient along the lead. The aluminum in the lead becomes mobile and diffuses within itself along the direction of the two gradients. The first effect is a thinning of the lead. In the extreme, the lead can become completely separated. Unfortunately, this event usually takes place after the circuit is in operation in the field, causing a failure of the chip. Prevention or moderation of electromigration is achieved by depositing an alloy of aluminum and 0.5 to 4 percent[3] copper or an alloy of aluminum and 0.1 to 0.5 percent titanium. Aluminum alloys containing both copper and silicon are often deposited on the wafer to resolve both alloying and electromigration problems.

Drawbacks of aluminum alloys are an increased complexity for the deposition equipment and process, and different etch rates, as well as an increase in film resistivity compared to the pure aluminum. The amount of the increase varies with the alloy composition and heat treatments but can be as much as 25 to 30 percent.[4]

Barrier metals

A method of preventing the eutectic alloying of silicon and aluminum metallization is by using a barrier layer. Both titanium-tungsten

(TiW) and titanium nitride (TiN) layers are used. TiW is sputter-deposited onto the wafer into the open contacts before the aluminum or aluminum alloy deposition takes place. The TiW deposited on the field oxide is removed from the surface during the aluminum etch step. Sometimes, a first layer of platinum silicide is formed on the exposed silicon before the TiW is deposited.

Titanium nitride layers can be placed on the wafer by all the deposition techniques: evaporation, sputtering, and CVD. It can also be formed by the thermal nitridation of a titanium layer at 600°C in an N_2 or NH_3 atmosphere.[5] CVD titanium nitride layers have good step coverage and can fill submicron contacts. A layer of titanium is required under TiN films to provide a high-conductivity intermediate with silicon substrates.

With copper metallization, the barrier is also critical. Copper inside the silicon ruins device performance. Barrier metals used are TiN, tantalum (Ta), and tantalum nitride (TaN).[6]

Refractory metals and refractory metal silicides

Although the limitations of electromigration and eutectic alloying have been made manageable by aluminum alloys and barrier metals, the issue of contact resistance may prove to be the final limit on aluminum metallization. The overall effectiveness of a metal system is governed by the resistivity, length, thickness, and total contact resistance of *all* the metal-wafer interconnects. In a simple aluminum system, there are two contacts: the silicon/aluminum interconnect and the aluminum interconnect/bonding wire. In a ULSI circuit with multilevel metal layers, barrier layers, plug fills, polysilicon gates and conductors, and other intermediate conductive layers, the number of connections becomes very large. The addition of all the individual contact resistances can dominate the conductivity of the metal system.

Contact resistance is influenced by the materials, the substrate doping, and the contact dimensions. The smaller the contact size, the higher the resistance. Unfortunately, ULSI chips have smaller contact openings, and large gate array chip surfaces can be as much as 80 percent contact area.[7] These two factors make the contact resistance the dominant factor in VLSI metal system performance. Aluminum-silicon contact resistance, along with the alloying problem, have led to the investigation of other metals for VLSI metallization. Polysilicon has a lower contact resistance than aluminum and is in use in MOS circuits (Fig. 13.5).

Refractory metals and their silicides offer lower contact resistance. The refractory metals of interest are titanium (Ti), tungsten (W), tantalum (Ta), and molybdenum (Mo). Their silicides form when they are

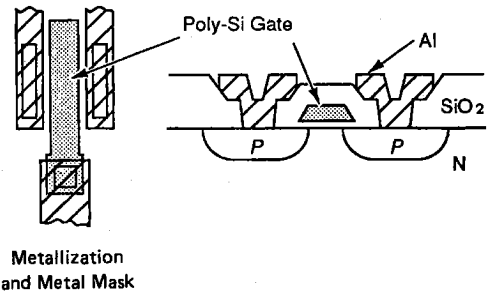


Figure 13.5 Silicon gate electrode extended for metallization lead.

alloyed on a silicon surface (WSi_2, $TaSi_2$, $MoSi_2$ and $TiSi_2$). The refractory metals were first proposed for metallization in the 1950s, but they stayed in the background due to a lack of a reliable deposition method. That situation has changed with the development of LPCVD and sputtering processes.

All modern circuit designs, especially MOS circuits, use refractory metals or their silicides as intermediate (plugs), barrier, or conducting layers. The lower resistivities and lower contact resistances (Fig. 13.6) make them attractive for conducting films, but impurities and deposition uniformity problems make them less attractive for MOS gate electrodes. The solution to the problem has been the polycide and silicide gate structures, which are combinations of a silicon gate topped by a silicide. The details of this structure are explained in Chapter 16.

A popular use of refractory metals is the filling of via holes in multilevel metal structures. The process is called *plug filling*, and the filled via is called a *plug* (Fig. 13.2). The vias are filled by either selective tungsten deposition through surface holes onto the first layer metal or by CVD techniques.[8] Of the available refractory metals, tungsten finds a lot of use as aluminum-silicon barriers, MOS gate interconnects, and for via plugs.

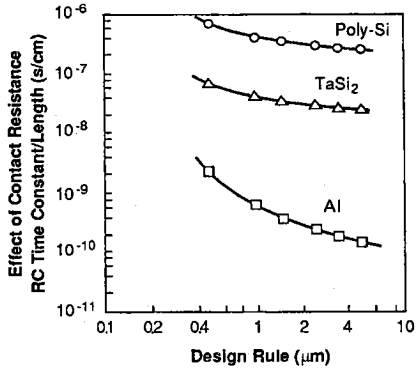


Figure 13.6 Effect of contact resistance on RC time constant.

One of the attractions of copper metallization is that copper can be the plug material, creating a monometal system that minimizes intermetal resistances.

Copper

Aluminum metallization ran into a performance barrier as circuits reached hundred-megahertz speeds. Signals must move fast enough through the metal system to prevent processing delays. Also, longer leads and smaller cross sections required for larger chips increase the resistance of the metal wiring system. As the number of contact holes increases, the small contract resistance between aluminum and silicon surfaces adds up to become significant. While aluminum provides a workable resistance, it is difficult to deposit in via holes with aspect ratios up to 10:1. To date, barrier metal schemes, stacks, and refractory metals have been employed to reduce aluminum metal system resistance. Additional resistance reductions needed for 0.25-μΩ-cm (and smaller) devices have renewed interest in copper as a conducting metal. Copper is a better conductor than aluminum, with a resistance of 1.7 μΩ-cm, compared to a 3.1 μΩ-cm value for aluminum. Copper is resistant to electromigration and can be deposited at low temperatures.[9] It also can be used as a plug material. Deposition can be by CVD, sputtering, electroless plating, and electrolytic plating. Drawbacks, besides lack of a learning curve, include etching problems, vulnerability to scratching, corrosion, and the requirement of barrier metals to keep the copper out of the silicon. Nevertheless, the overall benefits of copper led IBM, followed quickly by Motorola, to announce the availability of production copper-based devices in 1998.[10] Novellus Systems estimates that copper-based circuits will operate four times faster than aluminum-based ones and realize a 20 to 30 percent decrease in cost.

All current circuits are being developed with copper metallization and low-k dielectrics. The primary benefits are increased performance and a reduction in the number of metal layers required.

Low-k dielectric materials

In the dual-damascene illustration, the dielectric separating the two metals is silicon dioxide. However, this material presents a problem for high-performance circuits. The slowing of circuit signals results from the combination of the metal resistance (R) and capacitance (C). It is called the *RC constant* of the system. A major contribution to the capacitance factor is the dielectric constant of the material used to separate the metal layers, the intermediate metal dielectric (IMD).

Silicon dioxide has a dielectric constant (k) in the 3.9 range. Successful circuits will require k values dropping to the 1.5 to 2.0 range in the 2006 time frame (Fig. 13.7), according the SIA *International Technology Roadmap for Semiconductors*. In addition to the dielectric property the IMD must have a number of chemical and mechanical properties. They include thermal stability (subsequent metal processes can take an initial film through a number of heat steps up to the 450°C), good etch selectivity, being pin-hole free, enough flexibility to withstand on chip stresses, and compatibility with the other processes.

A number of low-k dielectric materials have been developed to meet ULSI circuit needs. They are listed in Figure 13.8 along with their dielectric constants. The major categories are oxide based materials, organic based and variations of each. The organics, based on poly(alylene)thers (PAE) or hydrido-organic siloxane polymers (HOSP), offer the advantage of spin-on applications. Spin-on processes can provide great uniformity and planarity, and they are less expensive than CVD processes.

Year	1997	1999	2001	2003	2006
Metal levels	5–6	6–7	7	7	7–8
Interconnect length (m/chip)	820	1480	2160	2840	5140
Dielectric constant (k)	3.0–4.1	2.5–3.0	2.0–2.5	1.5–2.0	1.5–2.0

Figure 13.7 Trends in metal and low-k dielectric requirements.

Metal system	Low-k material
Aluminum	parylene HSQ methyl silesequioxane F-doped oxide F-doped amorphous carbon Parylene-F (AF_4) Xerogel
Gold	polyimide BCB Xerogel

Figure 13.8 Low-k materials. *(Source:* Future Fab International, *no. 8, pp. 180.)*

The dual-damascene copper process

Transitioning from aluminum to copper metallization is not a simple matter of switching materials. Copper has its own set of problems and challenges. It is not easy etched by wet or dry techniques. Copper has a high electrical resistance connection with silicon. It diffuses easily through silicon dioxide and can enter the silicon structure. There, it can degrade device performance and create junction leakage problems. Copper does not adhere well to silicon dioxide surfaces, causing structural problems. These challenges led to the development of a unique process specifically designed to overcome the copper problems and produce a high-production process. It features a lithography process, the development of a low-k barrier/liner process, copper electrochemical plating, and a chemical mechanical polishing process.

In Chapter 10, a basic damascene process was introduced. The damascene concept is simple. A trench is formed in a surface dielectric layer using a photolithography process, and the required metal is deposited into it. Usually, the trench is overfilled, requiring a CMP step to replanarize the surface (Fig. 13.9). This process offers superior dimensional control, because it eliminates the variation introduced in a typical metal etch process.

In practice the process is a bit more complicated. Figure 13.10 illustrates a typical dual-damascene process that connects two metal levels. It starts with the first metal already in place. A layer of a low-k dielectric is deposited and planarized with a CMP process. A patterning step creates a via hole in the dielectric layer. A second patterning step results in the lowering of the dielectric and a "step back" on the

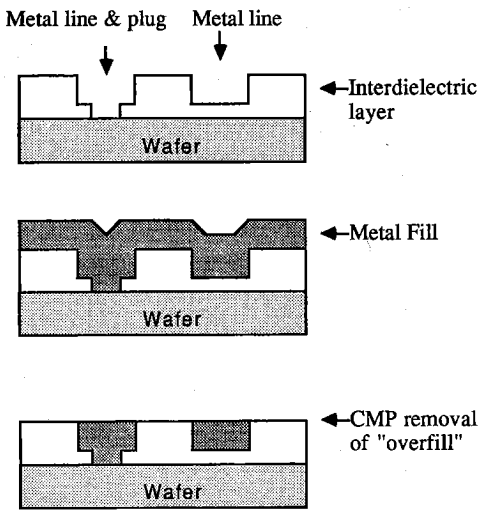


Figure 13.9 Dual-damascene (inlaid) process.

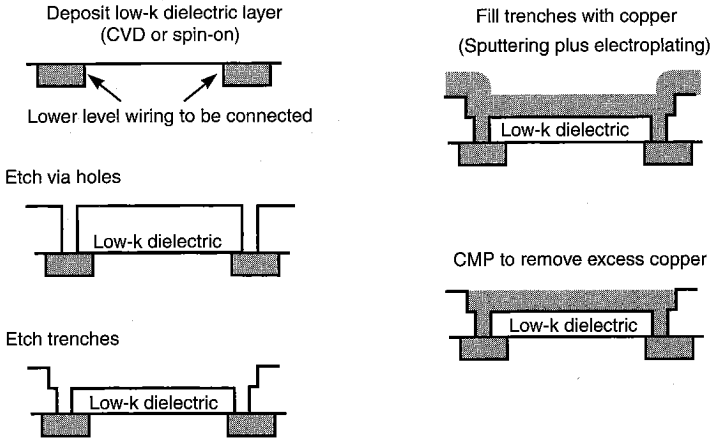


Figure 13.10 Typical dual-damascene process.

surface to allow a wider trench width. This pattern leaves a wider opening in the top laver to allow enough width for the copper stripe to carry the required current levels. This sequence offers the advantage of a one-step process to fill the via and form the copper metal lead. There are a number of variations on this basic dual-damascene process, each ending up with a narrow via and a wider trench opening ready for metal fill.

Barrier/liner deposition

As previously mentioned Copper diffuses easily through silicon dioxide and can cause electrical performance problems if it gets in the circuit components. This problem is addressed by depositing a "liner" layer in the via hole bottom and sides (Fig. 13.11). Typically, the materials used are tin (Ti) or tantalum (Ta), such as TiN, Ta, TaN, and TaSiN. Depending on the material, either sputtering or CVD deposition is used to create the barrier/liner. These vias are very high aspect and challenge the process to produce an uniform film over the entire via/trench inside surface.

Seed deposition

While copper can be deposited by sputtering, or CVD deposition, electrochemical plating (ECP) has emerged as the preferred deposition method. Producing a uniform, void-free copper film with ECP requires a starting "seed" layer in the via/trench hole. PVD techniques are used to deposit the copper seed in the via hole. The challenge, as in

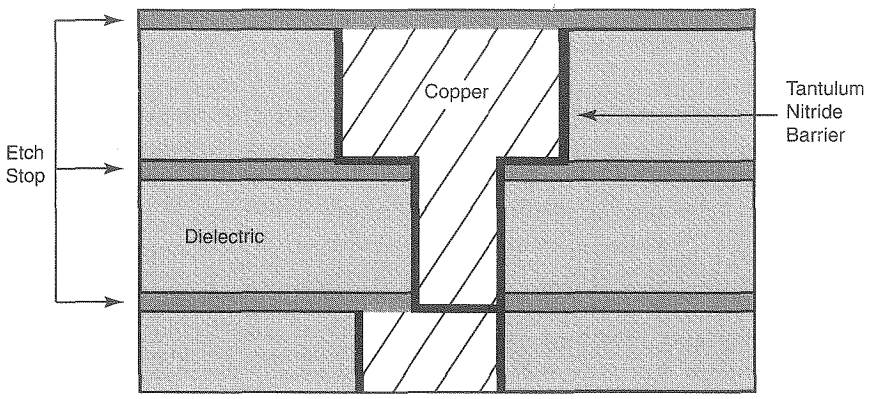


Figure 13.11 Single-level dual-damascene with tantalum nitride barrier. *(From Wolf, and Tauber,* Microchip Manufacturing, *Lattice Press.)*

the barrier/liner deposition, is producing a uniform layer in very high-aspect vias.

Electrochemical Plating (ECP)

Electroplating has emerged as a production copper deposition method due to its low temperature and low cost.[11] Low temperature is necessary due if low-k dielectric layers are used. The seed layer must uniformly coat the bottom and sides of the via/trench to ensure uniform physical and electrical properties of the copper metal lead. Electroplating of copper has been a mainstay of printed circuit board processing for decades (Fig. 13.12). The wafer is suspended in a bath

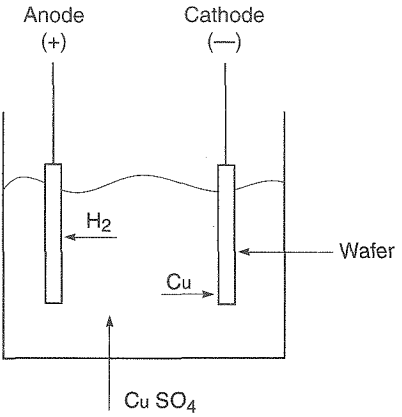


Figure 13.12 Schematic of electroplating of copper.

$$2\,Cu^{++} + 2H_2O \longrightarrow 2Cu(s) + O_2\,(G)$$

containing copper sulfate ($CuSO_4$) and is connected to an cathode (negative pole). With the application of a current, the bath components disassociate. The copper plates out on the wafer and hydrogen gas is liberated at the anode. One concern is the uniformity of the film across the wafer. The varied materials and structures on the wafer surface mitigate against uniform current distribution. Nonuniform film growth and density can be the result. Another concern is buildup of a bevel at the lip of the opening. This is addressed by a separate cleaning step after the plating. Most processes also include a post-ECP to stabilize the film prior to CMP. Nonuniform areas across the wafer surface will have different removal rates in the CMP process. Production level ECP systems will include wafer preclean, the plating sections, bevel removal, and annealing.

Chemical Mechanical Processing

Chemical mechanical processing (CMP) is used in several steps in the semiconductor process. In Chapter 3, its use for planarizing raw silicon wafers was described. In Chapter 10, we describe its use for planarizing in-process wafers to achieve a flat surface for lithography accuracy. The post-copper CMP is a similar process but with a different surface to be flattened. During copper plating, the via/trench hole is overfilled to ensure complete filling of the trench. Before proceeding to the next step, it is necessary to reflatten the surface by removing the copper overfill. The process and details are discussed in Chapter 10.

Doped polysilicon

The advent of silicon-gate MOS technology turned deposited polysilicon lines on the chip into conductors. For use as a conductor, the polysilicon has to be doped to increase its conductivity. Generally, the preferred dopant is phosphorus, due to its high solid solubility in silicon. Doping is by either diffusion, ion implantation, or *in situ* doping during an LPCVD process. Each of the methods produces a different doping result. The differences relate to the doping temperature's effect on the grain structure. The lower the temperature, the greater the amount of dopant trapped in the polygrain structure, where it is unavailable for conduction. This is the situation with ion implantation. Diffusion doping results in the lowest film sheet resistivity. *In situ* CVD doping has the lowest dopant carrier mobility due to grain boundary trapping.

Doped polysilicon has the advantage of a good ohmic contact with the wafer silicon and can be oxidized to form an insulating layer. Poly-

silicon oxides are of a lower quality than thermal oxides grown on single-crystal silicon because of the nonuniformity of the oxide grown on the rougher polysilicon surface.

Although polysilicon has a low contact resistance with silicon, it still exhibits too high a resistance the metal material(s). Creating a multimetal stack of the polysilicon and a silicide (such as titanium silicide). These are called polycide structures (see Chapter 16).

Metal Film Uses

MOS gate and capacitor electrodes

Most electrical devices depend on the passage of an electrical current to operate. Capacitors are an exception. These devices (see Chapter 16) require two conductive layers, called electrodes, separated by a dielectric. In most designs, the top electrode is a section of the conductor metal system. A discussion of the relationship of capacitor parameters is in Chapter 2.

MOS transistors are a capacitor structure, and the top electrode, called a *gate*, is a critical structure in MOS circuits.

Fuses

The development of thin-film fuse technology allowed creation of the programmable read-only memory (PROM) circuit. The fuse allows field programming of data in the memory section of the chip. In this role, the fuse is not a protective device, as in most electrical circuits, but is included specifically to be "blown" or disabled.

In the memory section of the chip, called the *array*, are a number of memory cells, each with a fuse between the cell and the main metallization system. The array is essentially a blank blackboard (Fig. 13.13). Information can be coded into the array in digital form (on/off) by having some of the cells operating and others not operating. Nonoperating cells can be created by *blowing* the fuse, thus removing it from the circuit. The same system is used to program logic arrays. (Fig. 13.14). Once the fuse is blown, the associated memory cell is permanently removed (electrically) from the circuit.

There are two primary fuse configurations. One consists of thin films of nichrome, titanium-tungsten, or polysilicon lying under two metal leads, which are patterned with thin "necks" that can be blown by a current pulse directed through the metal lines. Another fuse scheme employs a thin film of polysilicon or oxide in the contact hole. The fuse is blown when a high current is passed through the layer and destroys it by heating.

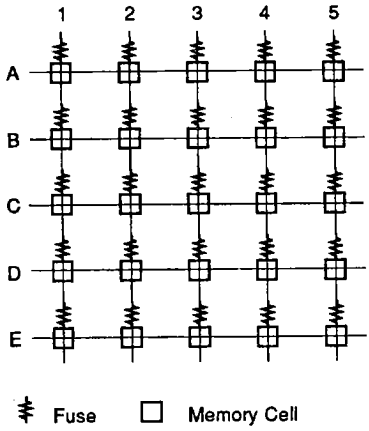


Figure 13.13 Schematic of fused memory array.

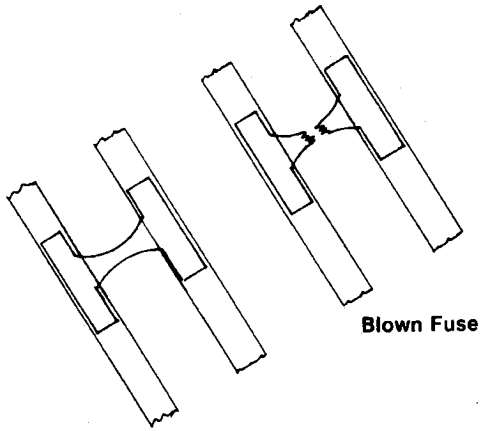


Figure 13.14 Thin film fuses.

Backside plating

Gold is sometimes evaporated onto the entire back of the wafer just prior to wafer sort. The gold functions as a solder in certain packaging processes (see Chapter 18).

Deposition Methods

Metallization techniques, like other fabrication processes, have undergone improvements and evolution in response to new circuit requirements and new materials. The mainstay of metal deposition up to the mid-1970s was vacuum evaporation. Aluminum, gold, and the fuse metals were all deposited by this technique. The needs of depositing multi-metal systems and alloys, along with the need for better step

coverage, led to the introduction of sputtering as the standard deposition technique for VLSI circuit fabrication. Refractory metal use has added the third technique, CVD, to the arsenal of the metallization engineer. Some of the basic issues of the deposition of metals, by all techniques, are discussed in the following section on vacuum evaporation.

Vacuum evaporation

Vacuum evaporation is used for the deposition of metals on discrete devices and circuits of lower integration levels. It is also used for the deposition of gold to the back of a wafer for die adhesion into a package.

Vacuum evaporation takes place inside an evacuated chamber (Fig. 13.15). The chamber can be a quartz bell jar or a stainless steel enclosure. Inside the chamber is a mechanism to evaporate the metal source, wafer holders, a shutter, thickness and rate monitors, and heaters. The chamber is connected to a vacuum pump(s) (see "Vacuum Pumps," p. 426).

Since aluminum is the most critical of the materials evaporated, we shall focus on its deposition. The vacuum is required for a number of reasons. First is a chemical consideration. If any air (oxygen) molecules were in the chamber when the high-energy aluminum atoms were coating the wafer, they would form aluminum trioxide (Al_2O_3), a dielectric that, if incorporated into the deposited film, would compromise aluminum's role as a conductor. A second requirement for vac-

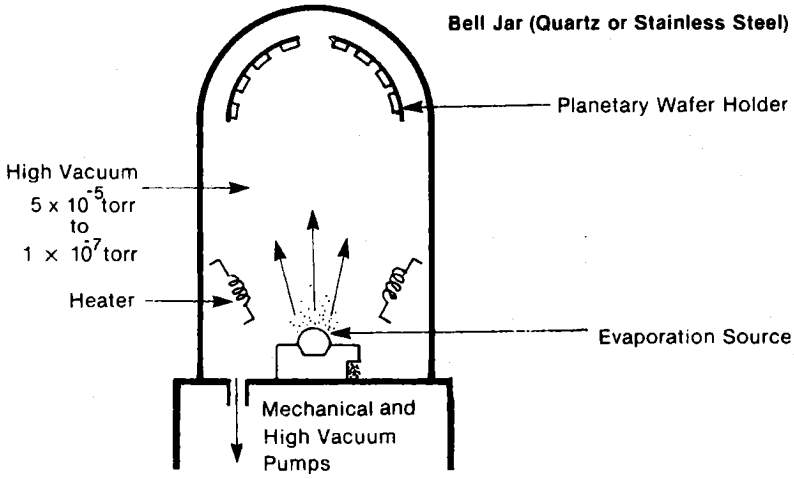


Figure 13.15 Vacuum evaporator.

uum deposition is uniform coating. The vacuum required for successful evaporation of aluminum is from 5×10^{-5} to 1×10^{-9} torr of pressure. (See Chapter 2 for a discussion on vacuum and pressure.)

There are systems that operate in the ultra-high vacuum (below 10^{-9} torr) range.[13] They offer the advantage of lowered background contamination in the system hardware, which reduces contamination of the wafers.

Evaporation sources. Before describing the various methods of causing a metal to evaporate, a review of basic evaporation theory is in order. Most of us are familiar with the evaporation of a liquid from a beaker. This happens because there is sufficient energy (heat) in the liquid to cause the molecules to escape into the atmosphere. Over time, some of them stay in the atmosphere; we call this *evaporation*. The same process of evaporation can be made to occur in solid metals. The requirement is to heat the metal to a liquid state so that the atoms or molecules evaporate into the surrounding atmosphere. Filaments and electron beam sources are the usual methods of evaporation. Filament evaporation is the simplest and is used for noncritical evaporations such as backside gold layers. The material, in wire form, is either wrapped around a coiled tungsten (or other metal that is able to withstand high temperatures) wire. A high current is passed through the tungsten wire, where it heats the deposition metal to a liquid and evaporates into the chamber, coating the wafers. Another version uses a flat filament with a dimple to hold pieces of the deposition material.

The need for evaporation control and low contamination led to the development of the electron beam evaporation source (Fig. 13.16) for aluminum. The system is called an *e-beam gun* or just *e-gun*. This evaporation source consists of a water-cooled copper crucible with a center cavity to hold the aluminum. At the side of the crucible is a

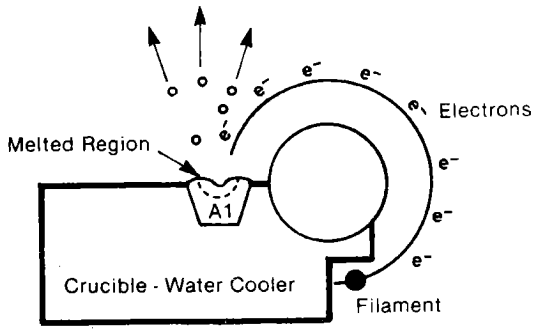


Figure 13.16 Electron gun evaporation source.

high-temperature filament. A high current is passed through the filament, which, in turn, "boils" off electrons. The negative electrons are bent 180° by a magnet so that the electron beam strikes the center of the charge in the cavity. The high-energy electrons create a pool of liquid aluminum in the center of the charge. Aluminum evaporates from the pool into the chamber and deposits on the wafers in holders at the top of the chamber. The water cooling maintains the outer edges of the charge in a solid state, thus preventing contaminants from the copper crucible from evaporating. Film thickness is controlled by shutters and by rate and thickness monitors. In-chamber monitors, located near or above the wafer holders, feed back information to the e-gun power supply, which controls the evaporation rate.

A major goal of any metal deposition system is good step coverage (Fig. 13.17). This is a challenge for vacuum evaporators, because the source is essentially a point source. The problem comes when material from a point source is shadowed by steps on the wafer surface. The result can be that one side of the openings in the surface oxide can be too thin or have a void. Planetary wafer "domes" rotating in the chamber are used to create uniform thicknesses (Fig. 13.18). Quartz heaters in the chamber aid step coverage by maintaining evaporant mobility as the atoms arrive at the surface. They "fill in" the steps by a capillary action.

Sputter deposition (PVD)

Sputter deposition (sputtering) is another old process adapted to semiconductor needs. It is a process first formulated in 1852 by Sir William



Figure 13.17 Step coverage.

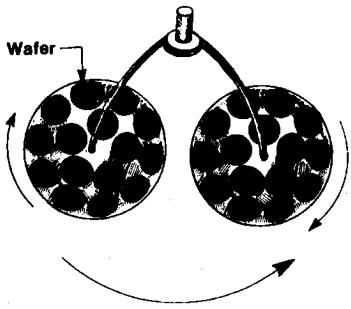


Figure 13.18 Planetary wafer holder.

Robert Grove.[12] Sputtering (in general) can deposit any material on any substrate. It is widely used to coat costume jewelry and put optical coatings on lenses and glasses. Discussion of the benefits of sputtering to the semiconductor industry is best left until the principles and methods of sputtering have been covered. Sputtering, like evaporation, takes place in a vacuum. However, it is a physical, not a chemical, process (evaporation is a chemical process) and is referred to as physical vapor deposition (PVD).

Inside the vacuum chamber is a solid slab, called a *target*, of the desired film material (Fig. 13.19). The target is electrically grounded. Argon gas is introduced into the chamber and is ionized to a positive charge. The positively charged argon atoms are attracted to the grounded target and accelerate toward it. During the acceleration, they gain momentum, which is force, and strike the target. At the target, a phenomenon called *momentum transfer* takes place. Just as a cue ball transfers its energy to the other balls on a pool table, causing them to scatter, the argon ions strike the slab of film material, causing its atoms to scatter (Fig. 13.20). The argon atoms "knock off" atoms and molecules from the target into the chamber. This is the sputtering activity. The sputtered atoms or molecules scatter in the chamber with some coming to rest on the wafer. A principal feature of a sputtering

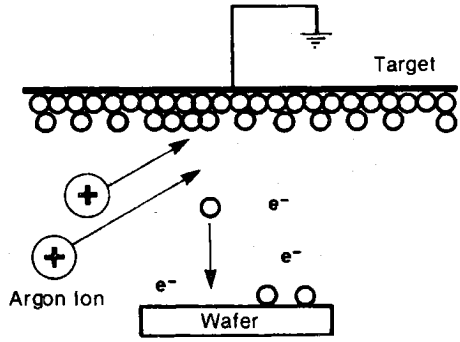


Figure 13.19 Principle of sputtering.

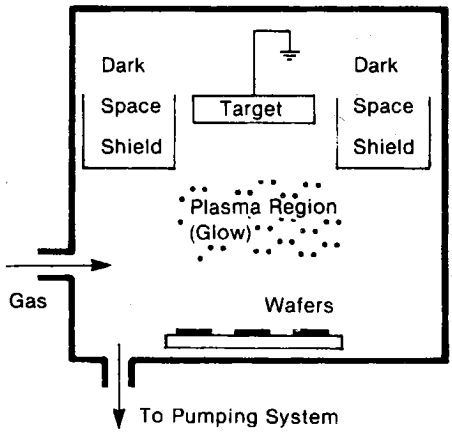


Figure 13.20 Typical sputtering equipment.

process is that the target material is deposited on the wafer without chemical or compositional change.

There are several advantages of sputtering over vacuum evaporation. One is the aforementioned conservation of target material composition. A direct benefit of this feature is the deposition of alloys and dielectrics. The problem of evaporating alloys was described in the preceding section. In sputtering, an aluminum and 2 percent copper target material yields an aluminum and 2 percent copper film on the wafers.

Step coverage is also improved with sputtering. Whereas evaporation proceeds from a point source, sputtering is a planar source. Material is being sputtered from every point on the target, with material arriving at the wafer holder with a wide range of angles to coat the wafer surface. Step coverage is further improved by rotating the wafer holder and by heating the wafer.

Adhesion of the sputtered film to the wafer surface is also improved over evaporation processes. The higher energy of the arriving atoms makes for a better adhesion, and the plasma environment inside the chamber has a "scrubbing" action of the wafer surface that enhances adhesion. Adhesion and surface cleanliness can be increased by grounding the wafer holder and sputtering the wafer surface for a brief time prior to the deposition. In this mode, the sputter system is functioning as an ion-etch (sputter-etch, reverse-sputter) machine, as described in Chapter 10.

Another technique to improve step coverage and uniform film formation in deep holes is a collimated beam (Fig. 13.21). Atoms come off of the target at many angles and tend to fill the sides of holes before filling the bottom. A collimator is a physical barrier plate similar to a honeycomb with round or hexagonal holes. It is grounded for electrical

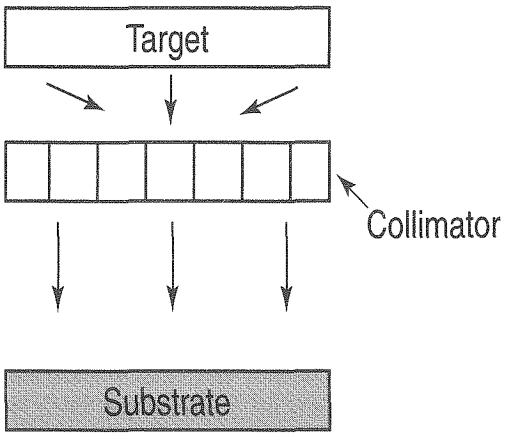


Figure 13.21 Sputtering with a collimator.

neutrality. Atoms arriving at the collimator at steep angles are caught on the sides, while straighter-angle atoms continue onto the wafer surface. The thickness of the collimator is a factor in the degree of collimation of the atom beams.

Perhaps the greatest contribution of sputtering is the control of film characteristics available by the balancing of the sputtering parameters of pressure, deposition rate, and target material. Sandwiches of material can be sputtered in one process with multiple target arrangements.

Clean and dry argon (or neon) is required to maintain film composition characteristics, and low moisture is required to prevent unwanted oxidation of the deposited film. The chamber is loaded with the wafers, and the pressure is reduced by pumps (pumped down) to the 1×10^{-9}-torr range. The argon is introduced and ionized. Control of the argon amount entering the chamber is critical due to its effect of raising the pressure in the chamber. With the argon and sputtered material in the chamber, the pressure rises to a level of about 10^{-3} torr. Chamber pressure is a critical parameter in the deposition rate of the system. After liberating material from the target, the argon ions, the sputtered material, gas atoms, and electrons generated by the sputtering process form a plasma region in front of the target. The plasma region is evident by its purple glow. The plasma region is separated from the target by a darkened region, known as the *dark space*.

Four sputtering methods are used in semiconductor applications. They are

- Diode [direct current (dc)]
- Diode [radio frequency (RDI)]

- Triode
- Magnetron

The first two methods, known as *diode sputtering*, are simple in concept. The target is connected to a negative potential with a positively charged anode present in the chamber. The negatively charged target ejects electrons, which accelerate toward the anode. Along the way, they collide with the argon gas atoms, ionizing them. The positively ionized argon atoms then accelerate to the target, initiating the sputtering process. The ionized argon (+) and the target (–) form a diode.

A secondary effect of the ionization process is the impact of the electrons on the gas atoms, resulting in the plasma region that is visible as the glowing purplish region just in front of the target. Dark spaces exist just in front and to the sides of the target. Sputtering efficiency is enhanced when the plasma is confined to the region between the target and wafers. This condition is enhanced by placing "dark space" shields to the sides of the target. The shields prevent target material from being sputtered from the sides—material that will never deposit on the wafers.

Another problem arises from the outgassing of contamination from the chamber walls while the chamber is under vacuum. This condition is called a *virtual leak*, as opposed to an actual leak of atmosphere into the system. Besides compromising the pressure level in the chamber, the contamination can be incorporated into the deposited film.

This latter problem is addressed by placing a small negative bias (charge) on the wafer holder. The bias creates ions on the wafer surface and has the effect of dislodging stray outgassed atoms from the growing film. Direct-current diode sputtering is used primarily to deposit metals.

Improved sputtering is gained by connecting the target to the negative side of a radio-frequency (RF) generator. The ionization of the gas takes place near the target surface without requiring a conductive target. Radio-frequency sputtering is necessary to sputter nonconducting materials (dielectrics) and is used also for conductors. Biasing is also used with the radio-frequency sputtering to achieve a cleaning effect at the wafer surface. Radio-frequency biasing offers the advantage of etching and cleaning of the exposed wafer surface. Etching and cleaning are achieved by putting the wafer holder at a different field potential from that of the argon, causing the argon atoms to impinge directly on the wafer. This procedure is called *sputter etch*, *reverse sputter*, or *ion milling*. The process removes contamination and a small layer from the wafer. Removal of contamination improves electrical contact between the exposed wafer regions and the film and improves adhesion of the film to the rest of the wafer surface.

In diode sputtering, a number of processes occur at or near the wafer surface. On impact of the argon atom, a number of electrons are created. These electrons cause heating of the substrate (as high as 350°C), which in turn can cause uneven film deposition. The electrons also create a radiation environment that can damage sensitive devices.

The heating produced with diode sputtering causes a serious problem with the deposition of aluminum. The heating causes residual oxygen in the target and in the chamber to combine with the aluminum to form aluminum oxide. The aluminum oxide is a dielectric and can compromise the conductive property of the deposited aluminum. More serious, a layer of aluminum oxide can form on the target surface, and the impinging argon atoms do not (in diode sputtering) have enough energy to break through the layer. In effect, the target becomes sealed, and the sputtering stops. Triode sputtering avoids some of the problems of diode sputtering.

The electrons necessary to ionize the argon are created by a separate high-current filament. In designs where the filament is located outside the deposition chamber, the wafers are protected from radiation damage. Films deposited by a triode method are more dense.

Another problem with diode sputtering is the electrons that escape into the chamber and do not contribute to the establishment of the plasma necessary for deposition. The situation is resolved in *magnetron* sputtering systems, which use magnets behind and around the target (Fig. 13.22). The magnets capture and/or confine the electrons to the front of the target. Magnetron systems are more efficient for increased deposition rates. The resulting ion current (density of ionized argon atoms hitting the target) is increased by an order of magnitude over conventional diode sputtering systems. Another effect is a lower pressure required in the chamber, which contributes to a cleaner deposited film. Magnetron sputtering leaves a lower target temperature,

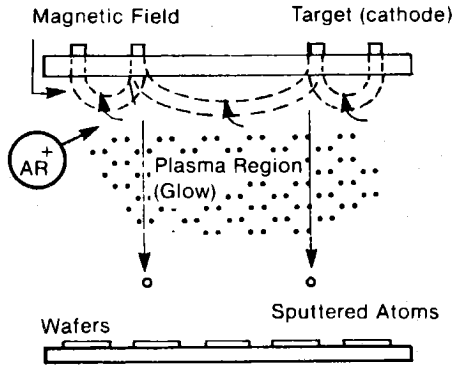


Figure 13.22 Magnetron sputtering.

which makes it a favorite for the sputtering of aluminum and aluminum alloys.

Production-level sputtering systems come in a variety of designs. Chambers are either batch systems or single-wafer in-line designs. Most production machines have load-lock capabilities. A load lock is an antechamber where a partial vacuum is created so that the deposition chamber can be maintained at vacuum. The advantage of a load lock is a higher production rate. Production machines are usually dedicated to one or two target materials, while development machines have a wider range of capability.

CVD metallization

Advances in LPCVD offer the deposition department a third choice for metal depositions. LPCVD offers the advantages of not requiring expensive and maintenance-intensive high-vacuum pumps, conformal step coverage, and high production rates. Perhaps the most often deposited CVD refractory metal film is tungsten (W).

Tungsten is used in a variety of structures, including contact barriers, MOS gate interconnects, and via plugs. The filling of via holes is a key to effective multi-metal systems (see Fig. 13.2). The dielectric layer is relatively thick and the via holes have to be relatively thin (high aspect ratio). These two factors make for a difficult continuous metal deposition to fill the vias without thinning of the metal in the via. Selective CVD-deposited tungsten plugs fill the entire via and present a planar surface for a subsequent conducting metal layer. For use as a barrier metal, tungsten can be deposited selectively by the silicon reduction of the gas tungsten hexafluoride (WF_6) by the reaction

$$2WF_6 + 3Si \rightarrow 2W + 3SiF_4$$

Tungsten can also be deposited selectively over aluminum and other materials from WF_6. The processes are called *substrate reduction.* Tungsten is also deposited from WF_6 by hydrogen reduction; the reaction is

$$WF_6 + 3H_2 \rightarrow W + 6HF$$

All of the depositions are performed in LPCVD systems at temperatures of about 300°C, which makes the processes compatible with aluminum metallization.

The depositions of tungsten silicide and titanium silicide proceed by the following reactions:

$$WF_6 + 2SiH_4 \rightarrow WSi_2 + 6HF + H_2$$

$$TiCl_4 + 2SiH_4 \rightarrow TiSi_2 + 4HCl + 2H_2$$

Vacuum Pumps

LPCVD, ion implantation, evaporation, and sputtering processes all take place in reduced-pressure (vacuum) chambers (see Chapter 2). Vacuum chambers provide process conditions free of contaminating gases. In the deposition processes, the vacuum increases the mean free path of the depositing atoms and molecules, which in turn results in more uniform and controllable deposited films. LPCVD takes place in the pressure range down to 10^{-3} torr (medium range) while the other processes take place at pressure ranges down to 10^{-9} torr high to ultra-high range). Medium range is reached with mechanical vacuum pumps. These same pumps are used to initially reduce the pressure in the high-vacuum process chambers. In this role, they are called *roughing pumps*. Additionally, mechanical vacuum pumps are used on the outlet end of high-vacuum pumping systems to assist in the removal of gas molecules from the pump to the exhaust system.

After the rough vacuum is established, a high-vacuum pump takes over to establish the final vacuum. The pumps used for this purpose are *oil diffusion*, *cryogenic*, *ion*, or *turbomolecular*. Whatever the type, all pumps are constructed of materials that will not *outgas* into the system and compromise the vacuum. Materials used are typically type 304 stainless steel, oxygen-free high-conductivity copper (OFHC), Kovar, nickel, titanium, borosilicate glasses, ceramics, tungsten, gold, and some low-vapor-pressure elastomers.[13] Pumps used to evacuate corrosive and toxic gases or reaction by-products must have corrosion-free inside surfaces. Also, care must be taken in servicing pumps with these types of applications.

Pumps are selected and used based on a number of criteria, including

- Vacuum range required
- Gases to be pumped (lighter gases such as hydrogen are more difficult to pump)
- Pumping speed
- Overall throughput
- Ability to handle impulsive loads (periodic outgassing)
- Ability to pump corrosive gases

- Service and maintenance requirements
- Downtime
- Cost

Recall from Chapter 2 that pressure in a system results from the activity of gas atoms or molecules in an enclosure striking the chamber walls with some force. Reduction of the pressure in a system requires the removal of the gas in the chamber. This is generally accomplished by the pump establishing a lower pressure, first within itself, which allows gas material in the process chamber to flow into the pump, where it is removed entirely from the system. At very low pressures, there is not much material in the chamber, and continued pressure reduction requires that the system be leak-free and not add to the pressure by its own outgassing. Some systems require traps to prevent material from the pump from *backstreaming* into the chamber. Cold traps are explained in the section on oil diffusion pumps.

Mechanical pumps

Mechanical oil rotary vacuum pumps are a basic design that traces to the 1640s, when Galileo and Torricelli were investigating the theory that air had weight. They are referred to as *blower*-type pumps. Air is removed from the pump by an eccentric rotating vane in a cavity (Fig. 13.23). As the vane rotates, it compresses and sweeps out the gas in front of it in the cavity, simultaneously leaving a reduced-pressure region behind it. The "pushed" material exits through a valve, while another valve opens to the cavity from the chamber and allows material in the chamber to flow into the cavity. As the vane rotates, more and more material is removed from the cavity, thus reducing the pressure in the chamber.

A critical part of the pump is the efficiency of the exit valve. An ineffective valve that leaks atmosphere back into the system will limit the

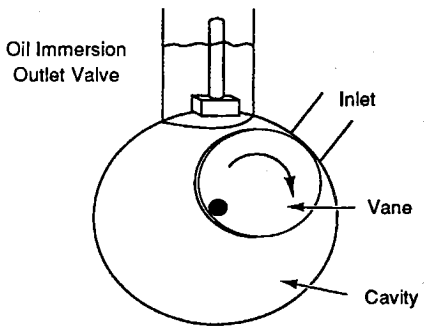


Figure 13.23 Mechanical rotary oil pump.

ultimate pressure level of the pump. This style of pump uses an oil-immersed valve that prevents backstreaming of the atmosphere.

Oil-filled pumps met their match with the advent of LPCVD and etch applications. These processes produce a lot of physical residues that contaminate the oil quickly. *Oil-sealed, oilless, scroll pumps,* and *roots pumps* (called *roots blowers*) are all push-type pumps that have replaced oil-mechanical pumps.[16]

Oil diffusion pumps

The oil diffusion pump (Fig. 13.24) was the mainstay of most semiconductor vacuum processes. The pump requires the services of an oil mechanical pump to first reduce the pressure in the chamber to the 10^{-3}-torr level. Either the same or a second mechanical pump is required at the outlet end. High vacuum, in the 10^{-8}-torr range (under production conditions), is achieved by a clever momentum transfer system. A low-vapor-pressure, hydrocarbon-based oil is heated in the base of the pump, where it rises up a structure called the *stack.* At the top of the stack is a series of downward-facing baffles. The hot oil molecules, which have gained speed and energy from the boiling, exit the stack in a downward direction. Outside the stack, they collide with gas from the chamber, causing them to be propelled toward the bottom of

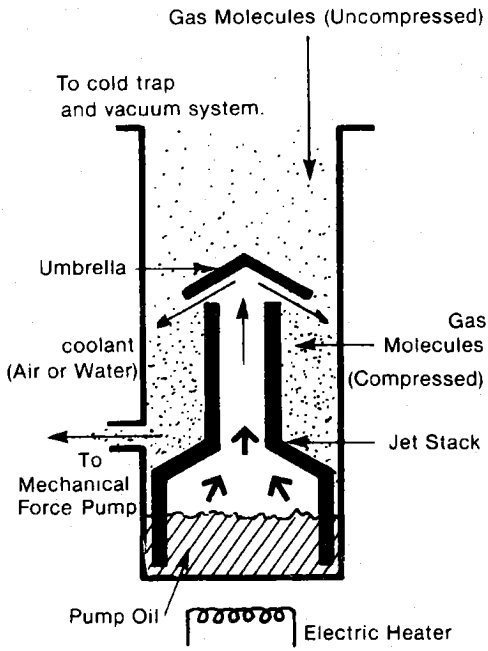


Figure 13.24 Oil diffusion pump.

the pump, where they are removed by a mechanical pump. The oil molecules return to the heated reservoir.

Two problems with oil diffusion pumps are the migration of oil molecules back up into the chamber and the inability of the pump to handle water vapor from the chamber. Both problems are resolved by the use of a *cold trap* between the pump and the chamber. A cold trap is similar in design to a liquid source bubbler. The fluid in the trap is liquid nitrogen, which reduces the temperature to –96°C. At this temperature, oil, contaminant and water vapor molecules are frozen to the inside walls of the trap and do not add to the pressure in the system.

Cryogenic pumps

Even with cold-trap technology, some processes cannot stand contamination from hydrocarbon oils such as are used in oil diffusion pumps. This situation has led to the use of cryogenic (cryo) pumps. A cryogenic pump (Fig. 13.25) uses the fact that gas molecules will "freeze" out on cold surfaces. The cold trap described in the preceding section and the frost that collects on the insides of refrigerators are examples of cryogenic activity.

Cryogenic pumps are designed with a central finned stack. At the low temperatures, gas from the chamber collects on the vanes, removing material from the system, which in turn reduces the pressure. The central stack, called an *expander*, is cooled as a compressor releases the liquid helium or nitrogen into it from the bottom. Cooling is by a phenomenon called *adiabatic expansion*. This is the same phenomenon that causes a pressurized can to cool when the nozzle is opened and the gas expands rapidly into the lower-pressure atmosphere. The top of the expander is at a higher temperature than the bottom, resulting in different gas molecules freezing out at different levels. Cryogenic pumps can efficiently pump water vapor, which turbo pumps cannot.

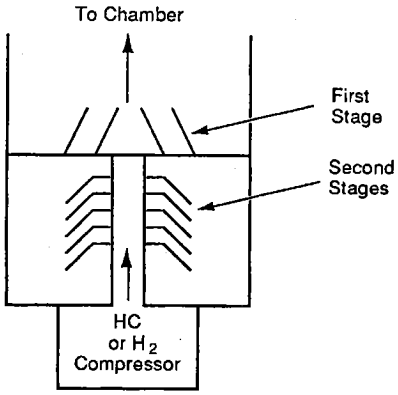


Figure 13.25 Cryogenic pump.

Cryopumps operate without cold traps or mechanical roughing pumps. Due to their action of trapping gas material on the fins, they are of a type known as *capture pumps* rather than the displacement-type pumps that move the chamber material to the atmosphere. The total capture nature of the pump and its oilless operation drastically reduce the possibility of contamination. However, when the pump is brought to room temperature, by mistake or for maintenance, the frozen gases are released, and care must be exercised to vent any toxic or flammable gases trapped on the vanes. The buildup of gases also affects the pumping speed. The pump's speed must be monitored and the system cleaned when the speed falls. Cryopumps are simpler to operate and maintain, since no cold trap liquids or messy oils are needed. Additionally, cryopumps can handle bursts of outgassing from the process chamber and feature a fast pumping speed.

Ion pumps

Another capture-type pump is the ion pump, also known as a *sputter ion pump* or a *getter pump* (Fig. 13.26). An ion pump operates in a manner similar to an ionization section in an ion implanter or sputter machine; only in this application, the atoms and molecules come from the chamber. A portion of those that drift into the ionization chamber are ionized to a positive charge by bombardment with electrons and attracted to a titanium cathode (negative potential). On collision with the titanium, some of the titanium is sputtered away and travels into the pump. The titanium atoms are chemically active enough to combine with other gases in the pump, which also accumulate on the pump walls. Again, material is removed from the chamber which reduces the pressure. Ion pumps are capable of pressures down to 10^{-11} torr, which is the ultra-high vacuum range.

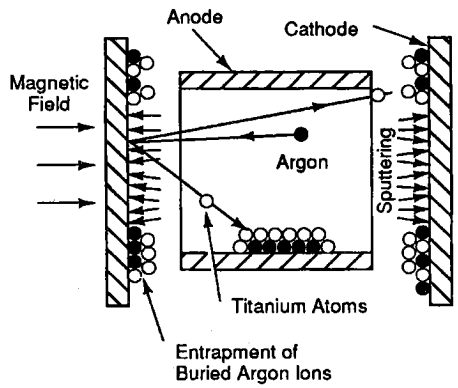


Figure 13.26 Principle of ion vacuum pump.

Turbomolecular pumps

Turbomolecular pumps are similar in design to a jet turbine engine. A series of blades (Fig. 13.27) with openings are mounted and rotated at very high speeds (24,000 to 36,00 rpm[14]) on a central shaft. Gas molecules from the chamber encounter the first blade and gain momentum from the collision with the rotating blade. The momentum direction is downward to the next blade, where the same thing happens. The net result is a removal of gas from the chamber. The use of a momentum transfer makes the pumping principle the same as an oil diffusion pump. Major advantages of turbomolecular pumps are a lack of backstreaming from oils, no need to recharge, high reliability, and pressure reduction into the high vacuum range. Drawbacks are a slower pumping speed compared to oil diffusion and cryogenic pumps and vibration and wear due to the high rotational speeds.

An addition to turbo pumps is a *drag type* pump. Molecules are bounced off a rotating drum or disk rather than vanes or stators.[15] These *combination* pumps can exhaust at high pressures. Use of turbo pumps with corrosive gas processes requires coating the rotors and staters and/or heating the pump to keep the gases from forming solid particles that deposit on the pump parts.

Summary

Figure 13.28 is an overview of the metals most used, their uses, and deposition methods.

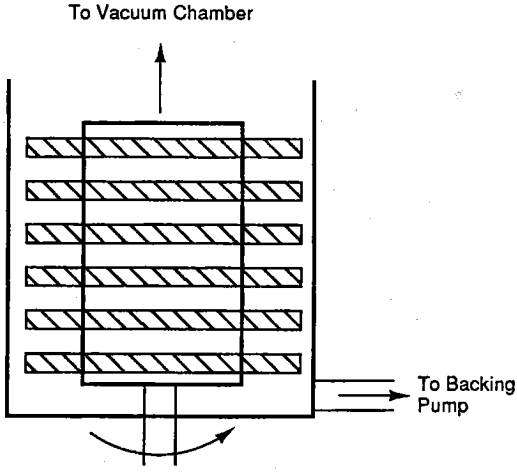


Figure 13.27 Turbomolecular pump.

	Use in Device				Deposition Method			
Metal	Conductor	Fuse	Barrier	Plug	Other	Evaporation	Sputtering	CVD	Plating
Aluminum	X					X	X		
Aluminum/Silicon	X					X	X		
Aluminum/Copper	X					X	X		
Aluminum/Silicon/Copper	X					X	X		
Copper	X			X				X	X
Titanium									
Tungston									
Titanium/Tungston		X							
Nichrome		X							
Doped Polysilicon	X						X	X	
Molybdenum			X				X	X	
Platinum Silicide			X	X			X	X	
Titanium Silicide			X	X			X	X	
Tungsten Silicide			X	X			X	X	
Molybdenum Silicide			X	X			X	X	
Gold					Die Attach Life Time- Control				

Figure 13.28 Summary of metallization materials and processes.

Review Questions

1. Discuss the properties required for a microchip metal system and how they apply to aluminum and copper metallization.
2. Explain electromigration and how it is controlled.
3. Explain aluminum and silicon eutectic alloying and two processes used to prevent it.
4. Why are refractory metal and their silicides favored for VLSI circuit metallization?
5. What is a silicide and how is it formed?
6. Draw a diagram of a two-metal system, identifying all parts.
7. Why are low-k dielectrics used with copper metallization systems?
8. Draw a diagram of and label the parts of a sputter machine.
9. What are the advantages of a sputter process compared to an evaporator process?
10. List the four types of high-vacuum pumps used in semiconductor processes and their principles of operation.
11. Discuss the RC constant effect on circuits and techniques to reduce it.
12. List three advantages of copper metallization over aluminum.
13. Sketch the steps in a dual-damascene process.

References

1. Staff, "Speeding the Transition to 0.18 μm," *Semiconductor International*, January 1998, p. 66.
2. S. Wolf and R. Tauber, *Silicon Processing for the VLSI Era,* Lattice Press, Sunset Beach, CA, 1986, p. 332.
3. P. Riley, S. Peng, and L. Fang, "Plasma Etching of Aluminum for ULSI Circuits," *Solid State Technology,* PennWell Publishing, February 1993, p. 47.
4. S. M. Sze, *VLSI Technology,* McGraw-Hill, New York, 1983, p. 347.
5. P. Singer, "New Interconnect Materials: Chasing the Promise of Faster Chips," *Semiconductor International*, November 1994, p. 53.
6. P. Singer, "Copper Goes Mainstream: Low-k to Follow," *Semiconductor International*, November 1997, p. 67.
7. P. Singer, "New Interconnect Materials: Chasing the Promise of Faster Chips," *Semiconductor International*, November 1994, p. 54.
8. D. M. Brown, "CMOS Contacts and Interconnects," *Semiconductor International*, 1988, p. 110.
9. D. Pramanikm and V. Jain, "Barrier Metals for ULSI," *Solid State Technology,* PennWell Publishing, January 1993, p. 73.
10. P. Singer, "Copper Goes Mainstream: Low-k to Follow," *Semiconductor International*, November 1997, p. 68.
11. A. Braun, ECP Technology, *Semiconductor Technology,* May 2000, p. 60

12. Y. Pauleau, "Interconnect Materials for VLSI Circuits," *Solid State Technology*, February 1987, p. 61.
13. G. Tisdale, et al., "Next-Generation Aluminum Vacuum Systems," *Solid State Technology*, May 1998, p. 79.
14. A. J. Aronson, "Fundamentals of Sputtering," *Microelectronics Manufacturing and Testing*, January 1987, p. 22.
15. J. Ballingall, "State-of-the-art Vacuum Technology," *Microelectronics Manufacturing and Testing*, October 1987, p. 1.
16. S. Wolf and R. Tauber, *Silicon Processing for the VLSI Era*, Lattice Press, Sunset Beach, CA, 1986, p. 95.
17. P. Singer, "Vacuum Pump Technology Leaps Ahead," *Semiconductor International*, Cahners Publishing, September 1993, p. 53.

Chapter

14

Process and Device Evaluation

Overview

The wafer-fabrication process requires a high degree of precision in process control, equipment operation, and material manufacture. One process mistake can render the wafer completely useless. One "killer" defect can ruin a die. Throughout the process, a variety of tests and measurements are made to determine both wafer quality and process performance. The tests take place on in-process wafers, test die and production die, and the finished circuit. Individual tests are described in this chapter. Statistical process control programs are addressed in Chapter 15.

Objectives

Upon completion of this chapter, you should be able to:

1. Explain the difference between resistance, resistivity, and sheet resistance.
2. Draw a sketch of the parts and current flow in a four-point probe.
3. Compare the principles and uses of color interference, fringe counting, spectrophotometers, ellipsometers, and stylus for film thickness measurements.
4. Compare the principles and uses of groove and stain, SEM, and spreading resistance for junction depth measurements.
5. List the methods and advantages of microscope and SEM inspection of wafer surfaces.
6. Draw sketches of diodes in forward and reverse bias and their companion current-voltage curves.

7. Explain the effect of surface current leakage on a junction performance characteristic.
8. Draw sketches of a bipolar and MOS transistor in operation and their companion current-voltage characteristics.
9. List the process steps for a capacitance-voltage measurement and the principle of contamination detection.
10. Describe the principle and use of atomic force microscopes.

Introduction

Characterization of processes and circuit parameters is required for production-line control and product stability. *Metrology* is the general term applied to the measurement of physical surface feature. Interests include pattern widths, film depths, defect identification and location, and pattern registration errors. Good characterization can warn of a process that is about to go out of control, and device characterization is essential to analyze circuit performance and conformance to customer specifications. Consequently, every process step has a rigid set of equipment and process parameters that are controlled (temperature, time, and so forth). After every significant process step, there is an evaluation of the result on the wafer or a test wafer. *Test wafers* are blank wafers or wafer pieces that are included in the process step for post-process measurements. Many of the tests are destructive and cannot be performed on the device wafers or cannot be performed on the actual components in the chip. In the process chapters, the important parameters for each process were identified (e.g., film thickness, resistivity, cleanliness). Here, the basic theory, applicability, and range of sensitivity of the test methods are examined.

Some are direct measurements, and some are indirect. One group includes electrical measurements of test wafers and on the actual devices. They measure the direct effect of some of the processes, such as ion implantation. Device performance measurements are usually inclusive of several processes, and the results are used to infer individual process parameter control. Another group directly measures physical parameters such as layer thicknesses and widths, composition, and others. This group includes defect detection. A third group measures contamination in and on the wafer and in materials.

Not surprisingly, test and measurement methods have changed along with the levels of integration and smaller image sizes. ULSI technology is ushering in nanometer- and angstrom-level inquiry, called the *nanoanalysis era.*[1] And the price of in-line testing is going up. Larger wafers and more dense circuits require more tests to prop-

erly characterize processes. High-volume processing requires real-time testing and analysis to guard against scrapping volumes of production wafers. Data management systems for ULSI circuits usually include on-board statistical analysis and data base management capabilities.

Wafer Electrical Measurements

Resistance and resistivity

The object of the fabrication process is to form, in and on the wafer surface, solid-state electrical components (transistors, diodes, capacitors, and resistors) that are wired together to form the circuit. Each of the components must meet individual electrical performance specifications if the entire circuit is to function. Throughout the process, electrical measurements are performed to judge the process and predict electrical device performance.

Resistivity measurements

The addition of dopants to the wafer, both during crystal growth and during the doping processes, alters the electrical characteristics of the wafer. The altered parameter is its resistivity, which is a measure of a material's specific "resistance" to the flow of electrons (Fig. 14.1). Whereas the resistivity of a given material is a constant, the resistance of a specific volume of the same material is a function of its dimensions and resistivity. This relationship parallels that of density and weight. For example, the density of steel is a constant, whereas the weight of a particular piece depends on its volume.

The units of resistance (R) are ohms (Ω) and the units of resistivity (ρ) are ohm-centimeters (Ω-cm). Because adding dopants to a wafer will alter its resistivity, measurement of resistivity is actually an indirect measure of the amount of dopants added.

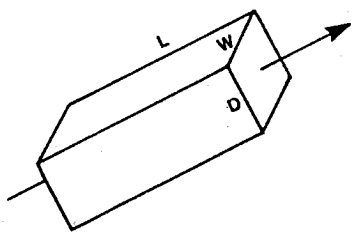


Figure 14.1 Relationship of resistance to resistivity and dimensions.

$$R = \rho \frac{L}{A} = \rho \frac{L}{WXD}$$

Four-point probe

The parameters of resistance, voltage, and current are governed by Ohm's law. The three parameters are related mathematically in the following way:

$$R = V/I = (\rho)L/A = (\rho)L/(W \times D)$$

where R = resistance
V = voltage
I = current
ρ = resistivity of sample
L = length of sample
A = cross-sectional area of sample
W = width of sample
D = depth of sample

Theoretically, the resistivity of a wafer can be measured with a multimeter (Fig. 14.2) by measuring the voltage at a constant current through a sample of known dimensions and calculating the resistivity. However, the resistance between the probes and the wafer material is too great to accurately measure the resistivity of semiconductors with their relatively low quantity of dopants.

The four-point probe is the instrument used to measure resistivity on wafers and crystals. It employs four thin, in-line probes connected to a power supply and voltmeter. The four-point probe consists of four thin metal probes arrayed in a line. The two outside probes are connected to a power supply, and the inside probes are connected to a voltage meter. During operation, the current passes between the outer probes, and the voltage drop (change) is measured between the inner

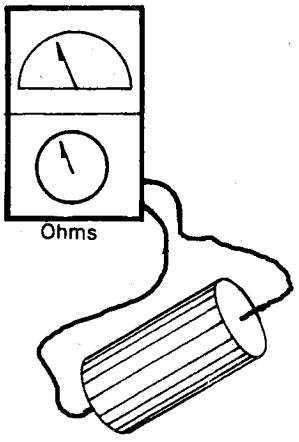


Figure 14.2 Multimeter.

probes (Fig. 14.3). The relationship of the current and voltage values is dependent on the resistance of the space between the probes and the resistivity of the material. The four-point probe cancels out the effects of probe-wafer contact resistance on the measurement.

Process and Device Evaluation

Using a four-point probe, the voltage and current are related to the resistivity by the following relationship:

$$\rho = 2\pi s V/I$$

where s is the distance between probes when s is less than the wafer diameter and less than the film thickness.

Sheet resistance

The four-point probe measurement just described is used to measure the resistivity of wafers and crystals. It is also used to measure the resistivity of thin layers of dopants added into the wafer surface by the dopant processes. When a four-point probe measurement is made on a thin layer of added dopants, the current is confined in the layer (Fig. 14.3). A thin layer is defined as a layer thinner than the probe spacing (distance between probes).

The electrical quantity measured on a thin layer is called *sheet resistance,* R_s. This quantity has the units of ohms per square (Ω/❑). The concept of ohms per square can be understood by considering the resistance of two squares of the same thin material of equal thickness (Fig. 14.4). Since the resistivity of ρ is the same for each piece, and $T_1 = T_2$, the sheet resistance is the same for each piece. Or, the resistance of the thin sheet is a constant for any square of the same material.

The formula relating sheet resistance to the voltage and current is

$$R_s = 4.53 V/I$$

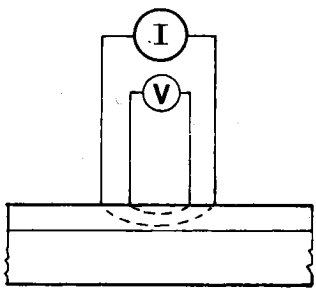


Figure 14.3 Four-point probe measurement of a thin layer.

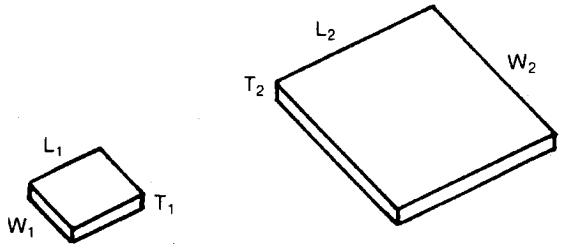


Figure 14.4 Resistance of a "square."

where 4.53 is a constant that arises from the probe spacing. Some companies elect to drop the constant 4.53 from the formula and just measure the V/I of a wafer as in Fig. 14.5.

Four-point probe thickness measurement

The thickness of uniform conducting layers on an insulating layer can be determined using a four-point probe. For thin films, the formula is

$$T = \rho_s/R$$

where T = layer thickness
ρ_s = resistivity
R_s = sheet resistance

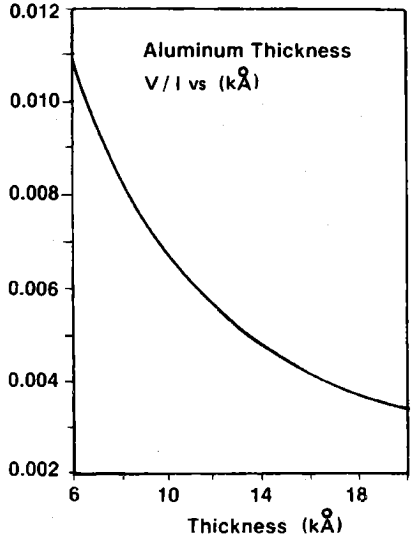


Figure 14.5 Voltage/current (V/I) versus thickness of aluminum.

Since the resistivity is a constant for pure materials such as aluminum (Fig. 14.5), the sheet resistance measurement is actually a measurement of the film thickness. This formula does not calculate the thickness of a doped layer, since the dopants are not evenly distributed throughout the layer.

Concentration/depth profile

The distribution of dopant atoms in the wafer (see Chapter 11) is a major influence on the electrical operation of a device. The distribution (or dopant concentration profile) is determined by several techniques. One is spreading resistance. After doping, a test wafer sample is prepared by the bevel technique. After the junction is exposed by the beveling, a series of electrical two-point probe measurements are made sequentially down the bevel (Fig. 14.6). At each point, the vertical drop of the probes is recorded and a resistance measurement made. The resistance value at each point changes with the change in dopants at each level.

A computer is used to perform calculations that relate the depth and resistance values to the dopant concentration at each level. The computer uses the data to construct a dopant concentration profile for the sample. This measurement is usually made periodically off-line or when electrical device performance indicates that the dopant distribution may have changed.

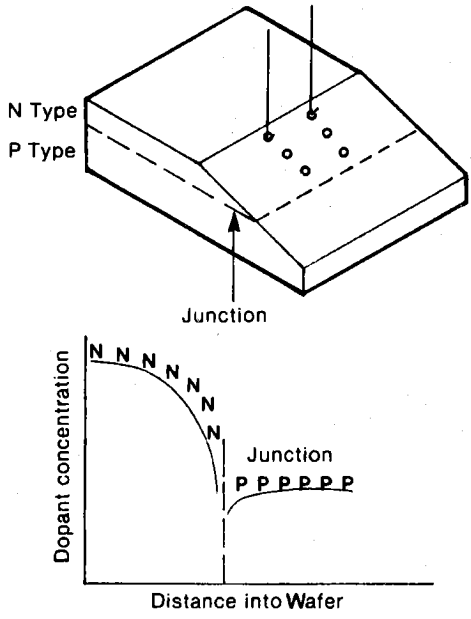


Figure 14.6 Spreading resistance.

Another method is to incrementally remove thin layers from the wafer surface with the anodic oxidation (see Chapter 7) technique. After the oxide is grown, it is removed by etching, and a four-point probe measurement is made on the new surface. The distance down into the wafer is a function of the oxide thickness and is related to the dopant concentration four-point probe sheet resistance values by a computer program.

Secondary ion mass spectrometry (SIMS)

SIMS is essentially a combination ion milling and secondary ion detection method. Ions are directed at the sample surface, removing a thin layer. Secondary ions are generated from the removed material, which contains the wafers material and the dopant atoms. The ions are collected and analyzed providing a calculation of the amount of dopant at each level, which in turn allows the construction of a dopant profile.[2]

Differential Hall effect (DHE)

Like the other two profiling methods, DHE requires sequential removal of the doped layer, down to the junction. As each layer is removed, the resistivity and the Hall coefficient (which relates to carrier mobility) are measured. Dopant concentration for each layer is calculated from the two measured parameters. A drawback to this method is the additional time to prepare a specially shaped test structure (a van der Pauw structure).[3]

Oxide rupture (BV_{ox} or rupture voltage)

An electrical measurement, BV_{ox}, is used as a measure of oxide quality. The test structure used is the same as for capacitance-voltage (C/V) analysis. But, in this case, the voltage is continually increased until the oxide is physically destroyed and current flows freely from the aluminum dot to the silicon. The maximum voltage that the oxide can withstand before breakdown is a function of its thickness, structural quality, and purity. Perhaps the most critical oxide layer is in the gate of a MOS transistor. This test is often called *gate oxide integrity* (GOI). With device geometries getting smaller and oxides thinner, high quality is essential for device operation.

Physical Measurement Methods

Product reliability and maintenance of yields requires on-line detection of defects, mistakes, and so forth to allow removal from the line of suspect material. Process control requires measuring the process re-

sults at each of the process steps and knowing the quantity, density, location, and nature of the various problems. This data comes from a series of measurements and evaluations that vary with the sophistication of the circuit in terms of image size, sensitivity to contamination, and density.

Tests are performed on test wafers or directly on product wafers. Product wafer tests may require revealing the subsurface structure by beveling, microsectioning, or using a focused ion beam (FIB) to remove portions of the circuit.

Layer Thickness Measurements

Color

Both silicon dioxide and silicon nitride layers exhibit different colors on the wafer. We know that, while silicon dioxide is transparent (glass is silicon dioxide), an oxidized wafer has a color. The color seen is actually the result of an interference phenomenon—the same phenomenon that creates the colors of rainbows.

The silicon dioxide layer on a silicon wafer is actually a thin transparent film on a reflecting substrate. Some of the light rays impinging on the wafer surface reflect off the oxide surface, while others pass through the transparent oxide and reflect off the mirrored wafer surface (Fig. 14.7). When the light rays exit the film, they combine with the surface-reflected ray, giving the surface an appearance of having a color. This phenomenon is the reason oxidized wafers change color as the angle of viewing is changed.

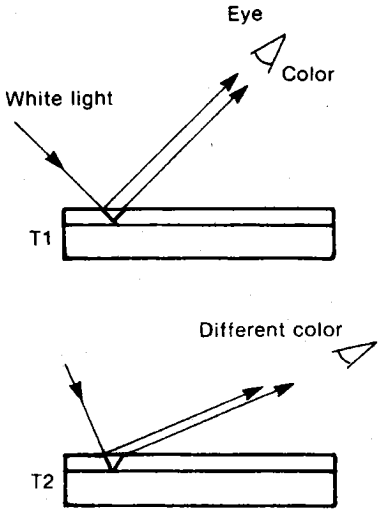


Figure 14.7 White light interference.

The exact color is a function of three factors. One, which is a property of the transparent film material, is the *index of refraction.* A second factor is the *viewing angle.* The third factor is the *thickness of the film.* The color of a thin transparent film becomes an indication of the thickness when the nature of the viewing light is specified (i.e., daylight, fluorescent), along with the viewing angle. The classic color versus thickness chart (Fig. 14.8) is a regular feature at oxidation and diffusion stations. Color alone is not an exact indication of thickness because of the consequences of the interference phenomenon.

As the film gets thicker, the colors change in a specific sequence and then repeat themselves. Each repetition of the color is called an *order.* To determine the exact film thickness, a knowledge of the color order is necessary. A principal use of color charts is for process control.

Film thickness, μm	Color* and Comments
	Order I
0.050	Tan
0.075	Brown
0.100	Dark Violet to Red Violet
0.125	Royal Blue
0.150	Light Blue to Metallic Blue
0.175	Metallic to Very Light Yellow
0.200	Light Gold or Yellow-Slightly Metallic
0.225	Gold with Slight Yellow Orange
0.250	Orange to Melon
0.275	Red Violet
0.300	Blue to Violet Blue
0.310	Blue
0.325	Blue to Blue Green
0.345	Light Green
0.50	Green to Yellow Green
	Order II
0.365	Yellow Green
0.375	Green Yellow
0.390	Yellow
0.412	Light Orange
0.426	Carnation Pink
0.443	Violet Red
0.465	Red Violet
0.476	Violet
0.480	Blue Violet
0.493	Blue
	Order III
0.502	Blue Green
0.520	Green (Broad)

*When viewed perpendicularly in daylight fluorescent light

Figure 14.8 Silicon dioxide thickness color chart.

Each oxidation or silicon nitride process is set up to produce a specified thickness. Naturally, the thickness will vary from run to run. Operators quickly become sensitive to the wafer color. When a variation occurs, a quick check of the chart will indicate if the film thickness is out of specification. Rarely is a process so far off that the film thickness is a whole order (same color, different thickness) out of specification. The accuracy of color chart thickness determination is limited to the accurate perception of the colors (what exactly is red-orange?). A typical chart is accurate to ±300 Å.

Fringes

When the order is not known, a fringe-counting technique can be used. When a test wafer edge is dipped in hydrofluoric acid for a few seconds, the acid quickly eats through the oxide at an angle, leaving the oxide exposed to view (Fig. 14.9). When the wafer is viewed in white light, colored fringes are formed between the wafer surface and the top of the film. Thickness determination is made by first determining the order of the film thickness. It is easy to see the repeated sequence of colors. If three blue-red fringes exist, the thickness corresponds to the surface color in the third order on the color-thickness chart.

A more accurate fringe-counting method uses monochromatic light as the viewing light. Monochromatic light consists of one color (wavelength), whereas white light is polychromatic (many wavelengths). The sample is prepared in the same way as it is for color fringe counting. However, in a microscope eyepiece, using monochromatic light, the fringes appear as alternating, evenly spaced black and white stripes (Fig. 14.10) with each fringe separation representing a specific vertical distance. Film thickness is determined by counting the number of fringes and multiplying by a correction factor. The correction factor is determined by the wavelength of the monochromatic light used. For sodium light, the wavelength is 5890 Å.

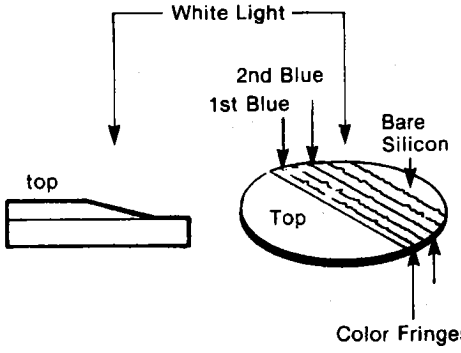


Figure 14.9 Color fringes.

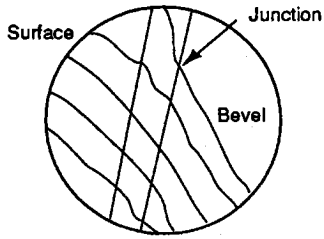


Figure 14.10 Junction depth determination by monochromatic fringes.

Spectrophotometers/reflectometry

Film thickness interference/reflectance measurement techniques can be automated. To understand the method, let's review the interference effects. Light is actually a form of energy. The interference phenomenon can also be described in terms of energy. White light is really a bundle of rays (different colors), each with different energies. When the rays interfere through the transparent film, the result is a ray of one color, one wavelength, and one energy level. It is our eyes that interpret the energy as a color.

In a spectrophotometer, which is an automatic interference instrument, a photocell takes the place of the human eye. Monochromatic light in the ultraviolet range is reflected off the sample and analyzed by the photocell. To ensure accuracy, readings are made under different conditions. The conditions are changed by either using another monochromatic light (to change wavelength) or changing the angle of the wafer to the beam. Spectrophotometers specifically designed for use in semiconductor technology have on-board computers to alter the measurement conditions and calculate the film thickness. With visible and ultraviolet (UV) light sources, these machines can measure films down to the 100-Å level.[4]

Index of refraction measurements are also made with these instruments. Spectrophotometers are also used to measure silicon film thickness. Because silicon is opaque to ultraviolet light, an infrared source is used.

Ellipsometers

Ellipsometers are film-thickness instruments that use a laser light source and operate on a different principle from that of a spectrophotometer. The laser light source is polarized. Polarization is the creation of a wave with all the rays traveling in only one plane. Polarization can be imagined by considering looking into the beam of a flashlight. In an ordinary beam, light rays come to your eyes in many planes, like an arrow with many feathers. A polarized beam has all of the light in only one plane, or an arrow with only one feather (Fig. 14.11).

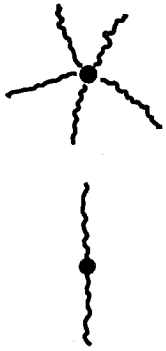


Figure 14.11 Unpolarized and polarized light.

In the ellipsometer, the polarized beam is directed to the oxide-covered wafer at an angle. The beam enters the transparent film and reflects off the reflective wafer surface. During its passage through the film, the angle of the beam plane is rotated. The amount of rotation of the beam is a function of the thickness and index of refraction of the film. A detector in the instrument measures the amount of rotation, and an on-board computer calculates the thickness and index of refraction.

Ellipsometers are used to measure thin oxides (50 to 1200 Å) and the index of refraction of the film. Their accuracy in this range is unequaled by any of the other techniques. The technique is also used for multilayer thin-film stacks. Ellipsometer accuracy and range is enhanced with the addition of multiple-viewing-angle capabilities, multiple-wavelength sources, and reduction in beam spot size.[5]

Oxides thinner than 50 Å can be measured by the corona-oxide-semiconductor (COS) technique described in the electrical measurement section.

Stylus (surface profileometers)

Some thin films, such as aluminum, cannot be measured by optical techniques. And in the case of aluminum and other very thin conductive films, the four-point probe thickness measurement is not sufficiently accurate. In these situations, a mechanical moving-stylus apparatus is used (Fig. 14.12). The method requires that a portion of the film be removed, creating a step on a test-wafer surface. This is normally done by a masking and etching step. The prepared sample is mounted and leveled on the pivoting stylus instrument stage.

After leveling, the measuring stylus is lowered gently onto one of the surfaces. The measurement is made as the stage is slowly moved under the stylus. As the stylus goes over the step, its physical position

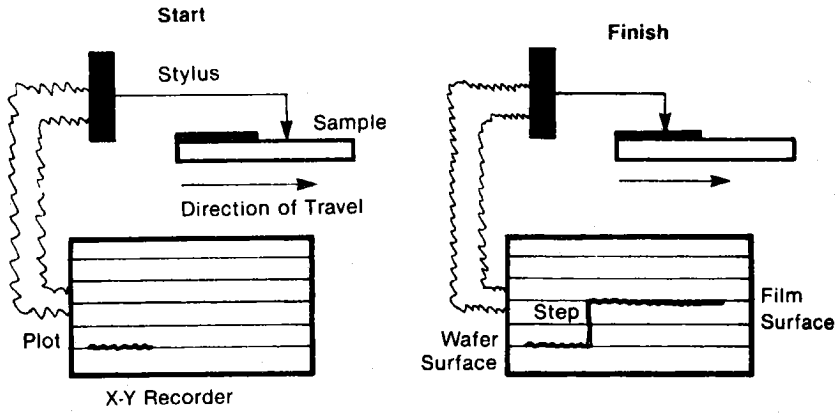


Figure 14.12 Step height measurement.

is changed. The stylus is linked to an inductor that generates an electrical signal in response to the stylus vertical position. This signal is amplified and fed into an *x-y* recorder.

While the leveled wafer is moving under the stylus, it does not move in the vertical direction, and no change in signal is produced. The trace on the *x-y* plotter is a straight line. When the stylus reaches the surface step, it changes position, causing a change in the signal output. This change is evidenced by a change of pen position on the *x-y* chart trace. The change in position is relative to the step height, which is read directly from the calibrated *x-y* chart.

Photoacoustic

A nondestructive thickness test relies on photoacoustic principles. In 1877, Alexander Graham Bell discovered that, under certain circumstances, the interruption of a light wave will cause a sound. In the semiconductor thickness application, a laser beam is converted to tiny sounds, which are in turn reflected off two surfaces on the wafer surface. By measuring the reflection delay between the two pulses, the thickness can be calculated.

Junction Depth

A critical device parameter is the junction depth of the various doped regions. This parameter is measured after each of the doping steps. The measurement methods described are all performed off line; that is, the test wafers or device wafers have to be taken to a measurement station or laboratory for the measurement.

Groove and stain

The traditional method of junction depth measurement is by the groove (or bevel) and stain technique. Grooving or beveling is a mechanical method of exposing the junction for viewing and measurement from the horizontal plane (Fig. 14.13). The extremely shallow depth of the junction requires either grooving or beveling of the wafer to expose the junction.

The junction itself is not visible to the naked eye. Two techniques, called *junction delineation,* are available to make it visible. Both techniques utilize the electrical differences between N-type and P-type regions. The first technique, the etch technique, starts with the placement of a drop of hydrofluoric acid and water mixture over the junction (Fig. 14.14). A heat lamp is directed onto the exposed junction. The heat and light cause holes or electrons to flow in each region. As a result of the flowing current, the etch rate of the HF-H_2O mixture is higher on the N-type region, making it appear darker.

The second delineation technique is electrolytic staining. A mixture containing copper is dropped on the exposed junction. Again, the heat lamp is directed onto the junction. In effect, a battery is formed, with the poles of the junction being the poles of the battery and the copper solution being the electrolytic connection. The current flowing in the liquid drop causes the copper to plate out on the N-type region side of the junction.

The final step, after exposure and delineation of the junction, is depth measurement. A number of methods may be used, including optical interference and SEMs.

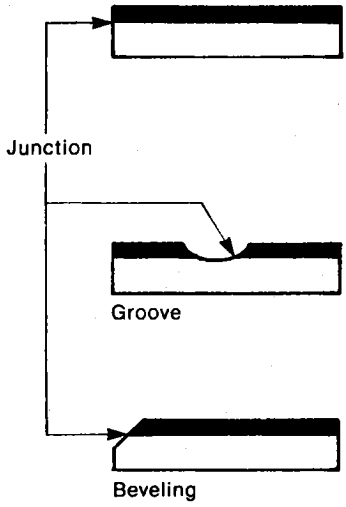


Figure 14.13 Exposure of junction by groove or beveling.

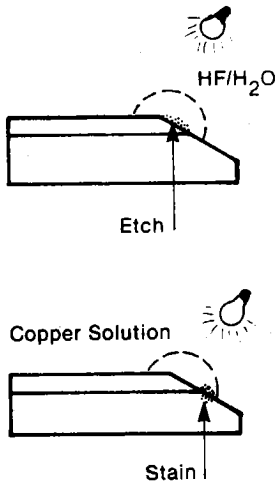


Figure 14.14 Etching or staining of junction.

Scanning electron microscope (SEM) thickness measurement

The SEM technique (described below) can also be used to measure junction depths and film thickness. The wafer is cleaved with the break position over the junction. The exposed wafer junction is delineated by one of the methods described.

The exposed cross section is positioned to the SEM beam at right angles to the wafer surface and a photograph is taken. The depth is determined from the photograph and the scale factor of the SEM (Fig. 14.15). SEM and the groove and stain methods provide a visual look at the junction area and lateral diffusion that other methods do not.

Spreading resistance probe (SRP)

Spreading resistance is a technique also used for measuring the junction depth. As the probes pass through the junction, they sense the change in conductivity type (N or P). This information, when plotted on the profile curve, also gives the junction depth (Fig. 14.16).

Secondary ion mass spectrometry (SIMs)

A SIMs instrument bombards the wafer surface with a source that produces secondary ions from species on the surface. It can detect the dopant(s) in the junction region. Coupled with careful removal of the surface through the junction and monitoring, when the dopant atoms disappear from the stream can be translated into a junction depth.

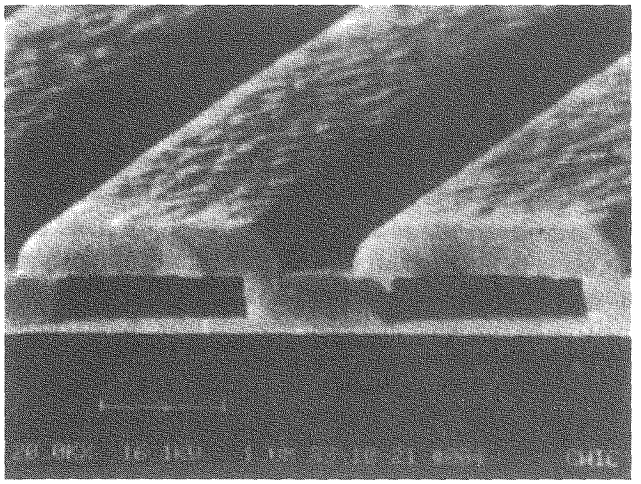


Figure 14.15 SEM declination of device cross section.

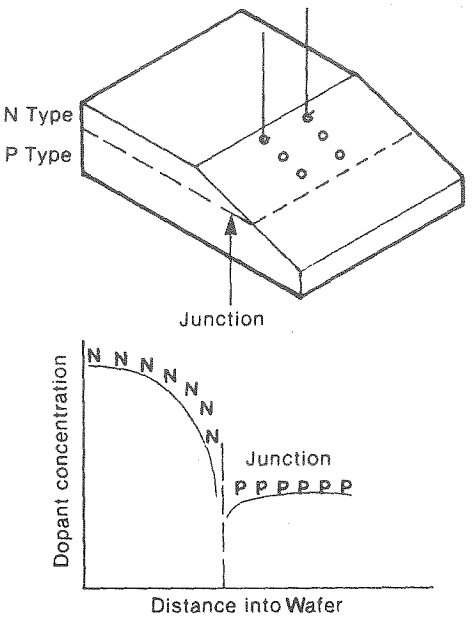


Figure 14.16 Spreading resistance.

Scanning capacitance microscopy (SCM)

One of the options with atomic force microscopy (AFM, see description below) is a probe that measures capacitance. Since the capacitance of a doped layer changes as the junction is approached, it can be used for

junction depth detection. Sample preparation is the same as for spreading resistance measurements.

SCM offers the advantage of nanometer precision[6] and may be the concentration profile and junction depth measurement technique for submicron junctions.

Critical Dimensions (CD) and Line Width Measurements

The exact dimensions required of each component in the circuit are controlled and influenced by *all* processes. Vertical dimensions are set by the doping and layering processes. The horizontal surface dimensions are produced in the lithography process. As part of that process, the critical dimensions are measured at both develop inspection and final inspection with microscopes.

Filar and image shearing dimension measurement

Two manual microscope methods are used to measure surface pattern dimensions: filar eyepiece and image shearing. Both have limited accuracy for small dimension lines but are found in laboratories and training labs. The filar measuring eyepiece is a dimension-measuring instrument that is fitted to a microscope. The eyepiece features a movable hairline that requires calibration to an outside standard. A calculation is made to "correct" the filar measurements to the actual dimensions on the standard.

Filar systems are easily automated. The hairline movement mechanism can be motorized and the correction factor programmed into an onboard computer, resulting in a direct digital readout of the actual width. Operator fatigue is minimized by a video monitor rather than requiring the operator to view the wafer through a high-power microscope.

An image-shearing attachment on a microscope is another method of critical-dimension measurement. A control on the unit allows the operator to optically separate (shear) the pattern into two images. To start the measurement, the two images are butted against each other (Fig. 14.17), then moved until the sheared images exactly overlap. The

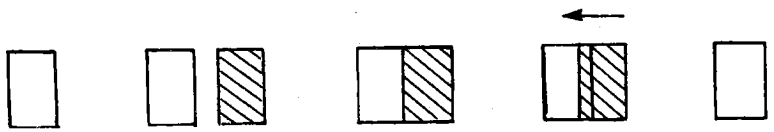


Figure 14.17 Single-image shearing.

difference between the starting and ending values is the width of the pattern. Like the filar unit, an image shearing unit must be calibrated to a standard.

Reflectance

Both the filar and image-shearing techniques require some operator decision, which can be a source of error. A third type of dimension measurement instrument is based on reflectance. Like the other two, the operator locates the pattern to be measured on a monitor. One edge of the pattern is positioned under a marker on the screen. The measurement is made automatically as a laser beam is swept along the direction of measurement and the reflection energy of the beam recorded (Fig. 14.18). When the beam comes to the edge of the pattern, it steps up (or down) to a new surface. The new surface causes a different reflectance of the beam, which is recorded by the detector. The width of the pattern is the difference between the starting point and the point at which there is a change in reflectance. This value is read out automatically. Operator decision is limited to determining the starting point, resulting in a more constant accuracy.

The limitation of this method is with patterns that have a number of steps in them, such as metal steps. The instrument is designed to read to the first change in reflectance that it senses. Another limitation is in measuring pattern edges that are sloped. The reflecting beams change gradually rather than abruptly, making a width determination difficult.

Some dimension-measuring microscopes use transmitted light for the measurement of mask and reticle critical dimensions. A light beam is passed from the bottom through the mask. The pattern to be measured is moved in front of the beam, and a detector above the mask senses when the light is blocked by the opaque pattern. The horizontal distance is measured until the light again is detected by the sensor.

Shape metrology and optical critical dimension (OCD)[7]

The SEM inspection tool described below is also used for very accurate measurement of line widths. In the nano-era, especially when using

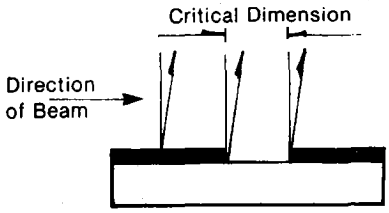


Figure 14.18 Reflectance CD measurement.

copper metallization, there is also interest in knowing and controlling the cross-sectional shape (3D shape metrology) of the hole or surface island. This is accomplished with sophisticated SEMs that direct beams off the top surface, the sides, and the bottom surface to reconstruct the exact shape. Measurements are made directly on the wafer. A draw back is OCD's inability to measure isolated lines. The great goal is to have OCD systems integrated directly into process tool, providing real-time measurement and process control (Fig. 14.19).

A summary of the various surface-inspection techniques and their use in a fabrication process is tabulated in Fig. 14.20.

Contamination and Defect Detection

Detection of contamination and visual defects is essential to high yields and process control. Particulate contamination is detected primarily by visual techniques including high-intensity lights, micro-

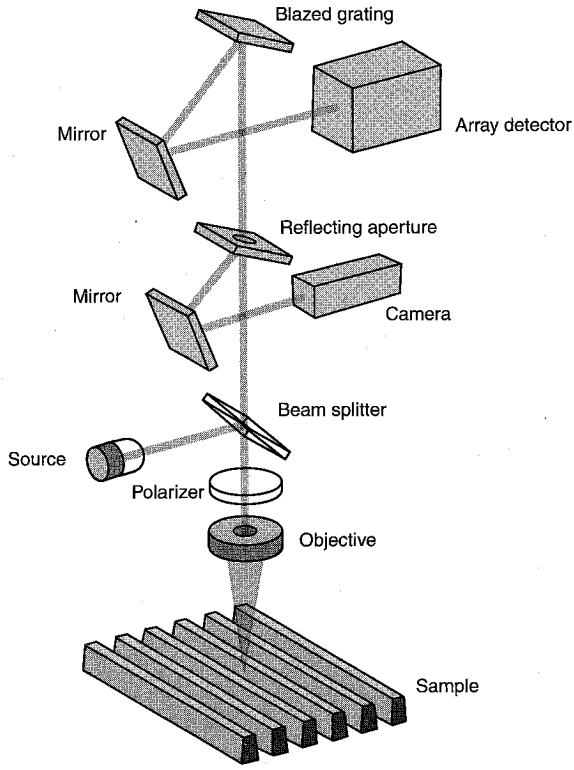


Figure 14.19 Schematic of OCD optics. (*Courtesy of* Semiconductor International.)

Method	Visual contamination	Surface Defects	Alignment	Contamination Elements	Contamination Compounds	C.D.'s
1x Incident White Light	X	X				
1x Incident U.V. Light	X	X				
Microscope-Light field	X	X	X			
Dark field	X	X	X			
S.E.M.	X	X	X			X
Auger				X		
E.S.C.A.					X	
Filar						X
Image Shearing						X
Reflectance						X

Figure 14.20 Overview of surface-inspection techniques.

scopes, SEMs, and automatic machines. Chemical contamination is detected and identified by both Auger and ESCA techniques.

Mobile ionic contamination in the wafer is detected by capacitance-voltage plotting and by interpretation of the transistor and diode electrical tests. Many other sophisticated techniques can be used to make these inspections. The ones described here are those employed in a typical wafer-fabrication line.

1 × visual surface inspection techniques

The first line of defense is to look at the wafers, which is a magnifying power of 1, or, in microscope terminology, a 1× power. Operators quickly become sensitive to the way "normal" wafers look. Even minor changes in the surface appearance are picked up by the experienced eye.

1 × collimated light

The resolving power of the naked eye (1×) can be assisted by using a high-intensity white light, such as the beam of light from a slide projector (Fig. 14.21). Particulate contamination is highlighted in the light beam when the wafer surface is viewed at an angle. The effect is similar to the highlighting of dust in the air by light streaming through a window.

1 × ultraviolet

In actuality, the eye cannot see ultraviolet light, but ultraviolet light from a mercury-vapor lamp emits blue, green, and even some red

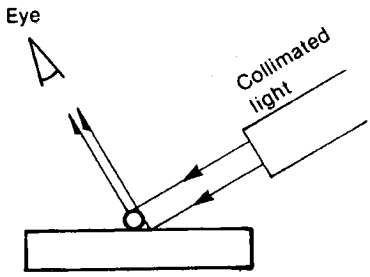


Figure 14.21 Collimated light inspection.

light. Because ultraviolet is harmful to the retina, a filter is frequently placed over the light source to block out the ultraviolet. The primary benefit of the ultraviolet lights used in fabrication areas is that they are very bright, which means that the intensity of the scattered light is greater, therefore increasing the detection of surface contamination.

Microscope techniques

Light-field microscope. The metallurgical microscope is the workhorse tool of surface inspection. The term *metallurgical* differentiates it from the standard microscopes found in biology labs. A biological microscope illuminates transparent samples by shining the light up through the sample. In a metallurgical microscope, the light is passed down to the nontransparent sample through the microscope objective (Fig. 14.22). The light reflects off the sample surface and is transmitted back up through the optics to the eyepieces. With white light illumination, the picture in the field of view exhibits the surface colors, which helps identify particular components on the wafer. The use of filters will change the surface colors.

A typical fabrication microscope is fitted with 10× or 15× eyepieces and a range of objectives from 10 to 100×. Increasing the total viewing power (eyepiece power times the objective power) reduces the field of

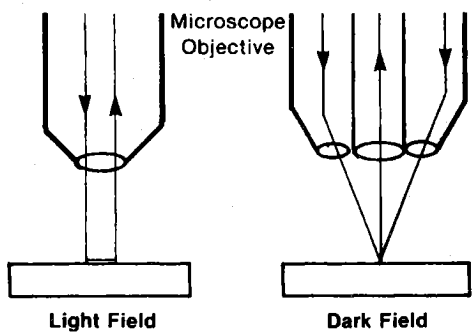


Figure 14.22 Light- and dark-field inspection.

view. This reduction requires more inspection time for the operator to look at the required sample inspection area on the wafer. The consequence is a slower inspection process. A trade-off power level, when inspecting LSI and VLSI devices, is 200 to 300× magnification.

The industry typically uses a microscope inspection procedure requiring the operator to look at three to seven specific locations on the wafer. This procedure is easily automated with motorized stages. Most of the automated microscope inspection stations feature automatic wafer placement on the stage, automatic focusing, and automatic binning of the completed wafers. At the touch of a button, the operator can direct each wafer into a boat for passed wafers or into one of several boats for different reject categories. Obviously, a microscope inspection procedure is used to judge surface and layer quality and (in masking) pattern alignment.

Just as image resolution in a patterning process is limited by the wavelength of light, so is bright light inspection. With a broadband white light source, the theoretical resolution limit is 0.30 μm.[8] Use of UV light sources and image processing can bring the resolution limit down to 80 nm.

Dark-field inspection. Dark-field illumination is achieved by fitting a metallurgical microscope with a special objective (Fig. 14.22). In this objective, the light is directed to the wafer surface through the outside of the objective body. It impinges on the surface at an angle and passes up through the center of the objective. The effect on the "picture" in the eyepieces is to render all flat surfaces black. Any surface irregularities, such as a step or pieces of contamination, appear as bright lines. Dark-field illumination is more sensitive than light-field to any surface irregularity. It has the drawback of limiting the ability to discern the nature of the surface irregularity. A passable surface dimple may look the same as a rejectable piece of contamination.

Dark-field resolution of defects is enhanced with the use of laser light[8] sources and multiple sources.

Confocal microscopes. Distinguishing fine detail with a normal microscope is hampered by stray light rays interfering as they bounce off the various surfaces and from the different plane depths on a wafer surface. Confocal light sources minimize the scatter and provide greater detail by limiting the returning light rays to a narrower plane. This is done by passing an intense white light or laser beam through holes in a disk (positioned between the light source and the wafer) spinning at high rpm. Only light rays from the inspection plane of interest bounce back through the holes. Confocal systems are capable of

imaging submicron dimensions.[9] Laser beams used in a confocal system cause the sample to fluoresce (glow). The resultant image, after x-y scanning and computer processing, has resolution in the nanometer range and can display layers of registered translucent samples.

Other microscope techniques. Optical technology is capable of providing many evaluation techniques beyond simple light- and dark-field viewing, such as phase contrast and fluorescence microscopes. Each allows the viewer to determine additional visual information about the surface. Phase contrast brings out surface irregularities in the vertical plane, and fluorescence-illuminated microscopes use ultraviolet illumination sources. In the ultraviolet light, organic residues (photoresist, cleaning chemicals) not easily visible in white light are brought into view. Their use and interpretation generally require technicians trained beyond the level of production operators.

Scanning electron microscopes (SEMs). Conventional optical microscopes are limited in their ability to provide accurate information about the wafer surface. First, their resolving power is limited by their optical light source. The ability of a viewing system to distinguish detail is related to the wavelength of the light (radiation used). The shorter the wavelength, the smaller the detail that can be seen.

Depth of field is another viewing factor. It relates to the ability of the system to keep two planes in focus simultaneously. A conventional photograph, with the subject in focus and the background out of focus, has a background beyond the depth of field limit of the camera. In a microscope, the depth of field decreases as the power (magnification) of the system is increased. If the power is increased to see the surface "closer," the operator may not be able to see the top and bottom surfaces in focus. Constant refocusing results in loss of information and a longer inspection time.

Magnification is the third limiting factor of optical microscopes. An optical system with white light illumination is limited to about 1000× magnification with conventional objectives. The oil-immersion technique pushes the limit up, but it is unacceptable, because it is too slow, too messy, and a possible source of contamination to the wafer.

All three limitations are overcome by using a scanning electron microscope. The microscope varies from an optical one in many aspects. The "illumination" source is an electron beam scanned over the wafer or device surface. The impinging electrons cause electrons on the surface to be ejected. These secondary electrons are collected and translated into a picture of the surface (Fig. 14.23) on either a screen or a photograph.

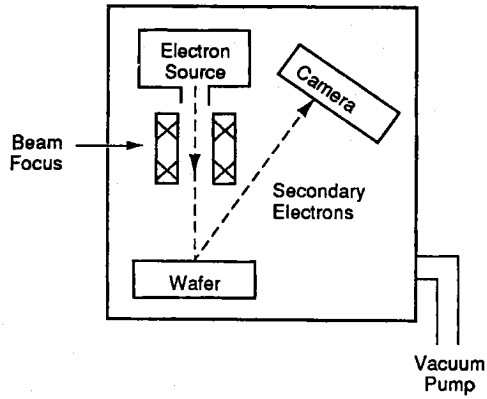


Figure 14.23 SEM analysis.

SEM analysis requires that the wafer and beam be in a vacuum. The electron beam has a much smaller wavelength than white light and allows the resolution of surface detail down to submicrometer levels. Depth of field problems do not exist; every plane on the surface is in focus.

Magnification is similarly very high, with a practical upper limit of 50,000×. A tilting wafer holder in an SEM allows the viewing of the surface at angles, which enhances the three-dimension perspective (Fig. 14.24). Surface details and features can be viewed at advantageous angles.

Some materials, such as photoresist, do not give off secondary electrons in response to e-beam bombardment. To inspect a photoresist

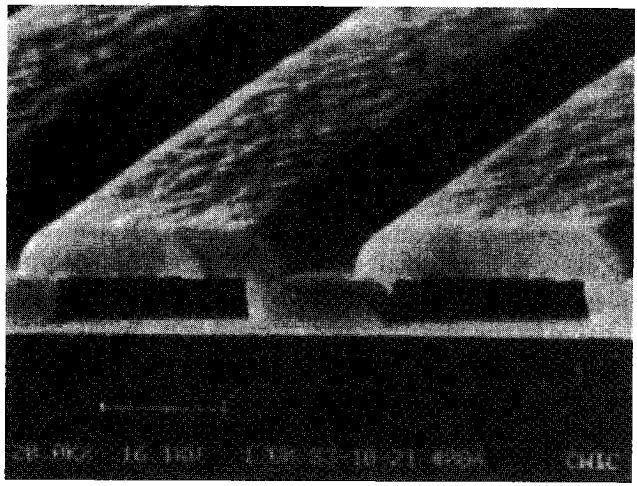


Figure 14.24 SEM declination of device cross section.

layer in an SEM, the resist layer is covered with a thin layer of evaporated gold. The gold layer conforms to the topography of the photoresist layer. Under e-beam bombardment, the gold gives off secondary electrons, thus resulting in an SEM picture that is an accurate reproduction of the underlying photoresist layer.

Transmission electron microscope (TEM). A standard SEM instrument has a resolution range of 20 to 30 Å,[10] which presents a barrier for inspection of ULSI devices. However, passing (transmission) the electron beam through a thin sample increases the resolution to 2 Å. As attractive as the increased resolution may be, sample preparation is time consuming and requires precision. TEM is best used as an offline laboratory tool.

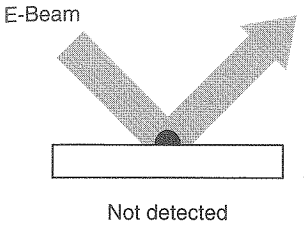
This principle is used to measure film thicknesses. In an SEM, the energetic e-beam causes X-rays to leave the surface. Fortunately, the X-ray energies are reflective of the materials (elements) presence on the surface and can be analyzed for additional information. This is accomplished when an X-ray spectrometer is added to the SEM to detect dispersed X-rays. This additional information is about the chemistry of the materials.

Automated in-line defect inspection systems

Automatic defect detection. The detection of ever smaller-sized particles on the wafer surface has led to the use of laser beam and e-beams as the detection illumination. Two advantages accrue from the use these beams. First is the obvious ability to detect smaller particle sizes (Fig. 14.25) due to the high intensity of the reflected light, which makes very small surface particles detectable (helium and neon lasers are the usual sources). E-beams have even smaller spot size and are the system of choice for nano-sized patterns, especially for copper/low-k dielectric metallization systems.

The second advantage is automation. Laser inspection equipment is easy to automate such that the inspection is automated from cassette to cassette, and the quantity and size of the contamination on the surface can be determined. The information is used to produce a map of the wafer surface showing the size, location, and density of surface contamination. This last factor is a big help in process control. Laser inspection of masks has become a standard production technique.

Some systems mix bright- and dark-field viewing and process the information into maps or data bases (Fig. 14.26). Advances in image processing have allowed automatic defect and pattern distortion detection instruments. The machines mate image processing and computer tech-



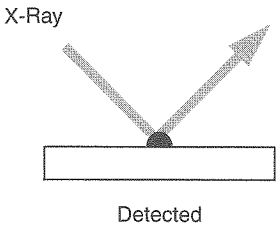


Figure 14.25 X-ray detection of small surface particles.

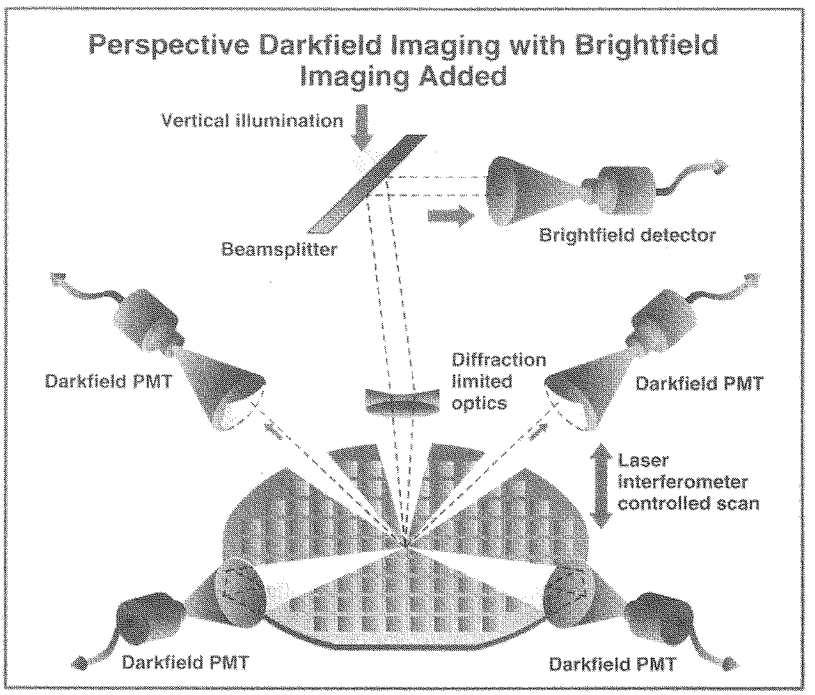


Figure 14.26 Mixed dark-field and light-field system. *(Courtesy of* Semiconductor International, *May 1997.)*

nology. They have scanners with laser or optical light sources that move over the wafer surface. In one version, the computer is preprogrammed with the design pattern for the circuit. In the *die-to-database* system, each die is scanned and the results compared to a stored image of the mask/reticle for the particular layer. The scanner looks for added or missing parts from the expected pattern. Anything missing from or on the die that is not in the data base is flagged for further inspection. The presumption is that, if an image is not in the data base, it is probably a defect of some kind. Defect locations are recorded and surface maps printed out. Engineers can go back to the wafer or mask and determine the exact nature of the problem.

Another system, called die-to-die inspection, compares adjacent die on the wafer or mask. A die is scanned and the pattern recorded in the computer memory. A second die is also scanned, and any deviations between the two are recorded. This system will not detect any repeated pattern defect that occurs on every die, but it will pick up random defects that have a very low probability of occurring in exactly the same spot on two adjacent die. In both machine types, the information from the surface is captured by charge coupled device (CCD) cameras or photomultiplier tubes.[12] For in-line inspection, a grave concern is calibration and standardization.

In addition to on-board verification of the electronics, most systems use a standard wafer to verify machine operation. Several inspection processes are used. While an automatic machine can find defects, the determination of which ones are "killer" defects is more difficult. With overall defect counting, it is possible that the machines may report a rise in defect count, yet the increase may contain only minor, nonkiller types. Verification by humans is still a critical part of any defect detection and management system.

The development of copper/low-k dielectric combinations has introduced a whole new host of defects to the process. Figure 14.27 lists a sampling of defects associated with these metal systems.

General Surface Characterization

Atomic force microscopy (AFM)

Several of the measurement techniques are multiuse, such as SEMs. A newer entry in the general category is *atomic force microscopy* (AFM). It is a surface profiler that scans a delicately counterbalanced probe over the surface (Fig. 14.28). Probe and surface separation is so small (about 2 Å) that atomic forces between the surface and probe materials actually affect the probe. Sophisticated electronics measure the relative position of the probe as it moves across the surface, thus compiling a three-dimensional map of the surface[13] (Fig. 14.29).

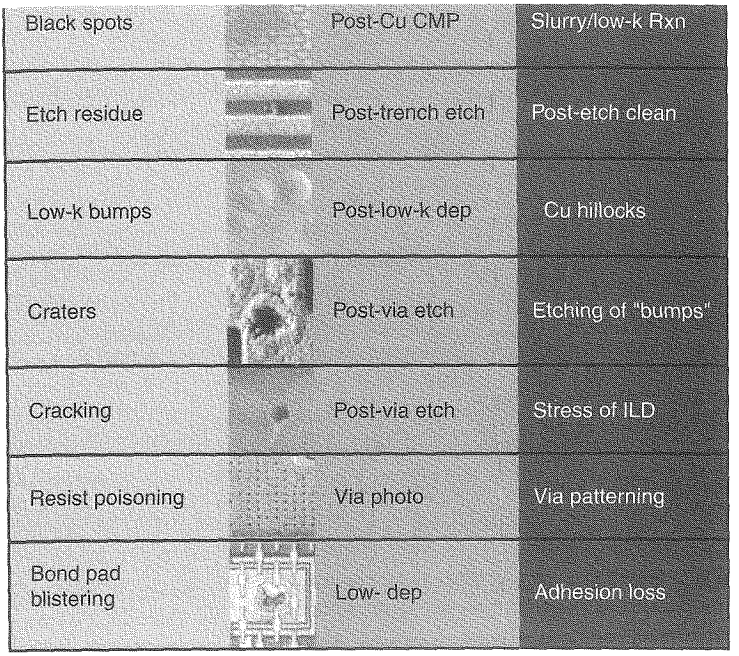


Figure 14.27 Sources of low-k defects. *(Courtesy of* Semiconductor International.*)*

AFM sensitivity is in the 1-Å range and can be operated in contact or noncontact modes. The first AFM production instrument was introduced in 1990. Projections are for it to become an integral part of the inspection arsenal. By its nature, AFM can characterize grain sizes, detect particles, measure surface roughness, and provide critical dimension measurements—all in three dimensions. In one novel use, an AFM probe has been combined with an optical microscope objective.[14]

Scatterometry

The search for fast, accurate, and nondestructive surface inspection tools led to the *scatterometry* metrology. Accuracy in optical systems is limited by the wavelength of the light. Roughly, particles or surface features smaller than the wavelength used cannot be detected. However, a scattered beam can give information about surface features smaller than the wavelength. A scatterometry system has the wafer at the center of curvature of a screen (Fig. 14.30). An incident laser beam is scanned over the surface and is reflected and "scattered" from the surface onto the screen. A camera with a microprocessor captures the screen image to reconstruct the surface that produced the particular

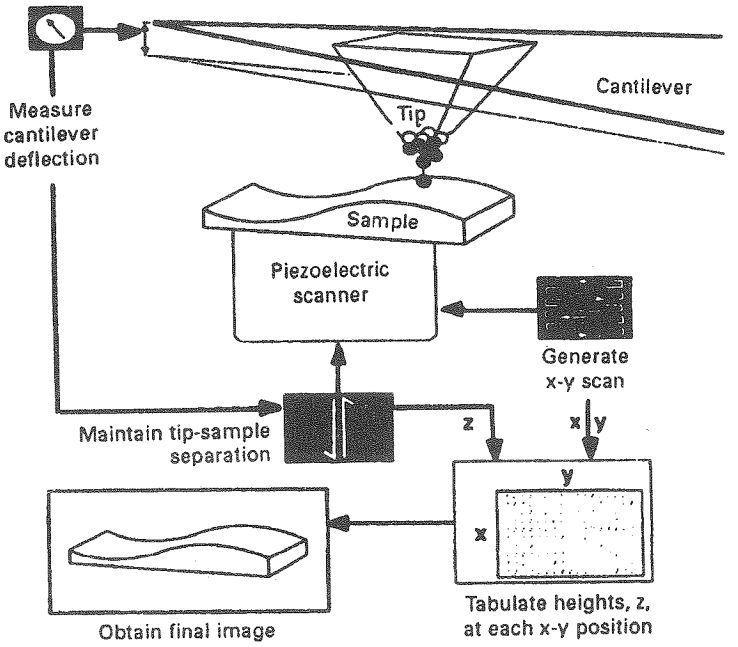


Figure 14.28 Atomic force microscope (AFM). *(Source:* Semiconductor International, *August 1993, p. 66.)*

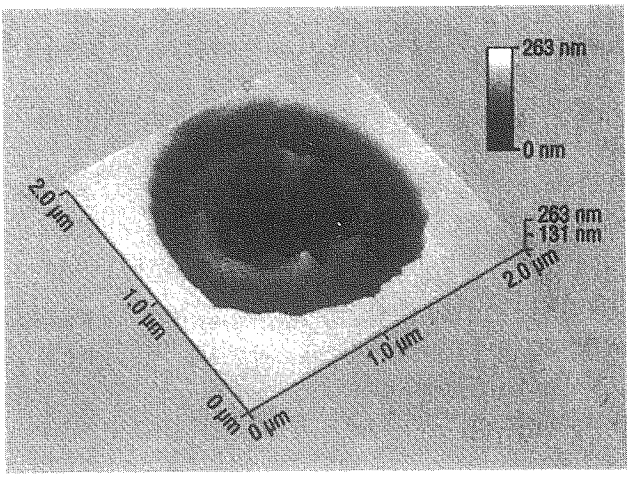


Figure 14.29 Atomic force microscope (AFM) surface map.

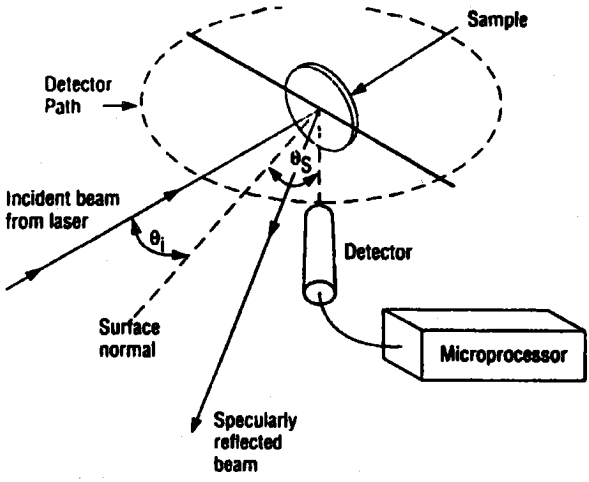


Figure 14.30 Arrangement of a scatterometer. *(Source: Solid State Technology, March 1993, p. 30.)*

pattern on the screen. Like AFM, this technique has the potential of measuring grain sizes, contours, and CDs. It can also measure latent images in undeveloped photoresist and characterize phase shift masks.

Contamination Identification

Pushing into the 0.18-μm era is requiring more detailed information about the nature of the various contaminants that end up on the wafer surface or in the deposited layers. Contamination types, forms, amounts, and other data are needed to maintain clean processes and product. In this section, the instruments used to collect this data are examined. In general, all of the techniques are based on a common phenomenon. Whenever a surface is excited with energy, there will be energy given off. The energy given off will have characteristics reflecting the interaction of the incoming energy wave with the material(s) on the wafer surface.

Auger electron spectroscopy (AES)

In an SEM, a range (spectrum) of secondary electrons is released by the impinging electron beam. One portion of this spectrum is the electrons that are released from the top several nanometers of the surface. These electrons, known as *Auger electrons,* have energies characteris-

tic of the element that emits them. Thus, sodium and chlorine each give off different Auger electrons.

The collection and interpretation of Auger electrons allows the identification of the surface materials, including contamination. In operation, the e-beam is scanned across the wafer. The ejected Auger electrons are analyzed for their energies (wavelengths) and printed out on an *x-y* plotter (Fig. 14.31). Energy peaks at specific wavelengths indicate the presence of specific elements on the surface.

Scanning Auger microanalysis (SAM) is limited to the identification of elements. This technique cannot identify the chemical state of the element, what it may be combined with, or its quantity on the surface. Salt (NaCl) contamination is identified only as the presence of sodium and chlorine on the surface.

Electron spectroscope for chemical analysis (ESCA)

Solving surface contamination problems often requires a knowledge of the chemical state of the contaminant. An Auger detection of chlorine on the surface does not reveal whether the chlorine is present as hydrochloric acid or a trichlorobenzene. Knowledge of the form of the chlorine expedites locating and eliminating the process source of contamination.

The electron spectroscope for chemical analysis (ESCA) is an instrument used to determine surface chemistry. The instrument works on principles similar to the Auger technique. However, X-rays instead of electrons are used for the bombarding radiation. Under bombardment, the surface gives off photoelectrons. Analysis of the photoelectron information leads to the determination of the chemical formula of the contamination. Unfortunately, the ESCA X-ray beam is wider than most integrated circuit features. The beam diameter limits the tech-

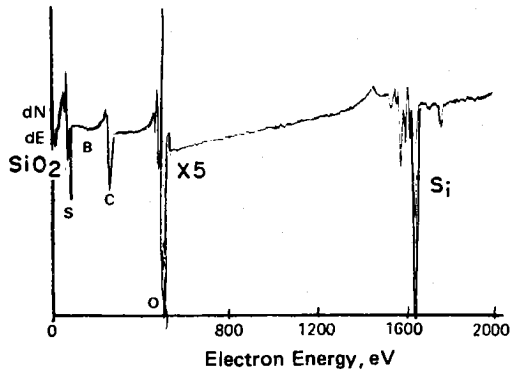


Figure 14.31 Typical Auger trace.

nique to a macro-surface analysis. By contrast, an Auger electron beam can zero in on specific bits of contamination.

Time of flight secondary ion mass spectrometry (TOF-SIMS)

TOF-SIMS surface analysis can use a number of incident radiation sources, such as an Nd YAG laser, a Cs ion beam, or a Ga ion beam.[15] For this method, the impinging energy pulses and ionizes surface material and produces secondary ions that are accelerated to a mass spectrometer. In the spectrometer, each ion's time of flight from the surface is measured. The time of flight is indicative of the species on the surface. An SIMS analysis can also determine the junction depth.

TOF-SIMS samples material depths of only several tenths of a nanometer.[16] Along with the ability to characterize both organic and inorganic contaminants, it is a mainstay of an analytical lab.

Vapor phase decomposition/atomic absorption spectroscopy (VPD-AAS)

VPD-AAS is used to detect trace inorganic contaminants. The sample is exposed to vapor phase hydrogen fluoride gas, which chemically reduces (decomposes) the contaminant to its fluoride salt. Next, the reduced material is dissolved in water or an acid and analyzed for elements with standard atomic absorption spectroscopy.

Evaluation of stack thickness and composition

The top of the wafer is becoming more and more complicated with stacks of various materials. Each material can be measured for thickness and composition from monitor wafers included in the deposition process. However, their effect in and on the circuit comes from the combination of thicknesses and compositions. A combination machine to determine these parameters, *in situ*, uses an electron beam and X-ray spectrometers (Fig. 14.32). An e-beam is directed at the stack, which in turn kicks off X-rays. A series of X-ray spectrometers arrayed around the sample analyze the rays. Because each thickness and material composition creates a unique X-ray signature, the machine's onboard computer can run an analysis and determine the thickness and composition of each component in the stack.[17]

Device Electrical Measurements

During the process, it is necessary to make direct measurements of the actual device parameters. These measurements are usually made

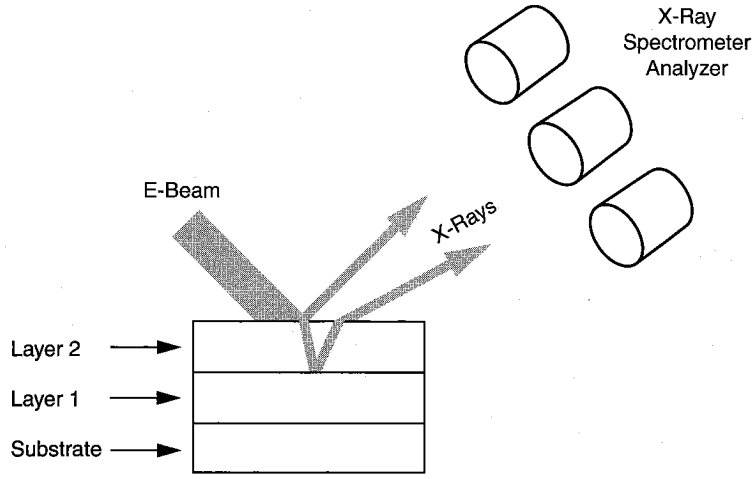


Figure 14.32 E-beam/X-ray detection of material stacks.

on special devices in the test die or on special structures in the scribe line areas.

A great deal of information about the process is available from these electrical tests. In this section, we explain the basic tests and several of the more obvious and frequent device failures. Often, these tests are referred to as *parametric testing.* The reference is to the measurement of the device electrical parameters as opposed to testing the overall function of the device/circuit, which occurs at wafer sort at the end of the fab process. A complete device and process troubleshooting guide is beyond the scope of this text.

Equipment

The basic equipment required to perform device electrical tests are (1) a probe machine (Fig. 14.33) with the capability of positioning needle-like probes on the devices, (2) a switch box to apply the correct voltage, current, and polarities to the device, and (3) a method of displaying the results. If many devices are to be measured, the probes may be fixed on a printed circuit board-like structure called a *probe card.* They are the same cards used in the wafer sort process. Display may be by a curve tracer or by a special digital voltmeter. A curve tracer is a specially designed oscilloscope set up to display voltage on the horizontal scale and current on the vertical scale. They show the characteristic voltage/current traces for each type of measurement. Example displays are shown in the following sections for each test. In high-production situations, the tests are performed digitally and the results

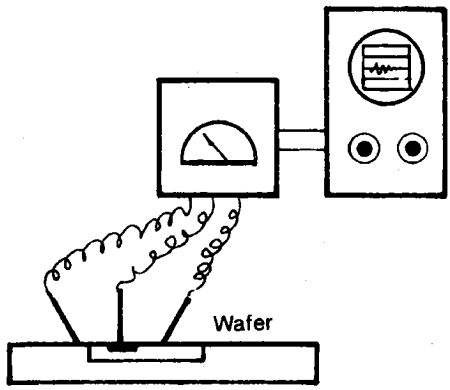


Figure 14.33 Device measurement equipment.

displayed on a readout. These measurement systems have the capability of storing and correlating data.

In advanced systems, the probe station may be automated to sequentially test several die. The switching can also be automated to allow the equipment to perform a series of tests in a predetermined sequence. Automated systems also include hard-copy printouts of the test results as well as analysis of the data. The individual tests will be explained, as performed on manual equipment. The shape and relationship of the traces displayed on the oscilloscope are very helpful for understanding device operation.

All the device measurements are made in basically the same way. A voltage is applied to the component contact probe, and the resultant current flowing between the contacts is measured. The results are displayed on the oscilloscope screen or display screen. The shape of the trace is governed by the dimensions of the device, the doping, and the presence of junctions. The current is modified by the resistance of the component or the presence of junctions. Properly working components will display known traces. The fundamental relationship is that of Ohm's law:

$$R = V/I$$

Resistors

Resistor measurements are made by contacting each end of the resistor and applying a voltage. Current passage between the probes will be governed by the nature of the material or, in electrical terms, its resistivity. During the test, the applied voltage is varied from zero to some higher value. On the screen, the voltage value is displayed on the x axis, with current measured on the y axis (Fig. 14.34).

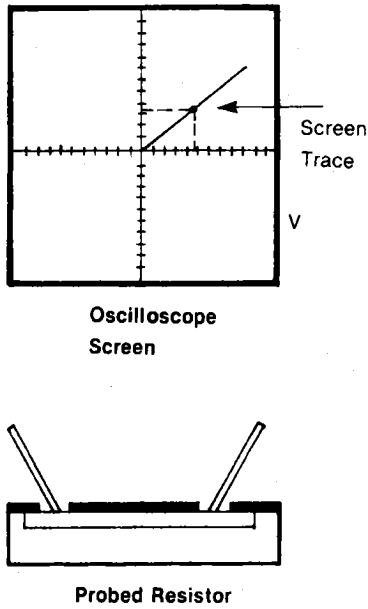


Figure 14.34 Resistor.

The value of the resistance is calculated by dividing one of the voltage values by the corresponding current value. One might ask, "Why not determine the resistance by simply measuring the voltage and current values with a meter?" In other words, why display the values on the screen? The answer lies in the quality of information gained from the trace. A resistor's V/I relationship should be a linear one (straight line). Any deviation from linearity could indicate a process problem, such as high contact resistance or a leaking junction.

Diodes

Diodes function as switches in a circuit. This means that a diode can pass current in one direction (forward bias) and not in the other (reverse bias). Checking diode operation in either direction requires measuring the diode with the proper polarities, as shown.

As the voltage is increased in the forward direction, current immediately starts flowing across the junction and out of the diode (Fig. 14.35). The initial resistance to that flow comes from contact resistance and a small resistance at the junction. After the resistances are overcome, there occurs a "full" flow of current through the diode. A diode is designed to have this condition occur at some minimum voltage value. If the diode forward voltage value exceeds the design value, it is out of specification. This voltage is called the *forward voltage.*

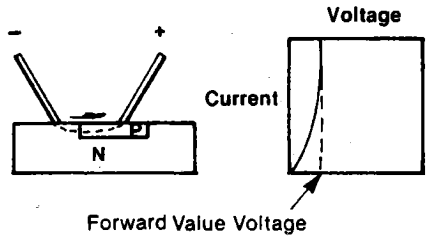


Figure 14.35 Diode forward bias.

In the reverse direction, the diode is designed to block current flow as long as the voltage stays below a specified value. In the reverse condition, a small current, called a *leakage current,* always flows across the junction (Fig. 14.36). Increasing the voltage increases the leakage current to a level at which the junction breaks down, allowing "full" current flow. This value of the voltage where full flow begins is called the *breakdown voltage* (Fig. 14.37). Circuits are designed to operate at a voltage level below the designed breakdown voltage of the diode to take advantage of the blocking nature of the junction. Lowered breakdown voltages are generally the result of processing mistakes or excessive contamination.

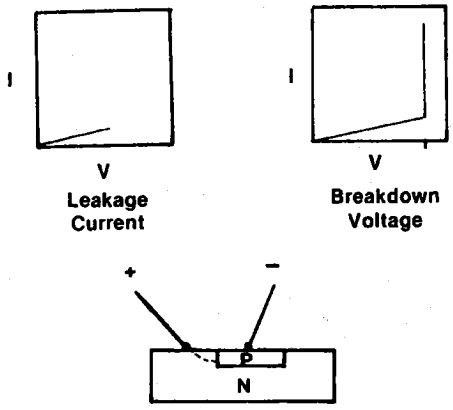


Figure 14.36 Diode reverse bias measurement.

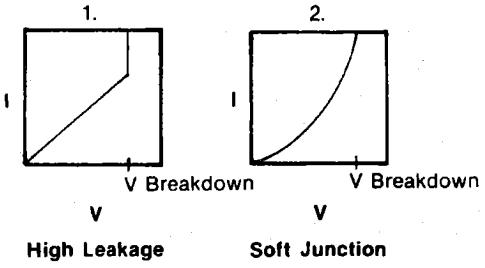


Figure 14.37 Junction leakage.

Exceeding the breakdown voltage does not (normally) permanently damage the junction. However, if the applied voltage is extremely high, the diode (junction) can sustain permanent physical damage from a high current flow.

A second parameter determined during this test is the current at the breakdown voltage. A small amount normally occurs as illustrated above. Contamination and/or improper processing can result in additional leakage current.

Trace 1 in Fig. 14.37 shows a diode with a small amount of leakage. The amount of current increases as the voltage is increased. Eventually, the breakdown voltage is reached, and the diode becomes fully conducting. In trace 2, gross leakage is demonstrated. The junction leaks current with every increase in voltage, and the problem is so severe that a breakdown level is never reached, and the diode never operates as a current block.

Bipolar transistors

Bipolar transistors, as explained in Chapter 16, are three-region, two-junction devices. Electrically, they can be thought of as two diodes back to back. Many tests are performed to characterize bipolar transistors. The individual junctions are characterized separately and the whole transistor operation measured. The junctions are probed for forward and reverse characteristics. The breakdown voltage (BV) tests are followed by the probing of the entire transistor.

Individual junction probe measurements are designated by the letters BV followed by lowercase letters that indicate the particular junction being probed. BV_{cbo} indicates the breakdown voltage measured between the collector and base regions. The "o" indicates that the emitter region is "open"—it has no voltage applied to it. BV_{cbo} is the breakdown measured between the collector and emitter.

The forward voltages of the two bipolar structure junctions are also measured. V_{be} is the forward voltage of the emitter-base junction, and V_{bc} is the forward voltage of the base-collector junction. BV_{iso} probes the collector-isolation junction for leakage currents.

A principal electrical measurement of a bipolar transistor is the beta (Fig. 14.38) measurement. This is a measurement of the amplification characteristic of the transistor. In a bipolar transistor, the current flows from the emitter to the collector, through the base (see Chapter 16). The base current is varied to change the resistance in the base region. The amount of current flowing out of the collector (from the emitter to base) is regulated by the base resistance.

The amplification of the transistor is defined as the collector current divided by the base current. This number is known as *beta*. Thus, a

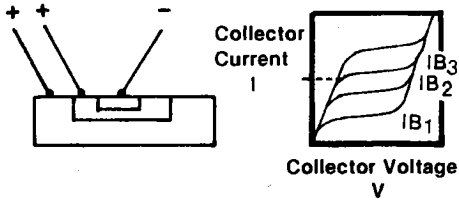


Figure 14.38 NPN transistor beta measurement.

beta of 10 means that a 1-mA (milliampere) base current will give rise to a 10-mA collector current. The beta of a transistor is determined by junction depths, junction separations (base width), doping levels, concentration profiles, and a host of other process and design factors. Measurement of beta is done by a variation of the BV_{ceo} measurement. A BV_{ceo} measurement at a specific base current is performed. In this mode, the emitter base junction is forward biased.

The collector characteristic of the transistor is displayed on the oscilloscope screen. The almost horizontal lines represent increasing base current values (IB_1, IB_2, and so on). With every increase in base current, a corresponding increase in collector current occurs. Calculation of the beta value takes place from the data displayed on the screen.

Collector current is determined from the vertical axis (dotted line). The base current is calculated by multiplying the number of horizontal lines (steps) by the scale value for each step (from the oscilloscope).

MOS transistors

MOS circuits are also made up of resistors, diodes, capacitors, and transistors. The first three are measured by the same methods used to measure bipolar circuit components. Like the bipolar transistor, the MOS transistor is composed of three regions, in this case called the source, gate, and drain (see Fig. 14.39 and Chapter 16). Measurement of this type of transistor consists of determining the reverse and forward values of the source and drain junctions. The functioning of the gate is determined by the threshold voltage test.

A MOS transistor has the source region forward biased. Because of the high resistivity of the gate region, the forward current does not

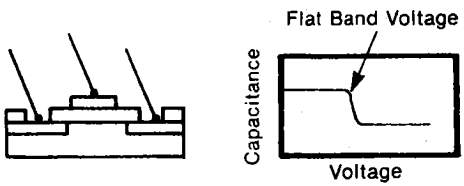


Figure 14.39 Threshold voltage measurement.

reach the drain. A voltage applied to the gate at a specified level (threshold) will cause enough charges to appear in the gate region to form a conducting channel that allows the source current to reach the drain region. Every MOS transistor is designed to operate at a specific threshold voltage. This value is measured using the capacitance-voltage technique. The gate voltage is continuously increased, while the capacitance of the gate structure is monitored.

A capacitor is a storage device. Initially, during the voltage-increase portion of the measurement, the capacitance does not change. At the threshold voltage level, the inversion layer forms and acts like a capacitor. Because two in-series capacitors have a combined lower capacitance than the sum of the two, the result is a drop to a combined lower capacitance. MOS transistors also exhibit amplification characteristics. The gain is defined as the source-drain current divided by the gate current. The source-drain characteristic for various gate currents is shown (Fig. 14.40).

Capacitance-voltage plotting

A variation of the threshold voltage test is used to test for the presence of mobile ionic contamination in the oxide. The test is performed on specially prepared test wafers. A thin oxide is grown on a "clean" silicon wafer. After oxide growth, aluminum dots are formed on the wafer by evaporation through a mask (Fig. 14.41). Dot evaporation is usually followed by an alloy step to ensure good electrical contact between the aluminum and oxide.

$$\text{Gain} = \frac{\Delta I_{DS}}{\Delta V_{GS}}$$

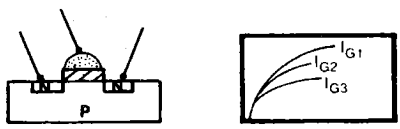


Figure 14.40 Gain characteristic of a MOS transistor.

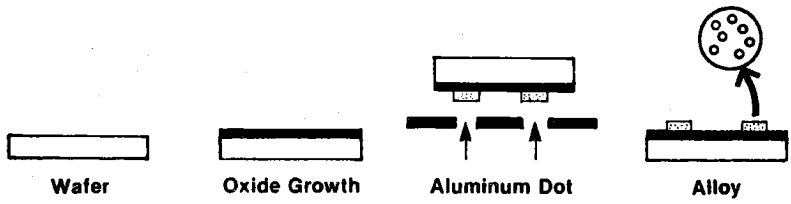


Figure 14.41 Preparation of C/V test wafer.

The "dotted" wafer is placed on a chuck, and a probe is placed on the aluminum dot. This structure is actually a MOS capacitor. A voltage is applied to the dot and gradually increased as the capacitance of the structure is simultaneously measured. The results are printed out on an *x-y* plotter with capacitance on the *y* axis and voltage on the *x* axis (Fig. 14.42).

At a voltage level known as the *threshold voltage* (or *inversion voltage*) charge starts to build up at the silicon surface. The charges "invert" the conductivity type from N-type to P-type. The inverted layer has a capacitance of its own. Electrically, the structure now has two capacitors in series. The total capacitance value of the two is less than the sum of the two by the relationship,

$$1/C_{\text{total}} = 1/C_{\text{oxide layer}} + 1/C_{\text{inversion layer}}$$

The trace on the *x-y* plotter drops vertically to the new capacitance level.

The second step in the process is to force the mobile positive ions in the oxide to the SiO_2-silicon interface. This is done by simultaneously heating the wafer to the 200 to 300°C level and placing a positive 50-V bias on the structure (Fig. 14.43). The elevated temperature increases the mobility of the ions, and the positive bias "repels" them to the oxide-silicon interface.

The last step in the process is a repetition of the initial *C/V* plot. However, as the voltage increases, inversion does not start at the

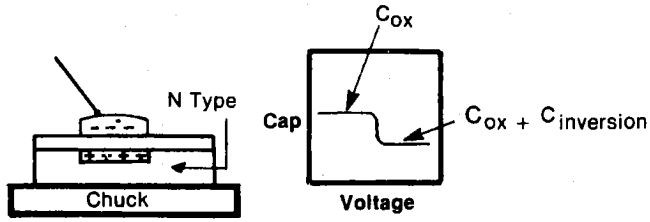


Figure 14.42 *C/V* plotting—first test.

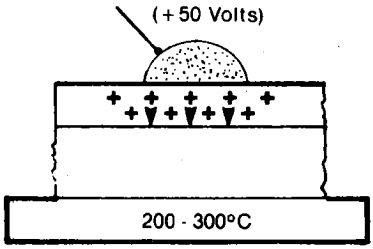


Figure 14.43 *C/V* plotting—ionic charge collection.

same level as in the initial test (Fig. 14.44). The positive charges at the interface require additional negative voltage to "neutralize" them before inversion can happen. The result is a C/V plot identical to the original but displaced to the right. The additional voltage required to complete the plot is known as the *drift* or *shift*.

The amount of the shift is proportional to the amount of mobile ionic contamination in the oxide, the oxide thickness, and the wafer doping. C/V analysis cannot distinguish the element (Na, K, Fe, and so forth) that was in the oxide—only the amount. Neither can this test determine where the contamination came from. It may have come from the wafer surface, any cleaning step, the oxidation tube, the evaporation process, the alloy tube, or any other process(es) the wafer has been through.

C/V analysis is usually a part of the evaluation of any process changes that may contaminate a wafer, such as a new cleaning process. To make the evaluation, C/V wafers are divided into two groups. One group receives normal processing as detailed above. The second group goes through the proposed cleaning process, usually between the oxidation and aluminum steps of the test. The drift on the standard processed wafers is compared with the drift of the experimental group. An increased drift on the experimental wafers would indicate that the proposed cleaning process actually contaminated the wafers with mobile ionic contamination.

Acceptable C/V drifts vary between 0.1 and 0.5 V, depending on the sensitivity of the device being made. C/V analysis has become a standard test in a fabrication area. The test is made after any process change, equipment maintenance, or cleaning that could have the potential of contaminating the wafers. The C/V plot provides a wealth of other information, such as flat-band voltage and surface states. The thickness of the gate oxide can also be calculated from the C/V test data.

Contactless C/V measurement. C/V monitoring as described above requires rigorous test wafer preparation, which is time consuming and expensive. Another method to determine unwanted charges (drift) and the other MOS gate parameters is with a contactless method called COS (Fig. 14.45). The MOS transistor method requires two electrodes separated by the gate oxide. A voltage on the top electrode (the metal)

1. Original Plot
2. Second Plot
3. Voltage Shift or Drift

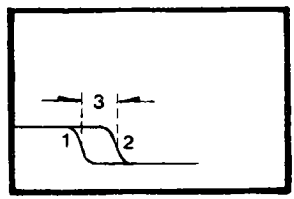


Figure 14.44 C/V replot.

produces a charge buildup at the metal/oxide interface. A similar result can be obtained by using a corona source (the *C* of COS) to build a charge directly on the oxide surface. Increasing the charge puts the MOS transistor through the same paces as a metal electrode device and the same information, charge (drift), flat-band voltage, surface states and oxide thickness, can be calculated. The method correlates with the standard measurements.[16]

Device failure analysis—emission microscopy

When a semiconductor device is operating, it gives off certain emissions of visible light. When there are certain problems, spots of visible light are given off at the trouble spot. For example, contamination-induced junction leakage will reveal itself as a light spot on the junction (at the surface). Microscopes fitted with sensitive detectors and charge coupled imaging devices (CCD) can locate and image these trouble spots. This method is particularly useful when electrical measurements indicate a failure in a sector of a circuit, but it cannot pinpoint the exact device(s) causing the failure.[17]

Review Questions

1. Changing the size of a resistor will change which of the following?
   a. Resistance
   b. Resistivity
   c. Both

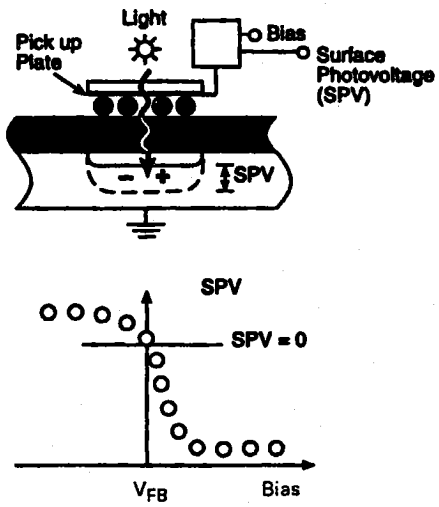


Figure 14.45 Noncontact surface charge measurement. *(Source: Keithley Instruments Quantox—TM.)*

2. In a four-point probe, the current is flowing in the inside probes (true or false).
3. The color of an oxide film on silicon is an indication of its exact thickness (true or false).
4. Which is the most accurate determination of a thin oxide layer (less than 1,000 Å)?

   a. Color

   b. Fringes

   c. Ellipsometer

   d. Spectrophotometer
5. Explain why an oxidized silicon wafer changes color when rotated.
6. Name a method to delineate an N-P junction for the groove and stain technique.
7. A collimated light is used to detect sodium contamination (true or false).
8. An SEM uses which magnifying power as compared with a light-field microscope?

   a. Higher

   b. Lower

   c. The same
9. Name two methods of measuring CDs.
10. If an oxide layer is contaminated with sodium, what is the effect on the C/V drift?

    a. C/V drift will be higher.

    b. C/V drift will be lower.

    c. C/V drift will remain the same.

References

1. B. McDonald, "Analytical Needs," *Semiconductor International,* Cahners Publishing, January 1993, p. 36.
2. S. Felch, "A Comparison of Three Techniques for Profiling Ultrathin p^+-n Junctions," *Solid State Technology,* January 1993, p. 45.
3. S. Felch, "A Comparison of Three Techniques for Profiling Ultrathin p^+-n Junctions," *Solid State Technology,* January 1993, p. 45.
4. P. Burggraaf, "Thin Film Metrology: Headed for a New Plateau," *Semiconductor International,* Cahners Publishing, March 1994, p. 57.
5. Thin Film Measurements, Rudolph Engineering, product brochure, 1985.

6. R. McDonald, "How will we examine IC's in the year 2000?" *Semiconductor International,* Cahners Publishing, January, 1994, p. 46.
7. A. Braun, "Metrology adapts to meet CD measurement needs," *Semiconductor International,* February 2002, p. 73.
8. A. Braum, "Optical Microscope Continues to Meet High Resolution, Defect Detection Challenges," *Semiconductor International,* December 1997, p. 59.
9. J. Baliga, "Defect Detection on Patterned Wafers," *Semiconductor International,* May 1997, p. 64.
10. *Product Description Bulletin,* Technical Instrument Company, San Francisco, CA, 1995.
11. S. Wolf and R. Tauber, *Silicon Processing for the VLSI Era,* Lattice Press, Sunset Beach, CA, 1986, p. 95.
12. P. Burggraaf, "Patterned Wafer Inspection Now Required!" *Semiconductor International,* Cahners Publishing, December 1994, p. 57.
13. L. Peters, "AFMs: What Will Their Role Be?" *Semiconductor International,* Cahners Publishing, August 1993, p. 62.
14. "Lithography News," *Semiconductor International,* December 1994, p. 24.
15. "Time-of-Flight Secondary Ion Mass Spectrometer," brochure, Charles Evans & Associates, Redwood City, CA, 1994.
16. A. Braum, E-Beam Techniques Measures Product Wafer Composition, Thickness, *Semiconductor International,* November, 2001.
17. P. Chu and R. Hockett, "New Ways to Characterize Thin Films," *Semiconductor International,* Cahners Publishing, June 1994, p. 143.
18. M. Peters, *COS Testing Combines Expanded Charge Monitoring Capabilities with Reduced Costs,* Keithley Instruments, Inc., product description paper.
19. T. Adams, "IC Failure Analysis: Using Real-time Emission Microscopy," *Semiconductor International,* Cahners Publishing, July 1993, p. 148.

Chapter

15

The Business of Wafer Fabrication

Overview

From its humble beginnings in laboratory-like production facilities to the automated wafer fabs of today, the semiconductor industry continues to evolve as the demands of new circuits, business factors, and global competition grow. While Moore's law still sets the product goalposts, fiscal considerations are setting the goals for productions facilities, equipment, and processes. Overall productivity and cost-of-ownership are the factors determining the bottom line. Strategies to increase the bottom line include maximizing equipment cost of ownership, automation, cost control, computer-automated manufacturing, computer-integrated manufacturing, and statistical process control.

Objectives

Upon completion of this chapter, you should be able to:

1. List the major cost factors that influence fabrication costs.
2. Describe the intent and factors of cost of ownership (COO) models.
3. List the advantages of statistical process control.
4. Identify the parts and use of a control chart.
5. List and discuss the different levels of automation.
6. List the factors that enter into an evaluation of a particular piece of equipment.
7. Sketch a "product manager"-oriented organizational chart.
8. Define the terms "CIM" and "CAM" and their use in a manufacturing setting.

Moore's law and the new wafer fabrication business

The semiconductor industry started supplying commercial products in the late 1940s. The manufacturing lines were little more than laboratories, and the workers were primarily trained technologists. By the 1970s, the manufacturing scene had changed to cleanrooms with highly specialized equipment attended by skilled production workers. A fabrication area of 2000 to 3000 ft^2 could be built for 2 to 3 million dollars.[1]

The VLSI era ushered in class 10 and 1 cleanrooms and newer, even more specialized and automated equipment and process control systems. The cost shot up to some 100 million dollars. By the mid 1990s, the industry was shipping large memory and microprocessor devices and facing 1 to 2 gigadollar (i.e., billion dollar) fab facility costs. The price tag is not so surprising, considering the factors driving the processes.

Advancements in integrated circuits progress along the prediction of Moore's law. Simply stated, it predicts a doubling of the number of transistors in a chip every 24 months. This leads to chips with higher capabilities and operating at faster speeds, as demanded by the marketplace. In reaction, the industry has developed strategies to keep pace with the higher challenges to the entire semiconductor production stream, from incoming materials to packaging.

The nano- and 300-mm era

Capability and speed advances are accomplished with driving chip component dimensions to the nano-level (one-billionth of a meter). However, smaller components crowd the chip surface, forcing additional layers of metal connectors above the chip surface, and more process steps. Smaller component sizes on the chip surface require shallower doping layers and thinner dielectric layers. These requirements also increase the number of process steps. Along with higher-density chips, capacity improvements lead to larger chip size. However, larger chips on the same diameter wafer result in lower productivity. The natural solution to this problem is larger-diameter wafers. The companion wafer diameter to the nano-era has become the 300 mm wafer (approximately 12 in).

The 300-mm diameter wafers have brought new production challenges and forced higher levels of automation. A number of factors drive the need for automation. At a practical level, the weight and value of a 25-wafer batch of 300-mm wafers in a carrier is very high. Asking operators to manually process this size wafer puts many dollars at risk. Maintaining productivity and yields with larger wafers creates higher costs. The equipment to process larger wafers with

tighter tolerances becomes more expensive. Improvements are needed to maintain uniformity across a larger wafer (or more tools processing single wafers). Of course, all of the materials and processing environment must become cleaner, since the smaller, more closely packed devices are more sensitive. And don't forget that, at these levels, the number of tests and characterizations increases to maintain product quality and process control in a wafer fabrication line that is rapidly moving large volumes of wafers. Of special note is the increase in process steps at the ULSI/nano level. Squeezing the devices closer together and making them smaller has introduced problems that were solved by the addition of new process steps, such as planarizing techniques to overcome topography generated imaging problems. The additional steps drive up the process and inventory expense. Some of these factors are noted in Fig 15.1, which shows the Semiconductor Industry Association's *International Technology Roadmap for Semiconductors.*

All of these cost increases have come as individual chip sale prices continue to erode through improved productivity and competition. To stay profitable, chip manufacturers have to continually improve their efficiency, yields and control costs at all levels. Reaching cost-effective operations generally means a throughput capability of 20,000 to 30,000 wafers per month. When the move to 300-mm wafers takes place, an advantage will be the option of smaller fabs (or *mini-fabs*) running 10,000 wafers per month and producing the same number of chips as a 200-mm fab.[2] This option comes from the larger area of the 300-mm wafers.

Despite all the changes in processes, costs, and markets, the overall financial measurement of a fabrication area has remained the same: it is *the cost per functioning die shipped out of fabrication.* When extended to a complete merchant facility with assembly capabilities, the

Year of Production	2001	2006	2012
Line width (nm)	150	100	50
Memory size	1 Gb	16Gb	64 Gb
Logic Bits/cm^2	380M	2.2B	17B
Chip Size-DRAM (mm^2)	445	790	1580
Max wiring levels	7	7-8	9
Mask layers	23	24/26	28
Defect density-DRAM (D/m)	875	490	250
Chip conections-I/Os	1195	1970	3585
Wafer diameter (mm)	300	300	450

Figure 15.1 Projection of wafer and chip parameters.

measure becomes *the cost per die shipped.* In the world of the megachips, the cost per transistor is becoming an indicator parameter.[3] These measurements apply to in-house wafer fabrication operations, foundry operations, and fabless chip companies.

The remainder of this chapter examines the various factors and strategies used to increase process/factory performance while controlling costs.

Wafer Fabrication Costs

A number of factors contribute to the cost of producing a functioning die (Fig. 15.2). They are generally divided into fixed and variable categories. Fixed costs are those that exist regardless of whether any die are being made or shipped. Variable costs are those that go up or down with the volume of product being produced.

Fixed costs include:

- Overhead—administration, facilities, and research
- Equipment
- Variable costs
- Materials—direct and indirect
- Labor
- Yield

Overhead

Overhead costs are all those incurred by the administrative and executive staff plus the cost of providing and maintaining a facility. A curious fact of company growth is that, beyond a certain level, the number of administrative personnel grows faster than the manufacturing

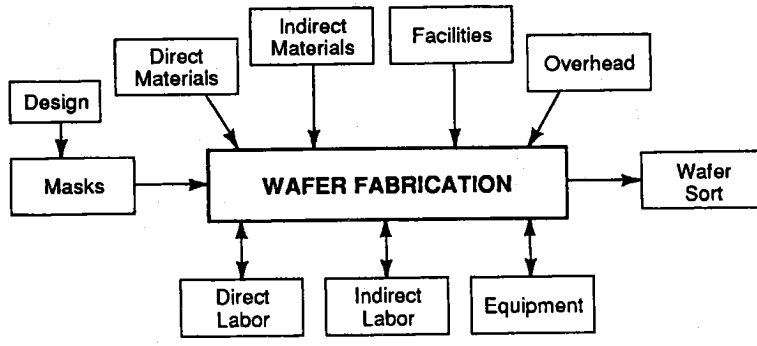


Figure 15.2 Fabrication cost factors.

workers. As companies grow, more information is generated internally, and more information must be handled from customers and suppliers.

To be effective, the information must be available to an ever growing staff. The two needs result in more and more staff processing "information" rather than product. Also, decision-making becomes more formalized (and costly) as more departments become stakeholders in the outcomes. Currently, some 50 percent of the workforce of the industrialized economies is involved in information processing.[4] A primary overhead expense is the design activity. With expensive CAD systems and a large professional design team, the cost of circuit design is considerable. Design cost goes up for ASIC fabrications, where a large number of individual circuits are designed and produced with a fast-turnaround requirement.

The facility cost and maintenance of the facility are major contributing costs. A fabrication area occupies only about 20 percent of the total facility area, yet it creates the majority of the expense. Air conditioning, chemical storage and delivery, and the cost of the cleanroom are all major expenses. Fabrication cleanroom costs for a ULSI facility are thousands of dollars per square foot. Actual floor costs are a factor of the cleanliness strategy chosen (see Chapter 4). A total cleanroom layout is more expensive than a hybrid/mini-environment approach. But in the latter, there is more expense for the equipment.

Co-development projects are emerging as the cost of product and process development increases. These partnerships are more important for processes whose effectiveness is heavily influenced by production factors. Often, a vendor cannot duplicate high-volume production parameters and needs access to wafer fab partners. Another trend is for semiconductor companies to co-develop new processes. As the industry has matured, more research is coming from universities and consortiums such as Semetech.

Materials

Manufacturing materials are divided into the categories of direct and indirect. Direct materials are those that go directly in or on the chip. This includes the wafer materials, the materials and chemicals needed to form the deposited layers and doped layers, and the packaging material costs. Indirect materials are the masks and reticles, chemicals, stationery supplies, and other materials that support the process but do not enter into the product.

Equipment

This cost is the equipment used directly in the fabrication of the devices and wafers. It shows up in the cost calculations as fixed overhead

or as depreciation. Depreciation is the loss of value of the machines as they wear out or become obsolete.

The transition from 200-mm wafers to 300-mm wafers has bumped up equipment costs. At the transfer stage, 300-mm processing is essentially the same as 200-mm processing. This means that the process tools are the same, except for size capability. However, the larger wafers generally require more handling time and more process time, resulting in productivity losses. These losses equate to more tools to maintain production quotas and more expense. Figure 15.3 compares the productivity of the primary process tools for the two diameters. New to the 300-mm process is greater need for in-line metrology measurement and monitoring. The downside of larger wafer diameters is bigger losses if wafers are misprocessed or have low yields. Now, 300-mm processing is proceeding with copper metallization, which brings surface factors and new low-k materials into the picture. Monitoring and controlling each of these steps at the process tool is critical if the whole system is to work at the end of the process. Critical dimension measurements and e-beam defect inspection systems have become in-process requirements and add to the equipment cost.

Interestingly, the move to larger wafers and more sophisticated processes has increased the typical life cycle of new processes. Figure 15.4 shows the history and projections of wafer size life cycles.

Labor

Labor also has a direct and indirect component. Direct labor takes in those workers actually handling the wafers and equipment. Indirect

Tool	200-mm baseline (1.0x)	300-mm productivity in 2001	Expected productivity in 2002, 2003	Major upgrade required
Photo	1.0x @ wafer passes/day	0.6x	0.9x	Yes
Etch	1.0x @ wph	1.0x	1.2x	No
Wet bench	1.0x @ wph	0.8x	0.9x	No
CVD	1.0x @ preventive maintenance	0.3x	1.0x	No
PVD	1.0x @ available time	0.9x	1.0x	No
CMP	1.0x @ available time	0.8x	1.0x	No
Furnace	1.0x @ wafers/day	0.7x	0.9x	No
Implant	1.0x @ wafers/day	1.0x	1.0x	No
Defect inspection	1.0x @ wph	0.6x	1.0x	Yes
Metrology	1.0x @ wph	0.7x	1.0x	Yes

Figure 15.3 300-mm tool productivity factors. *(Courtesy of* Semiconductor International, *August 2001.)*

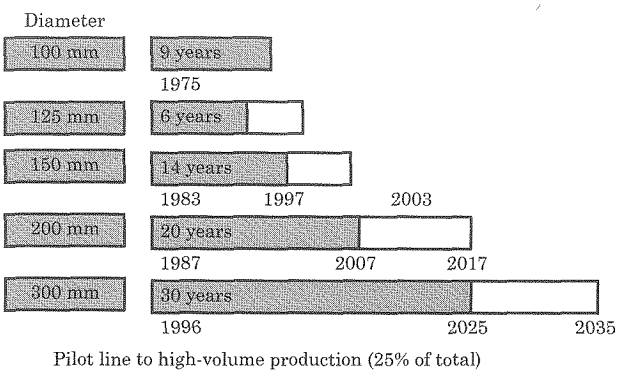


Figure 15.4 Wafer size life cycles. *(Courtesy of Future Fab International.)*

labor refers to support personnel such as supervisors, engineers, facility technicians, and office workers. Ironically, the new demands of process accuracy and productivity, using very sophisticated process tools, have returned the industry to the requirement that operators now have more technical training. Some companies, such as Intel, require all of their fabrication workers to have technician-level education. Coupled with higher levels of technicians and engineers to set up and maintain the tools, labor costs have increased.

Production cost factors

Typical contributions of each of the factors to the overall unyielded die projected for 300-mm diameter wafer costs are shown in the following table. The percentages will vary with the wafer and feature size, degree of automation, and number of process steps. Depreciation in the

Factor	Percent contribution	Costs ($)
Depreciation	35	1189
Labor	7	232
Maintenance	7	232
Consumables:		
direct materials	12	405
test wafers	6	203
indirect materials	26	890
Other	7	226
Total	100	$3378

chip industry is high as a result of the short useful lifetime of the expensive equipment. Direct materials include wafers. Others include the overhead costs.

In any discussion about wafer cost factors, one is naturally curious about the actual costs. This number changes from line to line, with the complexity of the devices and with the position of the product in the maturity cycle (Fig. 15.5). The more mature the product and process, the higher the yields and the lower the equipment depreciation factor. Newer products suffer from operator learning curves, equipment shakedown times, and development of new processes. Wafer volume and type are major influences on cost. High-volume products, such as DRAMS, generally have the lowest per-transistor or wafer cost because of high manufacturing efficiencies. ASIC wafers, on the other hand, have higher costs as a result of smaller production runs and the higher design and processing costs required for a varied product mix. The range of costs is from about $100 for discrete and simple ICs to several thousand dollars for advanced products.

Yield

The overall fabrication yield (see Chapter 6) determines how the various costs affect the final die cost. If the die yield is low, the cost per die goes up. Not only are the fixed costs distributed over fewer die, but the variable costs go up as more materials are required to get out the die. When the yield is calculated into the cost, the term used is *yielded die cost.* The cost of producing the wafers without considering die yield is the *unyielded die cost.*

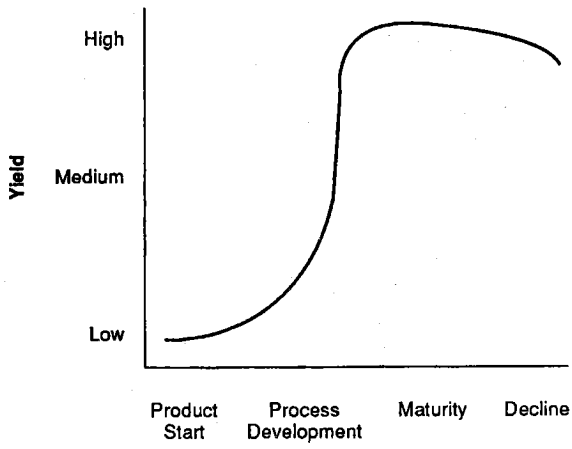


Figure 15.5 Wafer production loss versus product maturity.

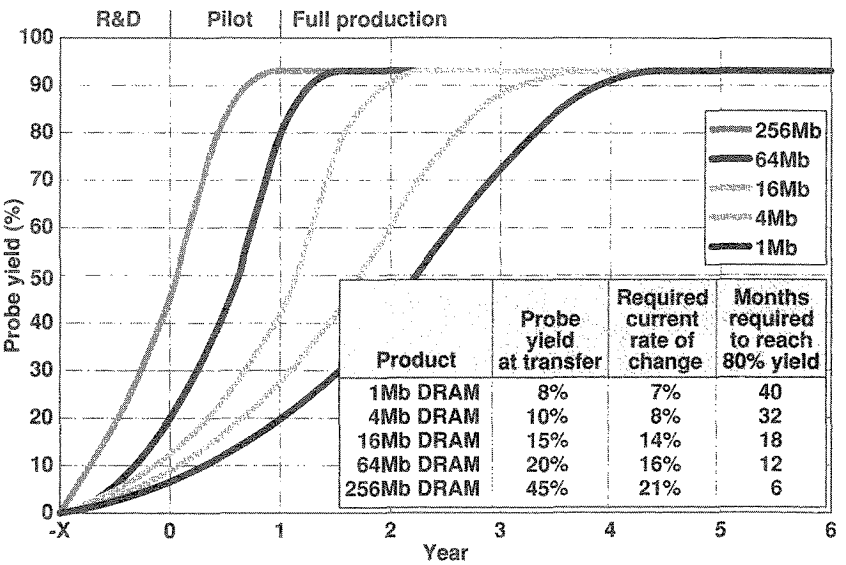
Die costs are a function of the wafer size, die size, and wafer sort yield. A $3000 wafer-manufacturing cost for a wafer with 300 die will translate into an unyielded die cost of $10 each. If the die sort yield is 50 percent, the die cost rises to $20 each ($3000 wafer cost divided by 150 functioning die). Increasing the sort yield to 90 percent reduces the die cost to $11.11.

Market pressures require that wafer fabrication operations reach wafer sort yields in the low 90 percent range at ever faster rates. Figure 15.6 shows the yield ramps for different levels of DRAM memory devices as they progress from R&D to full production.

Yield improvements

While it is true that increased attention is being given to traditional business factors, no less attention is paid to process and wafer yields. The effect of a yield improvement can be significant in terms of dollars. Consider the yield figures in Fig. 15.7. Line 2 shows that a wafer fabrication yield improvement of five percentage points increases the overall yield from 38 to 40.4 percent, a 1.2 percent increase. If the fabrication area starts 10,000 wafers per month and there are 350 die per wafer and the selling price is $5 per die, the increased revenue is

$$10{,}000 \text{ wafer starts} \times 350 \text{ die per wafer} \times 0.012\ (1.2\%) = \$210{,}000/\text{month}$$



Product	Probe yield at transfer	Required current rate of change	Months required to reach 80% yield
1Mb DRAM	8%	7%	40
4Mb DRAM	10%	8%	32
16Mb DRAM	15%	14%	18
64Mb DRAM	20%	16%	12
256Mb DRAM	45%	21%	6

Figure 15.6 Probe yields for DRAM circuits. *(Courtesy of* Semiconductor International, *January 1998.)*

Process	Change	Yield		Assembly	Overall
		Fab	Sort		
Base Line		0.80	0.50	0.95	0.38 (38%)
	Fab to 0.81	0.81	0.50	0.95	0.384 (38.4%)
	Fab to 0.85	0.85	0.50	0.95	0.404 (40.4%)

Figure 15.7 Yield impact of fabrication yield improvement.

Whether this amount of increased income is significant depends on the cost of effecting the improvement.

Yield and productivity

Yield has been the traditional measure of wafer fab success. High yields translated to lower production costs and higher profits. With wafer sort yields at the 90 percent level, the next cost-reduction factor is productivity. Improved production efficiency is measured by the number of wafers per square foot of the total fab area (or number of wafers per piece of equipment), or the number of wafers produced per operator. Increasing these factors means a lower manufacturing cost. Throughput is another factor. Increasing the number of wafers out per hour means a more efficient process with lower costs. These factors increase the bottom line if the cost of achieving them is less than the advantage.

Book-to-bill ratio

On the business side, one of the most watched industry factors is the book-to-bill ratio (b/b). *Book* refers to bookings, which is the dollar amount of sales orders received in a period. *Bill* refers to the dollar amount of billing invoices sent out. In a normal operation, chip orders are shipped and billed some time after the order is received. In good times, the delay reaches into months or longer if shortages occur. A high book-to-bill means that unfilled orders are accumulating, generally indicating a healthy economy. A low book-to-bill means that orders are drying up, and the factory is shipping out to inventory rather than building new product. This indicator is used to project inventory and production levels. A b/b ratio greater than 1 indicates a rising market, and less than 1 indicates a declining market. When the market becomes very good, a high b/b ratio is worrisome, since shipping times become stretched, and the user end of the electronic industry can become starved for chips.

Manufacturing strategies

Minimizing the cost per wafer (or transistor) involves a host of factors that are the result of decisions about how the fab area is laid out, contamination control systems, equipment/automation levels, process control programs, and material management schemes. These are discussed in the following sections.

Contamination control systems

There are two major contamination control systems in use (see Chapter 4). Some manufacturers elect to build and maintain a total cleanroom environment with class 1 or better air quality. This approach has the highest facility expense and requires rigorous maintenance and contamination control programs. Wafer isolation technology (WIT), or the use of mini-environments, is the other approach. It requires a large investment in clean boxes and equipment interfaces, but the fab area can be at a lower class value.

Equipment

Over the years, semiconductor processing has changed from a laboratory activity to full manufacturing using sophisticated and dedicated equipment. The development of semiconductor-specific equipment has allowed the advancement into the ULSI era and the spread of the technology throughout the world. IBM coined the term *tool* for semiconductor process machines, and it has become part of the industry vocabulary. As the tools have become more sophisticated, the prices have risen to the point where they are a major expense in any new facility. An ASA/Dataquest calculation shows a worldwide chip business of $804 billion and an equipment/materials business of $77 billion.[6]

Equipment factors

A number of considerations (Fig. 15.8) go into the selection of a particular process tool. They can be roughly divided into two broad categories: performance and economic. Performance factors relate to the ability of the tool to produce the required results. Economic factors relate to price, costs, and support factors. Figure 15.7 shows how these four factors have shifted as the industry has moved into the manufacturing era.

Performance has fallen as a selection criteria; nevertheless, any tool must have the basic capability to meet the process requirements. If the machine cannot routinely produce the right product, it is of no use. Other performance factors are repeatability, flexibility, upgrad-

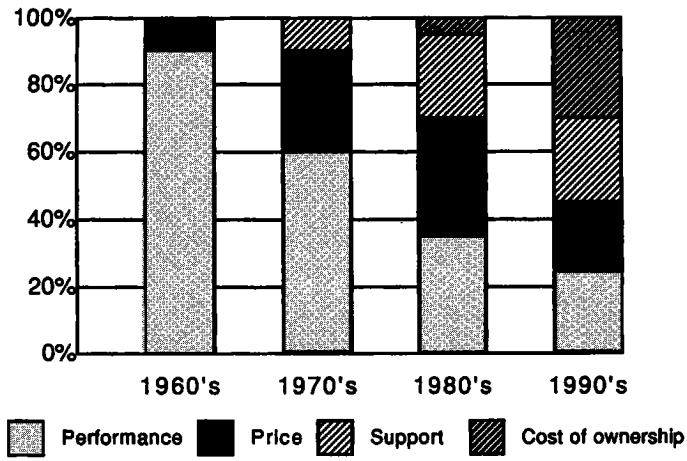


Figure 15.8 Changes in equipment purchase factors. *(Source:* Semiconductor International, *May 1993, p. 58.)*

ability, ease of operation, setup considerations, and reliability/downtime factors. Repeatability is the ability to produce the same result every time and over long periods of time. Flexibility relates to the ease with which a machine can be switched to run a variety of products and processes. This is an important factor for ASIC and other small-run fab lines. Upgradability is the ability of the machine to handle future process requirements, such as increasing wafer diameters. For example, a lot of 300-mm process tools are basic 200-mm versions, retrofitted to handle the larger diameter. This factor is very important for fabrication lines doing ASIC work, where staying abreast of the latest product technology is vital. In fabrication areas dedicated to long production runs of the same (or similar) products, the ability to upgrade is less important.

Ease of operation and setup factors address the issues of minimizing operator mistakes by good design and user friendly controls. Setup issues relate to the amount of time required to bring the tool on line (tests, calibrations, and so on) and the related loss of production time.

Most companies run their equipment two to three shifts per day, sometimes six or seven days a week. Scheduled maintenance and unscheduled breakdowns stall product flow and run up expenses. The factors are: *scheduled maintenance frequency and time, mean time to failure* (MTF), and *mean time to repair* (MTR). Given the expense of most machines, having backups is a costly luxury. It falls on the vendor to provide a machine that runs for long periods of time (long MTF), can be repaired quickly (short MTR) when failure occurs, and does not

need frequent and lengthy routine maintenance. Cost also includes the price of materials needed for the machine process. Machines that waste materials also waste money.

Vendor support has emerged as a critical factor. The advancement of process technology and equipment sophistication has forced close cooperation and alliances between the chip manufacturers and their suppliers. Reducing the cost of machine development requires input from the operating process engineers, and the detailed nature of the machines requires on-time backup from the vendors. Management guru Tom Peters sees modern corporations as having "fuzzy edges." By this he means that, more and more, a corporation must rely on its vendors and customers for information and co-development of needed equipment and materials. This is certainly true of the semiconductor industry.

Cost of Ownership

Shifting equipment purchase decisions away from purely technical to business-related factors has driven the development of cost of ownership (COO) models. These models attempt to bring together all of the relevant factors driving the total cost of ownership of a tool, process, or facility over its expected lifetime. Beyond the initial equipment purchase price, tools differ in the amount of expensive floor space they occupy, the amount of power and materials they require for operation, the yield of in-spec wafers, maintenance, repair and failure rates, and so forth. The COO formula is one developed by Semetech to evaluate equipment purchases.[7]

$$C_W = \frac{\$F + \$V + \$Y}{L \times \text{TPT} \times Y_{\text{TPT}} \times U}$$

where C_W = cost of the finished wafer, based on the particular cost associated with the tool/process for the lifetime of the tool
$\$F$ = fixed costs, which include the equipment purchase prices, facility costs, initial modifications, and so forth
$\$V$ = all material, labor, and process costs generated when the tool/process is operating
$\$Y$ = the cost of wafers scrapped due to tool scrap and defect induced losses
L = the lifetime of the tool in hours
TPT = the wafer throughput rate as reduced from the maximum by maintenance requirements, setup, test wafer monitoring, and so on, expressed in wafers per hour
U = the tool utilization factors that reduce available process time from the maximum

Each of these equation terms is calculated from formulas that consider the subfactors for a particular process. The COO formula provides the method to determine trade-offs for various tool factors. For example, a tool may have a high initial cost that is offset by lowered operating costs or a long time between failures. Or, a tool may provide a high yield but require so much adjustment and calibration that additional machines would be required to meet production schedules.

Automation

As most industries mature, the technology becomes stabilized, and the market drives up demand. These two conditions are precursors to process automation. Since 1940, automation of oil refineries has reduced the number of workers by a factor of 5.[8] Automation of semiconductor processes has been in progress since the 1970s, when process tools were designed to accept wafers in cassettes. Since then, automation has marched along toward the fabled dream of the total "peopleless lights-out" fab. Automation stages start with the process tool and extend to the factory level.

Process automation

The first level of automation is of the process itself. Most semiconductor equipment, by definition, automates a part of a process. Photoresist spinners automatically dispense the primer and resist at the correct speeds and for the correct time. Automatic gas-flow controllers dispense gases to the tool in the right amount, at the right pressure, and for the right time. Process automation brings consistency to the process and the product by reducing reliance on operator skills, training, morale, and fatigue.

Most tools are controlled by a set of instructions programmed in an on-board computer. The program is called a *recipe*. The recipe is loaded into the machine by the operator or from a central host computer.

Wafer-loading automation

The next level of automation is the loading and unloading of the wafers. The industry has settled on the wafer cassette as the primary wafer holder and transfer vehicle. Cassettes are placed on the machine by various mechanisms. Elevators and/or wafer extractors or robots feed the wafers into the particular process chamber, spin chuck, and so on. In some processes, such as process tubes, the entire cassette is placed in the process chamber. This level of automation is referred to

as the "one-button" operation. With one button, the operator activates the loading system, and the wafers are processed and returned to the cassette. At the end of the cycle, the machine sounds an alarm or turns on a light, and the operator removes the cassette.

Some machines have buffer storage systems that maximize the machine efficiency by always having fresh wafers (or reticles for the imaging tools) available for processing. These are called *stockers*. The operator places the cassette(s) on the machine loader and pushes a start button, after which the machine takes over the processing.

At the 300-mm diameter wafer level, and with single-wafer processing, wafers are transported in individual holders.

Clustering

Mating two or more process steps in a single unit is another level of automation. Generally, this level is called *clustering*. The industry has been "clustering" for a long time. Photoresist spinners were long ago mated to soft-bake modules and other track equipment groupings.

Recent clustering designs (Fig. 15.9) have been driven by both technical and economic forces. On the technical side, some clustered processes make better product by, e.g., keeping a silicon wafer clean after etch and before metal deposition. Another process advantage is sequential deposition of different materials in the same chamber. In these cases, the deposition process is better, because the wafer is not exposed to air between steps. Certainly, anytime a wafer loading/unloading step can be eliminated, both cleanliness and cost are favorably effected. For vacuum processes, time and cleanliness factors are affected when two or more processes can be performed with only one vacuum pump down. Clustering in which two or more sequential processes are performed is called *integrated processing*.[9] On the economic

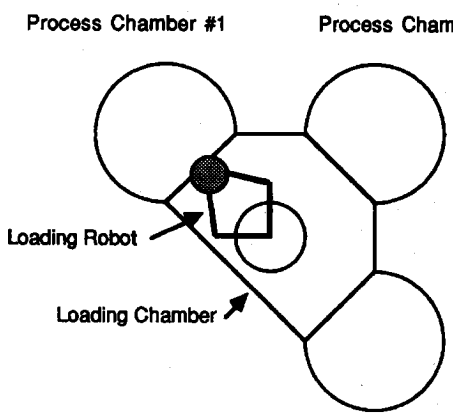


Figure 15.9 Three-chamber cluster tool.

side, some types of processes that are clustered for increased throughput are called *parallel processing.*

Despite the obvious attractions of clustering, there are drawbacks. Clustering for critical process advantages is easy to justify. But a cluster of same type processes requires interlocks, electronics, and software more sophisticated than individual tools. And a shutdown for maintenance or repair idles a larger part of the production capacity. Another barrier occurs when cluster modules cannot be provided by the same vendor. Customers prefer one responsible vendor, and vendors are slow to pair up when responsibilities are clouded.

Wafer-delivery automation

The third level of automation is when wafers are automatically brought to, loaded on, and removed from the machines. In addition to the production advantages of automation, there are ergonomic, safety, and cost benefits. These accrue from the weight of a cassette of larger diameter wafers, which can reach the 18 to 20 lb (6.7 to 7.5 kg) level. Operators injuries are a possibility, and the financial loss of dropping a cassette of expensive wafers can be staggering. Early delivery systems used traveling robotic carts that duplicated human delivery (Fig. 15.10). Called *automated guided vehicles* (AGVs), the carts travel along the aisles and dispense wafer cassettes when the machines need them. Another version is the rail guided vehicle (RGV), where carts follow a track laid out between the process tools and the stockers. These approaches have the advantage of being retrofittable to fabrication lines where the equipment is lined up in rows.

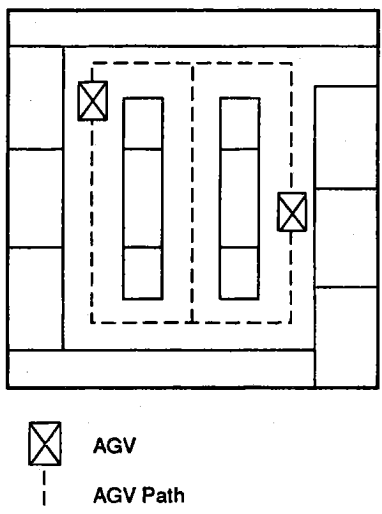


Figure 15.10 Automated guided vehicle wafer delivery system.

Another approach is the use of an overhead rail[10] (also called a gantry). The wafer cassettes arrive at the process tool area (Fig. 15.11) where a secondary system (usually a robot) removes them from the overhead rail and places them in the tool buffer. This system works best with machines that are grouped in bays rather than in the traditional linear layout.

Closed-loop control-system automation

The industry is just embarking on the last level of total automation, the closed-loop feedback system. There are two aspects. Some tools have on-board sensors that measure critical process parameters. Through feedback electronic circuits, the tool adjusts itself to maintain process operation specifications. This stage of automation is difficult to achieve for many processes. It requires measuring sensors that operate in hostile environments, such as heated-tube furnaces and sputtering chambers. One area of progress is in the area of mask-making, where recognition systems are advanced enough to compare the manufactured plates with the design criteria.

Second are tools that have, or are connected to, automatic inspection or measuring subsystems. The subsystem measures important parameters in real time (as they are happening), compares them with a standard, and feeds the information back. It may trigger a shutdown of the out-of-spec tool or make process parameter adjustments. The development of closed-loop process and machine control is necessary before the goal of a "peopleless, lights-out" wafer fab can be achieved.

Factory-Level Automation

The advent of higher levels of process and tool automation and inventory control systems requires higher levels of centralized control and information sharing. Most companies have computer-based management information systems (MISs) handling the paperwork and details

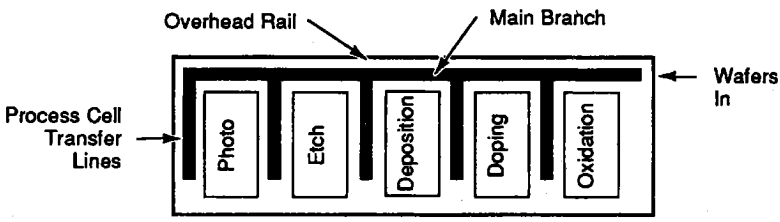


Figure 15.11 Overhead rail wafer delivery system.

of employment and finances. These systems are being expanded to the entire manufacturing environment in a process called *computer-integrated manufacturing* (CIM).

CIM is the computerization of all plant operations and the integration of those operations into one computer design, control, and distribution system. The processes involved in CIM are all related and interdependent, as illustrated in Fig. 15.12. The major activities of CIM are business functions, product design (mask and circuit), manufacturing planning (inventory, shop floor priorities, etc.), manufacturing control, and the fabrication processes. A complete CIM system is interactive at all levels. This means that each of the five functional areas input data to the system in real time and that the information is available to all who need it.

Computer-aided design (CAD) and computer-aided manufacturing (CAM) are two subsystems within the CIM system. The role of CAD has been discussed in terms of mask and circuit design. CAM is the part of the system that does the planning and control of the manufacturing operation. A CAM system includes a computer network and the automated process tools and material delivery systems.

In concept, a CIM system kicks into operation when a customer order is received. The computer logs the order and initiates the CAD system to start the design (if it is a custom order). It also triggers the ordering of needed materials in the right quantities and schedules

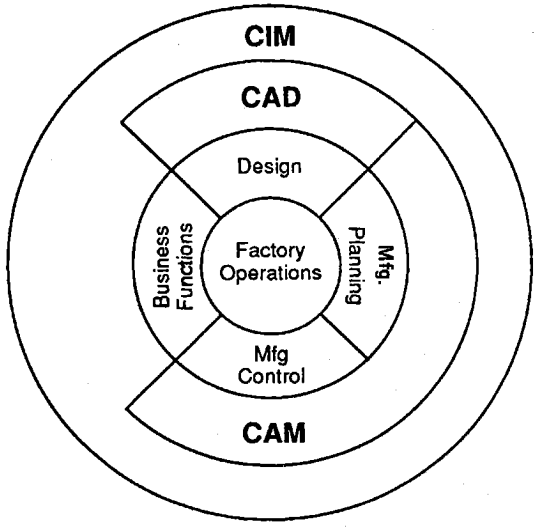


Figure 15.12 Provinces of factory computer control systems.

their delivery times. This subsystem goes by the name *computer-aided process planning* (CAPP). Through the CAM program, the individual process recipes are downloaded to the individual equipment computers. Once processing begins, the CAM system controls the WIP and makes necessary priority decisions to meet shipping schedules. It also keeps track of equipment performance and schedules repairs and maintenance.

An important feature of the CAM system is yield monitoring and reporting. Important measuring systems are connected directly to the factory computer. If a poor yield problem appears, the system reports it to the engineering and facility staff and (if necessary) will reorder materials to make up the losses. At the end of the production run, the system calculates costs and yields and schedules shipment to the customer.

Other packages included in CIM systems are facility monitoring, process modeling, and security systems. Facility monitoring might include power consumption and environmental factors inside the plant. Some facility CAM systems include monitoring the levels of liquids and gases in storage with automatic alerting and/or reorder. Monitoring and precise control of these factors can save appreciable amounts of money. Process modeling is a system of testing a particular design against the known process variations. The computer can run many variations that simulate the changes expected in fabrication. Good modeling can identify weak points in the design or process before wafers are committed to the line. Security monitoring may include the entering and exiting of employees (and/or intruders) and the securing of expensive finished product or materials. A security system may also include fire and other hazard control.

Equipment Standards

Given the many suppliers of materials and equipment and the very stringent technical demands of chip manufacturers, the need for standards is obvious. In many industries, different manufacturers attempt to establish their "standard" as the industry standard, as a way of maintaining a competitive edge. Fortunately, the semiconductor industry informally has settled on many standards—for example the, use standardized cassettes.

In 1973, Semiconductor Equipment and Materials International (SEMI) established a standards program. Supplier and user personnel come together and establish standards by a consensus process.

At the automation level, SEMI has published communication protocols for equipment interfaces.

Fab floor layout

The tradition fab floor layout of lining up the equipment in rows has given way to the bay or tunnel system. This system groups sections of the process in separate rooms, which minimizes contamination from personnel traffic and cross contamination from other process areas. Automation needs is forcing a rethinking of the most efficient way to move material while maintaining cleanliness.

Robots and overhead gantries have become the material mover of choice. Also driving robot use is the increasing weight of wafer lots (larger diameters), especially in SMIF type transfer modules.[12] Robotic design and performance challenges come to the forefront in vacuum tools, especially if the process uses corrosive gases.

One equipment layout option is the process island. The various tools for a particular process segment are grouped around a single loading/unloading robot (Figure. 15.13). The tools may be single or clusters.

Batch versus single-wafer processing

For the first two decades of process development, the drive was to ever larger batches of wafers. Simple economies of scale were behind the drive, i.e., gaining more product output per process sequence. However, the arrival of larger-diameter wafers and VLSI/ULSI-level circuits has changed the picture. Larger wafers require larger process chambers, and this, in turn, pushes the limits of uniformity. Many processes, particularly the deposition steps, are easier to control in a smaller (single-wafer) chamber. Fast cycle time fabs (ASIC and small-

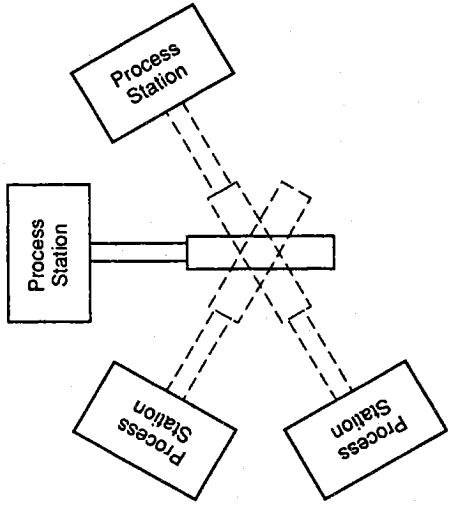


Figure 15.13 Robot wafer delivery in process island.

volume devices) can use the flexibility of single-wafer processing to good advantage.[13]

Offsetting improved process control is a loss of productivity compared to batch processing. Larger-diameter wafers help somewhat (300-mm diameter wafers have almost four times the area of 150-mm wafers). Single-wafer process productivity is improved by tool designs that, e.g., allow one vacuum pump down for an entire batch while wafers are processed individually, and other improvements.

On the downside, single-wafer processing requires higher levels of repeatability. Assuming the same size wafers, a single-wafer processor will have to perform 25 times compared to a one-batch 25-wafer batch process. The equipment is also more expensive. Each single-wafer processor requires essentially the same components as a batch tool.

Green fabs

Additional pressure on wafer fabrication processes is generated by environmental laws and environmental concerns. A wafer fabrication operation uses many hazardous chemicals and produces waste products. In-plant controls of chemical handling, storage, and use are a part of fab *environmental, safety, and health* (ESH) programs. Active research is aimed at development of processes that use less chemistry and chemicals that are less harmful. The SIA roadmap shows water consumption dropping from 30 gal/wafer to 2 gal/wafer in 2012.[14] The roadmap calls for lowered energy use per wafer and the elimination of polyvinyl chlorides (PVCs). Along with safety, there are cost savings related to lower-hazard chemicals in terms of the costs of on-site treatment and storage, transportation, and acceptance at special waste facilities.

Statistical Process Control (SPC)

This text, for the most part, has presented the process technology with few mathematical formulas. The next few sections on statistical process control will break from that approach. SPC is a powerful and needed tool to maintain process control and improve yields. Unfortunately, SPC is based on statistics, which is applied with the language of mathematics. We will attempt to convey the background, use, and interpretation of commonly used statistical tools with illustrative charts and graphs. No formulas will be presented.

The first question to address is, "What is the object of process control?" The answer is simple: to produce a product that falls within its design and operational specifications and do it at a high enough rate to be profitable. Challenging this simple goal are processes that are in-

fluenced by a multitude of parameters, including the conditions already on the wafer.

Processes do not run in control without monitoring and adjustments. Process control provides the information to make the necessary changes. Statistical process control does it in a way that relies on established mathematical principles that govern the type of processes used to make chips. Process control techniques can vary from simple to very complex. The simplest (and most familiar) is the calculation of the average of a group of numbers (called a *population*). We also know that the extremes of a population (range), along with the average, give us an idea of the distribution of the data within the population.

For example, the two groups in columns A and B of Fig. 15.14 have the same average. If the numbers were the sheet resistances of wafers from furnace A and B, we could easily come to the decision that furnace A has the most in-control process because of its tighter distribution. This fact is illustrated when the data is plotted (Fig. 15.15). The plots are called *histograms* and visually display data distributions that a simple average calculation will not reveal. Average calculations and histograms are statistics in action. Histograms are usually the first step in determining whether a process is in control.

Their power comes from a mathematical distribution known as the Gaussian distribution. Named after the famed mathematician Karl Friedrich Gauss (1777–1855), its origin is interesting. Gauss set out to reconcile the different star positions reported by different astronomers. His approach was to make all the necessary corrections to the observations, taking into account the time of the year and from where on the Earth the observations were made. He expected that, when all the corrections were made, there would be agreement from all the position calculations on the position of a particular star. After

Reading no.	Run A	Run B
1.	25	28
2.	24	25
3.	26	23
4.	23	26
5.	25	25
6.	26	22
7.	25	23
8.	24	25
9.	25	27
10.	26	25
Average	24.9	24.9
Range	26 - 23 = 3	28 - 22 = 6

Figure 15.14 Sheet resistance reading (Ω/❑).

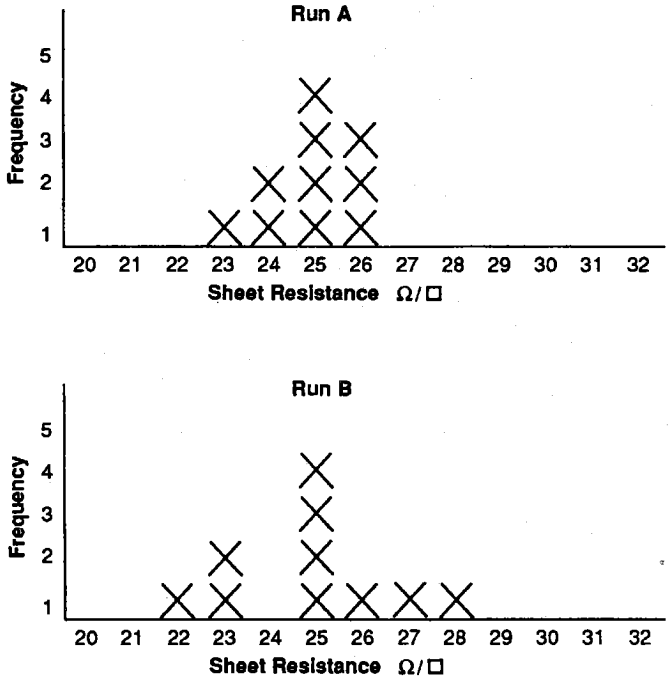


Figure 15.15 Frequency distribution of sheet resistances.

all, reason dictates that a star can occupy only one position at a time, and we should be able to determine that position.

However, the final data did not confirm his hypothesis. After all corrections were made, there were still a number of locations for each star. Fortunately, Gauss did not scrap the project but went on to establish the basis for the field of statistics and distribution probabilities. If the math-aversion readers will hang on, they will find that the concept of probability is not so difficult. The way Gauss analyzed the data was to plot the various calculated locations for a star. He calculated the center point (average) and drew a circle that encompassed the star location the farthest from the center. He reasoned that there was a 100 percent probability (a probability of 1) that the real star location was within that circle. He also reasoned that the probability of finding the star within the confines of smaller circles was less than 1. In fact, the smaller the circle, the lower the probability that the star was within those boundaries.

It also turns out that processes produce data distributed like the example given. The mathematical distribution is known as the Gaussian distribution. A good example is the height of blades of grass in a lawn.

If all the blade heights are measured and plotted on a histogram, the distribution will be the familiar bell-shaped curve, also called a *normal curve* (Fig. 15.16). From a probability consideration, it predicts that there is a higher probability that any given blade of grass will have a height closer to the average (center values) and that there is a lower probability that any given blade will be very short or very tall. The mathematical conditions that result in the distribution include human heights, IQ distributions, and (in most cases) semiconductor process parameter distributions, such as sheet resistance.

A first step in process control is to make a histogram of the particular process parameter and determine if the distribution is a normal distribution. If it is not, the chances are good that there is something wrong in the process. If the distribution is a normal one, the next step is to compare the range of the distribution with the design limits for the particular parameter (Fig. 15.16). This comparison is made to determine if the natural process distribution limits fall within the design limits. If they do not, the process must be fixed or some percentage of the parameter readings (and the wafers) will always be out of specification.

Another useful statistical tool is the Pareto chart. This is a form of the histogram, but with the x axis divided into unrelated sections rather than a continuum of some parameter such as sheet resistance. Defect inspection results are good candidates for Pareto charts. On the x axis is a list of the defect categories. A mark or vertical column over each defect type indicates the frequency of occurrence of each of the defect types. The operator or process engineer can see from the chart which defects are occurring most often and what processes must be improved.

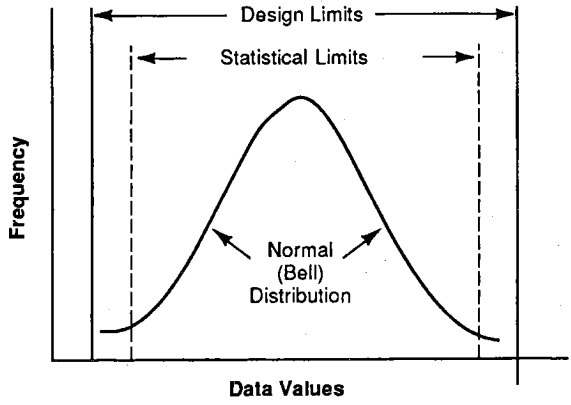


Figure 15.16 Frequency distribution of sheet resistances.

So far, the statistical methods explained are the after-the-fact histories of a process. A most powerful statistic of real-time process control is the *X-R* control chart (Fig. 15.17). The chart is constructed in two plots. The lower plot shows the moving range *R,* which is descriptive of process stability and the parts, with the *y* axis representing the parameter values. The top graph has a horizontal line for the historic average of the parameter. On each side of the average are control limit lines that are calculated from the historic data.[15] A control limit represents the limits that the individual data values will range between when the process is in control. Also on the chart are the process or design limits that represent the extremes the individual data points may have before being rejected. The bottom graph is constructed by calculating and plotting the amount that each data point varies from the average. When plotted, these values give further visual evidence of the amount of control in the process.

The value of the *X–R* bar control charts is their predictive powers. A process in control will produce data points that tend to vary in a regular pattern about the average (top of Fig. 15.17). The mathematics of a controlled process predicts this regular fluctuation. It also predicts *when* a process is going out of control *before the data points exceed the control limits* (bottom of Fig. 15.17). The data points in part B have shifted to the top of the control limit range. This is an unnatural pattern for an in-control process. When this situation occurs, the produc-

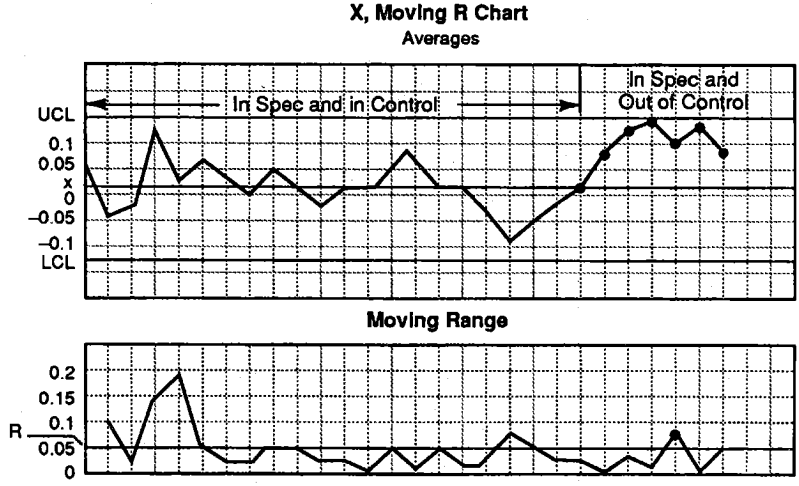


Figure 15.17 Moving *R* chart contains the averages of measurements, *x*, in the upper plot. The lower plot shows the moving range, *R*, which is descriptive of process stability.

tion operators, who maintain the charts as the data is produced, alert the proper personnel so the process can be brought back into control *before the data points exceed the control or design limits* and wafers have to be scrapped. A number of more sophisticated controls used in processing are beyond the scope of this text.

Another powerful statistical tool is multivariable experiment analysis. Most measured quality control parameters (sheet resistance, line width, junction depth, and so on) are influenced by a number of variables in the process. Line width, for example, varies with the resist solution, film thickness, exposure radiation time and intensity, baking temperatures, and etch factors. Any one or all can contribute to an out-of-spec condition. Multivariable evaluations allow the engineer to run tests that separate and identify the contribution(s) of each of the individual variables.

Designing an SPC system for a process requires selection of the proper statistical tool. Another decision revolves around the proper "indicator" population. Profit demands that all the die on all the wafers from all the batches, day in and day out, meet the specifications. However, picking the parameter population is not always an easy chore. Depending on the process, there are variations across the wafer, there are variations from wafer to wafer within a batch, and there are variations from tool to tool. Since every die cannot be measured, selecting the right sample point and sampling level becomes a demanding task.[16]

Inventory Control

A critical issue in fabrication cost control and yield is inventory level and control. As the number of processing steps has increased, so has the length of the processing time and the number of wafers in process (WIP). The problem is that the company pays for the wafers when purchased and doesn't receive payment until the finished devices are shipped. This period can vary from two to eight months for a production line making similar circuits. Fabrication lines doing ASIC circuits have an even heavier burden when many different circuit types are going through the system. To get an idea of the burden, consider a CMOS-type process with 50 major steps and 4 substeps each, for a total of 200 processes. The high cost of the equipment generally requires some buffer inventory to ensure that the machines are operating at maximum efficiency. If each buffer has 4 cassettes of 25 wafers each, the total inventory of WIP is 40,000 wafers. At a cost of $100 per wafer (large diameter), the total inventory burden becomes $4 million.

Excessive WIP affects productivity by having the capability of hiding process and equipment problems.[17] With a lot of inventory at the

stations, wafers can keep flowing out the back end while parts of the process are shut down. With a lower WIP, these problems are readily apparent and force the solution of problems. WIP also influences the overall fabrication yield. The collective experience of the industry is that the longer wafers are in the process the lower their wafer sort yield.

Just-in-time (JIT) inventory control

Just-in-time inventory control is a philosophy based on the objective of, "Make only what is required, only as required."[18] The system is simple in concept. All buffer inventories are reduced to an absolute minimum, from the storeroom to the machine buffers. To work effectively, excellent vendor relationships must be established. The incoming materials have to be of the highest quality, as JIT leaves little cushion for extensive incoming inspections and returns. Second, the vendor is asked to maintain ready-to-ship materials at its facility. In effect, the vendors are asked to hold the inventory previously held by the chip manufacturer, a situation that they are not always happy about. The chip manufacturers have to be very efficient in assessing their quantity, quality, and delivery needs, and must have a system that expedites material to the proper process tool and detects quality problems quickly.

JIT also has applications in the process flow. Some companies will commit wafers to a process only if the work can flow unimpeded through all of the substeps. This system gives higher yields and, if on a properly balanced line, increases throughput time, even though parts of the line are idle. This system is called *demand-pull*. Wafers "upstream" are worked on only when there is demand from downstream process stations about to run dry of wafers.

An effective JIT procedure can reduce the number of operators in the fabrication area, since a good proportion of their time is spent sorting, staging, and delivering work to the process tools. Fabrication-area layouts can change from the traditional linear arrangement of process stations. A linear layout benefits a line making only a few different product types. Wafers are introduced into the front of the line and physically move on a first in-first out (FIFO) basis. Problems arise when particular lots of wafers must be moved quickly. Usually tagged as *hot lots*, they are given priority processing at each station. Besides the control problems associated with keeping track of the lots, hot lots have the effect of making regular product sit in a queue waiting for the hot lots to clear. This layout is particularly cumbersome for ASIC lines where shipping dates and product types have constantly changing priorities.

JIT-CAM systems offer the advantage of knowing where all the lots are in the production cycle. The computer can stage the work to minimize disruption and keep a steady flow of product. Furthermore, JIT-CAM systems teamed with automatic delivery systems can be more efficient when the process stations are grouped rather than strung out in a line. This concept, called *work cells*, is more efficient, especially when a cell can run a number of different products. When a machine is inoperable in a linear layout, the line comes to a halt, and inventory builds up in front of the machine that is down.

Quality Control and Certification—ISO 9000

SPC and other product/process/personnel quality programs fall under the general category of quality control. In the semiconductor industry, there are two broad quality segments: quality control (QC) and quality assurance (QA). QC generally refers to all of the techniques used to monitor and controls the *process* and wafer quality while they are in the process. QA refers to the efforts to monitor and assure that the customer is receiving product that meets required specifications.

In 1987, a system of quality requirements for organizations was introduced: the ISO 9000 series. Developed by the International Organization for Standardization (ISO), the requirements are more like guidelines, sometimes called an "umbrella standard,"[19] rather than a strict cookbook set of standards.

The ISO 9000 guidelines/standards guide the development and critique of a complete quality system for product, process, and management. Individual companies decide on how best to comply and implement programs. With certification, customers and vendors know that there is a comprehensive quality program backing the product. The European Community (EC) requires that all companies doing business in the EC have ISO 9000 certification. In the U.S., the American National Standards Institute (ANSI) and American Society for Quality Control (ASQC) have developed equivalent standards. Assisting semiconductor suppliers is a program sponsored by the Semiconductor Equipment and Materials International (SEMI).[20]

Line Organization

Most fabrication areas are organized around the product-line concept. In this concept, fabrication areas are built to accommodate products with similar processing needs. Thus, there are bipolar lines and CMOS lines, and so forth. This arrangement makes for more efficient processing, since most of the machines are in use most of the time, and the staff can gain experience in processing a few products.

The staff of these lines are also fairly self-contained. The primary responsibility falls to the fabrication or product manager (Fig. 15.18). Reporting to this individual are an engineering supervisor, a production manager (or general supervisor), a design department, and the equipment maintenance group. The production manager is responsible for producing the finished wafers to specification, to cost, and to schedule. The engineering group is responsible for the developing of high-yield processes, documentation of the processes, and the daily sustaining of the line process. Both the production and engineering staffs are divided into groups focusing on a particular part of the process. This organization has the virtue of high focus on the fabrication area's primary goal of producing chips at a profitable level.

As the processes become more automated and arranged in process cells, small group organizational teams and responsibilities are emerging. A cell is attended by the operator(s), the equipment technicians, and the process engineer(s). These small groups make floor-level decisions with the information provided by the CIM system. However, few companies have formalized this arrangement with an organization structure, and the teams tend to exist as cross-department cooperatives.

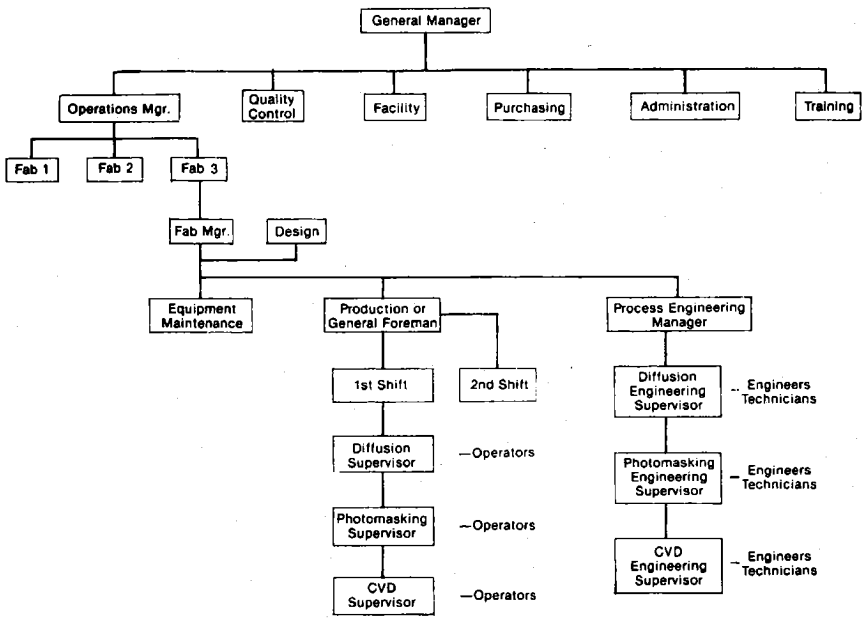


Figure 15.18 Typical semiconductor produce line organization.

Review Questions

1. List the major factors determining the manufacturing cost of a wafer.
2. Rank the factors according to their level of influence.
3. Describe how WIP impacts profit.
4. Discuss how statistical process control can enhance wafer yields.
5. List the four levels of automation in a fabrication process.
6. Describe the philosophy and operation of a JIT inventory system.
7. What is the difference between CIM and CAM?
8. What is a process cell, and how does it compare with a traditional line layout?
9. List the production, process, and yield advantages of a totally automated fabrication area.
10. Draw an organizational chart of a product-line-managed fabrication area.

References

1. J. G. Harper and L. G. Bailey, "Flexible Material Handling Automation in Wafer Fabrication," *Solid State Technology,* July 1984, p. 94.
2. D. Lam, "Minifabs Lower Barriers to 300 mm," *Solid State Technology,* January, 1999, p. 72.
3. W. Maly, "Growth of Silicon Cost and Resulting Changes in Manufacturing Paradigms," *Semiconductor International,* Cahners Publishing, June 1994, p. 158.
4. S. M. Sze, *VLSI Technology,* McGraw-Hill, New York, 1983, p. 5.
5. D. Seligson, "The Economics of 300 mm Wafers," *Semiconductor International,* January 1998, p. 52.
6. Jim Owens, "Semetech," Industry Strategy Symposium, *Semiconductor Equipment and Materials Institute* (*SEMI*), 1994 October 1994.
7. P. Burggraaf, "Applying Cost Modeling to Stepper Lithography," *Semiconductor International,* Cahners Publishing, February 1994, p. 40.
8. S. Shinoda, "Total Automation in Wafer Fabrication," *Semiconductor International,* September 1986, p. 87.
9. P. Singer, "The Thinking Behind Today's Cluster Tools," *Semiconductor International,* August 1993, p. 46.
10. *Ibid.*
11. *Ibid.*
12. *Ibid.*
13. M. Moslehi, "Single-Wafer Processing Tools for Agile Semiconductor Production," *Solid State Technology,* PennWell Publications, January 1994, p. 35.
14. R. Kerby and L. Novak, "ESH: A Green Fab Begins with You," *Solid State Technology,* January 1998, p. 82.
15. D. M. Campbell and Z. Ardehale, "Process Control for Semiconductor Manufacturing," *Semiconductor International,* June 1984, p. 127.
16. W. Levinson, "Statistical Process Control in Microelectronics Manufacturing," *Semiconductor International,* Cahners Publishing, November 1994, p. 95.

17. K. Levy, "Productivity and Process Feedback," *Solid State Technology,* July 1984, p. 177.
18. *Ibid.*
19. E. Hnatek, "ISO 9000 in the Semiconductor Industry," *Semiconductor International,* July 1993, p. 88.
20. P. Dunn, "The Unexpected Benefits of ISO 9000," *Solid State Technology,* March 1994, p. 55.

Chapter

16

Introduction to Devices and Integrated Circuit Formation

Overview

Integrated circuits are composed of individual conductors, fuses, resistors, capacitors, diodes, and transistors. The operation and formation of the basic each is explored. The formation of the major integrated circuits from the components is also presented in this chapter.

Objectives

Upon completion of this chapter, you should be able to:

1. Sketch and identify the structural parts of the individual components of an integrated circuit.
2. Explain the role and different isolation structures used for integrated circuits.
3. Sketch and identify the operation of a bipolar and MOS transistor.
4. List the types and advantages of the different MOS gate structures.
5. Sketch and identify the parts of a bi-MOS circuit.

Semiconductor-Device Formation

The previous chapters have focused on the individual processes used to make semiconductor devices (also referred to as *components* or *circuit components*) and integrated circuits. It is assumed that the reader has already read about (or is familiar with) the processes and has a

good understanding of the basic structure and electrical performance of the individual components as explained in Chapter 14. There are literally thousands of different semiconductor device structures. They have been developed to achieve specific performances, either as discrete components or in integrated circuits. However, there are basic structures required for each of the major device and circuit types. In this chapter, these basic structures are examined. Mastering them is essential to understanding the many variations and innovative structures that abound in the semiconductor world. The circuit components are:

- Resistors
- Capacitors
- Diodes
- Transistors
- Fuses
- Conductors

Resistors

Resistors have the effect of limiting current flow. This is accomplished by the use of dielectric materials or high-resistivity portions of a semiconductor wafer surface. In semiconductor technology, resistors are formed from isolated sections of the wafer surface, doped regions, and deposited thin films.

The value of a resistor (in ohms) is a function of the resistivity of the resistor and its dimensions (Fig. 16.1). The relationship is

$$R = \rho LA$$

where ρ = resistivity
L = length of resistive region
A = cross-sectional area of the resistive region

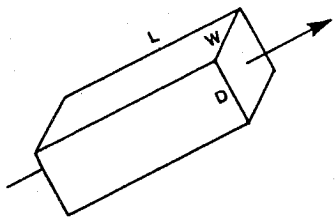


Figure 16.1 Relationship of resistance to resistivity and dimensions.

The area *(A)* becomes $W \times D$, where W = width of the resistor and D = depth of the resistive region. For doped resistors, the length and width are the surface pattern openings and the depth is the junction depth.

It should be obvious that every doped region is also a resistor, and the basic resistor formula governs electrical flow. A conductor is simply a resistor with a low resistance. The conceptual importance of Ohm's law is that the electrical resistance of any region in the device or circuit is altered by any change in dimensions or change in the doping level.

Doped resistors. Most of the resistors in integrated circuits are formed by a sequence of an oxidation, masking, and doping operation (Fig. 16.2). A pattern is opened in the surface oxide. Typical resistor shapes are dumbbells (Fig. 16.3) with the square ends serving as contact regions and the long skinny region in between serving the resistor function. The resistance of this region is calculated from the sheet resistance of the region and the number of squares (□s) contained in the region. The number of squares is calculated by dividing the length by the width.

After doping and a subsequent reoxidation, contact holes are etched in the square ends to contact the resistor into the circuit. A resistor is a two-contact, no-junction device. The term *no-junction* means that

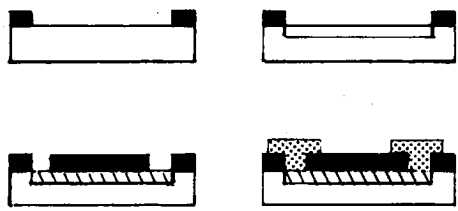


Figure 16.2 Diffused resistor formation.

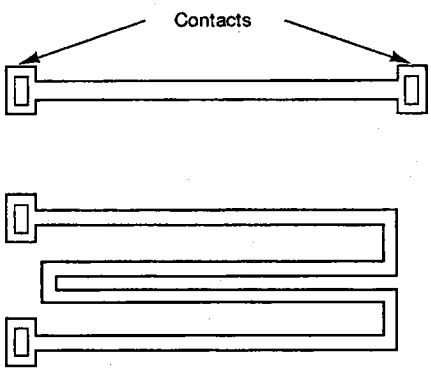


Figure 16.3 Resistor shapes.

the current flows between the contacts without crossing an N-P or P-N junction. But the junction serves to confine the current flow in the resistive region.

Resistors doped by ion implantation have more controlled values than those in diffused regions. Doped resistors can be formed during any of the doping steps performed during the fabrication process. A bipolar base mask will have the base pattern and a set of resistor patterns. In MOS circuits, resistors are formed along with the source/drain doping step. The resistor has the same doping parameters (sheet resistance, depth, and dopant quantity) as the transistor part.

EPI resistors. A resistor can be formed by isolating a section of an epitaxial region (Fig. 16.4). After surface oxidation and contact hole masking, what is left is a three-dimensional region functioning as a resistor.

Pinch resistors. Ohm's law shows that the cross-sectional area of the resistor is a factor in its value (Fig. 16.5). One way to reduce the cross-sectional area (and increase the resistance) is to dope the resistor region and then do another doping of the opposite conductivity type. This occurs in bipolar processing when a resistor region is formed during the P-type base doping with a "pinched" cross section formed from a subsequent N-type region formed along with the emitter doping.

Thin film resistors. Doped resistors don't always have the resistance control needed in some circuits and are poor performers in radiation environments. Radiation, such as found in space, generates unwanted holes and electrons that allow the current to leak across the confining junction. Resistors formed from deposited thin films of metal do not have this radiation problem.

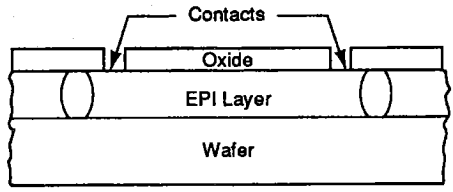


Figure 16.4 Epitaxial layer resistor.

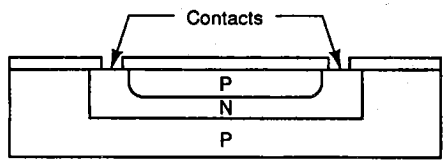


Figure 16.5 N-type resistor "pinched" with P-type doped region.

The resistors (Fig. 16.6) are formed by either film deposit and masking sequences or by lift-off techniques. After resistor formation of the wafer surface, it is "wired" into the circuit by contact between the resistor ends and the leads of the conducting metal. Nichrome, titanium, and tungsten are typical resistor metals.

Capacitors

Oxide-silicon capacitors. Silicon planar technology is based on a silicon wafer with a grown silicon dioxide layer on top or on a silicon dioxide layer grown and an epitaxial layer. If a conducting metal lead lies on top of the oxide, a simple capacitor is formed (Figure 16.7). Recall that a capacitor is formed with a dielectric layer sandwiched between two electrodes. In fact, this structure is a metal-oxide-metal (MOS) capacitor structure. However, the oxide thickness has to be thin enough (about 1500 Å[1]) for the structure to act as a capacitor. The top electrode is also called a *cell plate.* A bottom electrode is also called the *storage node.*

A capacitor is a device that stores charge. A battery is a capacitor. When a voltage is applied to the metal, a charge builds up in the wafer layer under the oxide (Fig. 16.7). The amount of charge is a function of the oxide thickness, the dielectric constant of the oxide, and the area, as defined in the metal top plate. Capacitors of this structure are called parallel-plate, monolithic, or MOS (after the metal oxide semiconductor materials of the sandwich).

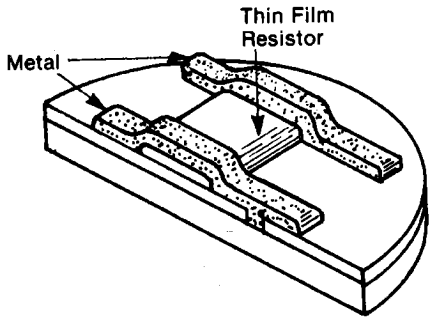


Figure 16.6 Formation of a thin film resistor.

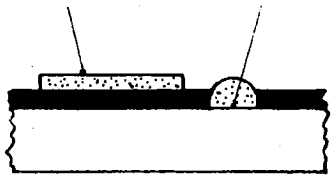


Figure 16.7 Monolithic capacitor.

In dense circuits, an oxide/nitride/oxide (ONO) dielectric sandwich is used. The combination film has a lower dielectric constant, allowing a capacitor area smaller than a conventional silicon dioxide capacitor. Capacitor function acts two ways in ICs. In some circuits, capacitors are formed specifically to store charge. However, capacitor structures are formed whenever metal lines lie over a dielectric on top of a layer of silicon (or other semiconductor material). In this situation, we do not want the capacitor to store charge that may interfere with the circuit operation. In this situation, the dielectric layer needs to be thick enough to prevent capacitor charge storage or the use of high dielectric materials (high-k). See Chapter 12 for reference.

Junction capacitors. A capacitor is formed at every junction in a device. When a voltage is applied across any junction, carriers on each side of it move away from the junction, leaving a depleted region (Fig. 16.8). This depleted region acts as a capacitor in the device or circuit.

The value of this junction capacitance must be taken into account when the circuit is designed. Some circuits actually use junction capacitors as part of the circuit design. In some circuits, the natural junction capacitance has the effect of slowing up the circuit operation. This is due to the time required to "fill up," or charge, the depleted region before current flows. A finite time is also required for the various junction capacitors to discharge. Both of these times affect circuit switching and operational speeds.

Trench capacitors. Preservation of wafer surface area is always a design criterion. One of the problems with oxide-metal capacitors is their relatively large area. Trench (or buried) capacitors solve the problem by creating a capacitor in a trench etched vertically into the wafer surface (Fig. 16.9). The trenches are etched either isotropically with wet techniques or anisotropically with dry etch techniques. The trench sidewalls are oxidized (the dielectric material) and the center of the trench filled with deposited polysilicon. The final structure is "wired" from the surface, with the silicon and polysilicon serving as the two electrodes with the silicon dioxide dielectric between them. Other dielectric materials may be used in place of the silicon dioxide to increase performance.

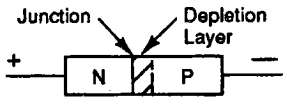


Figure 16.8 Depletion layer junction capacitor.

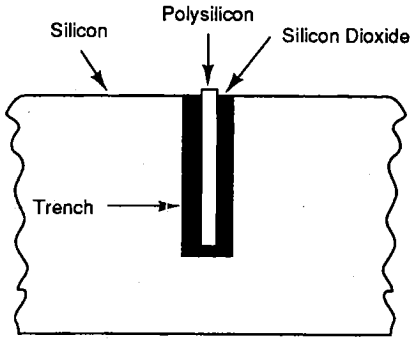


Figure 16.9 Trench capacitor.

Stacked capacitors. Another alternative for conserving surface real estate is to build *stacked capacitors* on the wafer surface. This effort has been driven by the need for small, high-dielectric capacitors for dynamic random access memory (DRAM) circuits. The storage part of a DRAM cell is a capacitor. In DRAM cells, the bottom electrode is usually polysilicon or hemispherical grain polysilicon (HSG). Capacitor shapes can be planar, cylindrical, or with "fins."[2]

A typical structure is shown in Figure 16.9. Capacitor dielectric materials under consideration include tantalum pentoxide (Ta_2O_5) and barium strontium titantate ($BaSrTiO_3$ or BST). The latter is a *ferroelectric* material.[3] Ferroelectric refers to iron containing materials and, in the world of electronics, represent dielectrics with improved speeds over conventional silicon-compatible materials. Another ferroelectric material is $PbZ_{1-x}T_xO_3$ or PZT.

The top electrode materials may be TiN, WN, Pt, polysilicon, or one of the other semiconductor conducting materials.

Diodes

Doped diodes. A diode is a two-region (two-contact) device separated by a junction. A diode either allows current to pass easily or acts as a current block. Which function it performs is determined by the voltage polarity, called *biasing* (Fig. 16.10). When the current voltage is the same as in the diode region, the diode is in forward bias, and the current flows easily. When the polarities are reversed, the diode is reverse-biased, and the current is blocked. A reverse-biased diode can be forced into a conducting state by raising the current voltage until the junction goes into *breakdown*. This condition is temporary; when the voltage is reduced, the diode once again becomes a blocking device (see Chapter 14). Diodes are used in circuits to steer the current around the circuit. By proper choice of the circuit current polarities

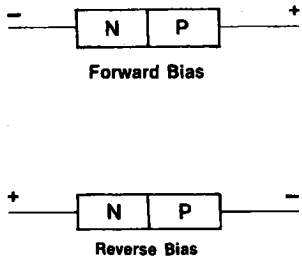


Figure 16.10 Forward and reverse biasing.

and the correct diode polarities, the current is allowed to pass into some branches of the circuit and is blocked out of others. A planar diode is formed from a doped region and two contacts on either side of the junction where it intersects the surface (Fig. 16.11). Diodes are usually formed along with transistor doping steps. Thus, in bipolar circuits, there are base-collector diodes and emitter-base diodes. In MOS circuits, most of the diodes are formed with the source-drain doping step.

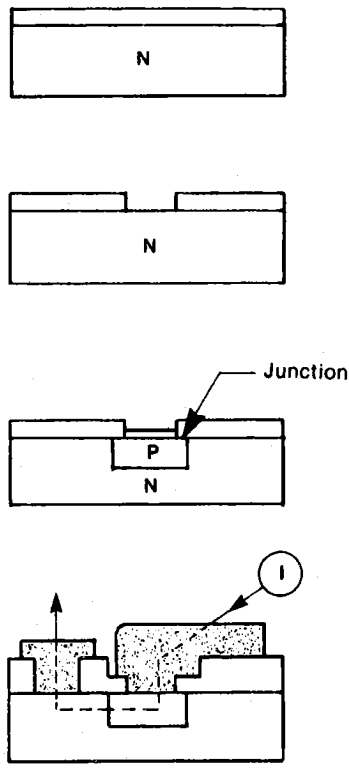


Figure 16.11 Formation of a P/N planar diode.

Schottky barrier diodes. In 1938 (ten years before the invention of the transistor), W. Schottky[4] discovered that whenever a metal is in contact with a lightly doped semiconductor, a diode is formed (Fig. 16.12). This diode has a faster forward time (it responds faster) and operates with a lower voltage than a doped silicon junction diode. Metal contacts to highly doped regions (greater than 5×10^{17} atoms per cubic centimeter) are regular ohmic contacts. This is the situation for the majority of contacts in a silicon circuit. This Schottky diode effect is taken advantage of in some NPN bipolar transistors. The structure and effect are explained in the section on bipolar transistors.

Transistors

Transistor operational analogy. A transistor is a three-contact, three part, two-junction device that performs as a switch or an amplifier. An often used analogy to explain the role of the parts and the operation of a transistor is the water flow system in Fig. 16.13. The flowing water represents current flow. In this system, one part is the source of the water (the tank), the valve controls the flow, and the bucket collects the water. The system can be operated as a switch simply by turning the valve on and off. It can even be imagined in an amplifier role. Consider the valve as a high-mechanical advantage miniature water wheel activated by a small external stream of water. A small trickle onto the valve wheel could open the valve to allow a large flow through the system. If the whole system was enclosed so that an observer saw only the trickle going in and a large flow coming out, such an observer might conclude that the system was *amplifying* the water trickle.

Bipolar transistors. The same basic parts and functions are present in solid-state transistors. A bipolar transistor is shown as both a simple block arrangement and in a double-doped planar form in Fig. 16.14. The current flows from the emitter region (tank) through the base (valve) into the collector (bucket). When there is no current to the base, the transistor is turned off. When it is on, the current flows. It only takes a small current to turn the base on enough to allow current

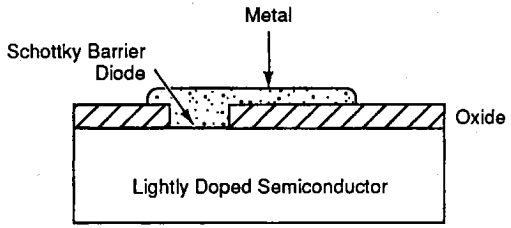
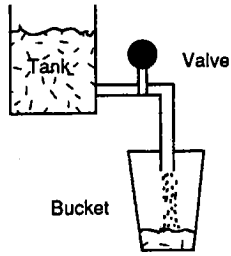


Figure 16.12 Schottky barrier diode.



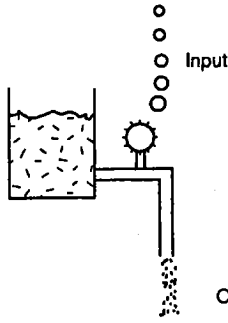


Figure 16.13 Water analogy of transistor operation.

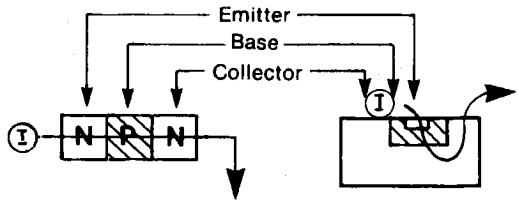


Figure 16.14 Bipolar transistor operation.

flow through the whole transistor. The size of the base current regulates the larger amount of current flowing through the transistor (called the *collector current*). There is an amplification of the base current to the collector current. The base current in effect changes the resistance of the base region. In fact, the term *transistor* comes from an early term for a bipolar transistor: *transfer resistor.* During operation, both positive and negative currents flow in the base, hence the term *bipolar*—literally, two polarities.

The amplification property, called *gain* or *beta*, is numerically the result of dividing the collector current by the base current (see Chapter 14). For efficiency, the emitter region has a higher doping concentration than the base, and the base doping is higher than the collector. A typical doping concentration versus distance plot is shown in Fig. 16.15.

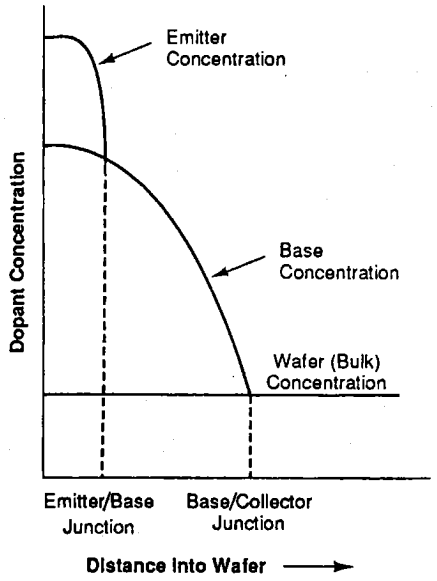


Figure 16.15 Dopant concentration profile of bipolar transistor.

Most bipolar circuits are designed with NPN transistors. NPN represents the respective conductivity types of the emitter, base, and collector. Some applications require PNP transistors, with many of them being formed laterally (Fig. 16.16). NPN transistors are more efficient because of the ease (higher mobility) of electrons in the N-type regions.

Bipolar transistors feature fast switching speeds. The speed is governed by a number of factors, of which the most important is the base width. Applying common sense, it stands to reason that the shorter the distance an electron or hole has to travel, the less time it will take.

Bipolar transistors can switch on and off in as little as a billionth of a second. To achieve this speed, the transistor is maintained at an "on"

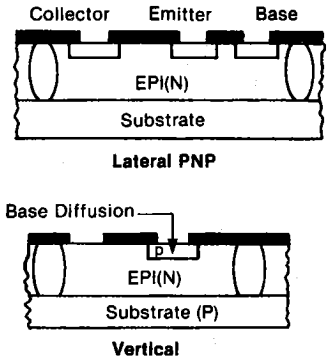


Figure 16.16 Lateral and vertical PNP transistors.

state. This means that the base always has power applied, which is a downside of bipolar-based circuits. Another penalty for this necessary condition is a buildup of heat in the transistor. This heat will eventually affect circuit operation and is the reason for the cooling fans and air-conditioned rooms of earlier bipolar-based computers.

Schottky barrier bipolar transistors. The Schottky barrier diode principle mentioned in a previous section is put to use in some bipolar transistors (Fig. 16.17). The construction requires that the base contact be extended into the collector region. When covered with metal, a Schottky diode is formed between the base and collector, which results in a faster-responding transistor. In a circuit, the time required to turn the transistor on and off (switching speed) is critical when millions of transistors are operating.

Field-effect transistors (FETs)

Metal-gate MOSFET. As early as 1948, William Shockley noted another type of transistor operation—a field-effect-actuated transistor current flow. That effect was developed into the field-effect transistor, with the MOS structure now the most popular design. A MOS transistor, like a bipolar (Fig. 16.18), has three regions, three contacts, and two junctions, but in a different structure. There is a similar analogy to the water system described previously. Current travels from the source region (tank), through the dielectric gate material (valve), and into the drain (bucket) before exiting the device.

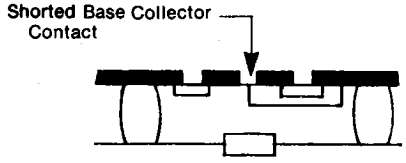


Figure 16.17 Schottky barrier bipolar transistor.

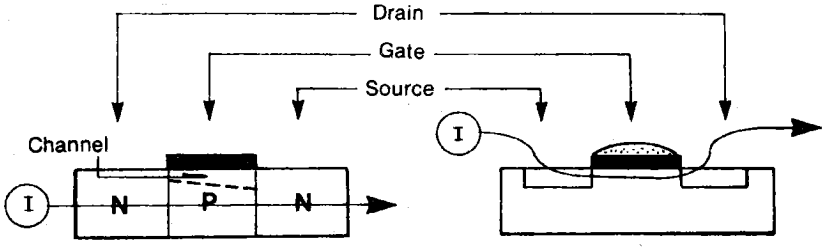


Figure 16.18 MOS transistor operation.

A MOSFET gate controls current flow by a different mechanism from that of a bipolar base. The MOS structure shown in Fig. 16.19 is a simple metal-gate type capacitor and operates the same as a capacitor. When a voltage (the gate voltage) is applied to the gate through the gate metal, a *field effect* takes place in the surface of the semiconductor. The effect is either a buildup of charge or a depletion of charges in the wafer surface under the top plate. Which event occurs depends on the doping conductivity type in the wafer under the gate and the polarity of the gate voltage.

The buildup or depletion of charge creates a channel under the gate that connects the source and drain. The surface of the semiconductor is said to be *inverted.* The source is biased with a voltage, and the drain is grounded relative to the source. In this condition, a current starts to flow as the inverted surface creates an electrically connecting channel. The source and drain are essentially shorted together. Applying more voltage to the gate increases the size of the channel, allowing more current to flow through the transistor (see Chapter 14). By controlling the gate voltage, a MOS transistor can be used as a switch (on/off) or as an amplifier. However, MOS transistors are voltage amplifiers, unlike the current amplification of bipolar transistors.

If the source and drain are N-type formed in a P-type wafer, the channel must be of N-type for conduction to occur. This type of MOS transistor is called N-channel. MOS transistors with P-type sources and drains are P-channel. Most high-performance MOS circuits are built around N-channel transistors due to the higher mobility of electrons in silicon. The mobility makes N-channel transistors faster, and they consume less power than P-channel circuits. They often are referred to as NMOS transistors. Figure 16.20 shows the major steps in the formation of an N-channel metal-gate MOS transistor.

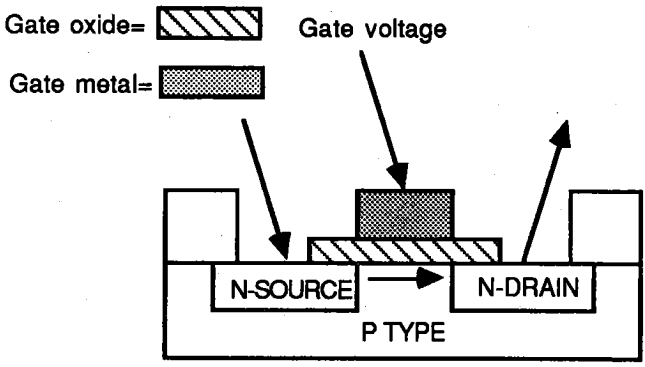


Figure 16.19 Metal gate MOS transistor.

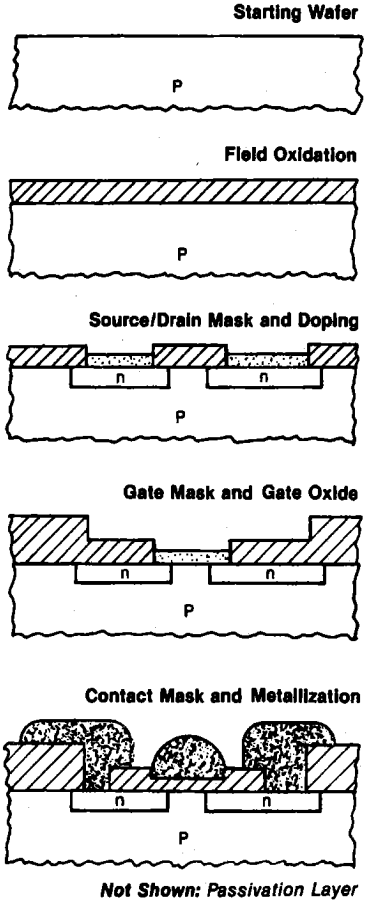


Figure 16.20 N-channel metal gate MOS process.

Silicon gate MOS. A certain amount of voltage must be applied to the gate metal before the channel forms. This voltage is called the *threshold voltage* or *V*. The value of the threshold voltage is an important and critical circuit parameter. A lower *V* means fewer power supplies and faster circuits.

A primary parameter that determines the threshold voltage is the *work function* between the gate material and the doping level in the semiconductor. The work function can be thought of as a kind of electrical compatibility. The lower the work function, the lower the threshold voltage, the lower the power required to run the circuit, and so on.

Deposited doped polysilicon has a lower work function than aluminum as a MOS gate material and has become the standard gate electrode material for MOS transistors. The formation of the transistor is

shown in Fig. 16.21. The polysilicon is heavily doped N-type to reduce its resistance. Thus doped, it serves as the gate electrode and as a circuit conduction line. A polysilicon gate can withstand subsequent high-temperature processing without degradation.

An additional benefit of the silicon-gate process is the self-aligned gate. In the metal-gate process sequence, a hole for the gate oxide must be patterned between the source and drain. To ensure that the source and drain are bridged by the gate, overlap for alignment tolerances must be allowed. This results in some overlap of the gate into the source and/or drain. The overlap becomes a region of unwanted capacitance. In the silicon-gate process, the gate is formed first and acts as a mask to locate the source and drain. Thus, whatever the gate placement, the source and drain *self-align* to it.

Other factors, besides the gate metal material, affecting the gate threshold voltage and operation are:

- Gate oxide thickness
- Gate material (dielectric constant)

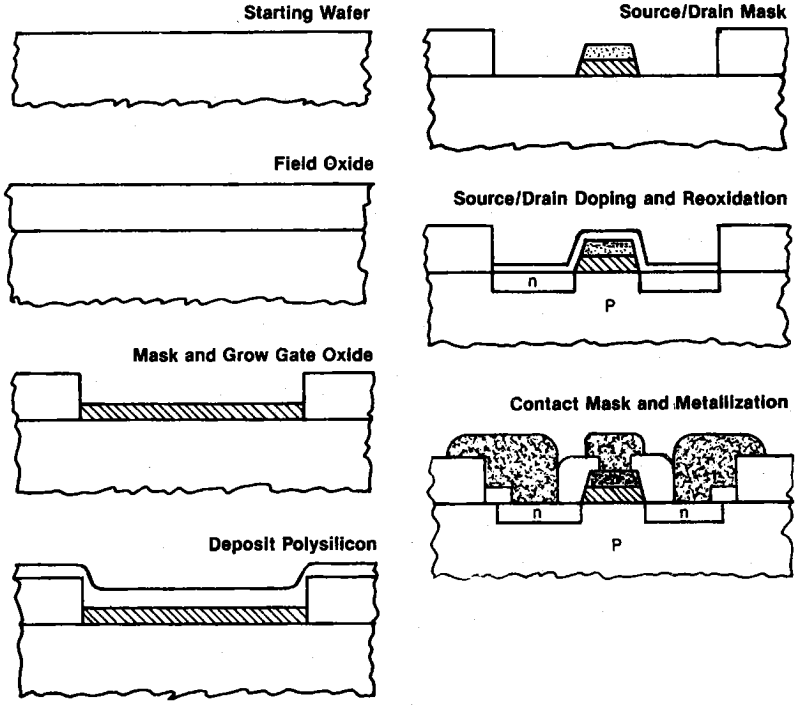


Figure 16.21 Silicon gate MOS process steps.

- Source-drain separation (channel length)
- Gate doping level
- Sidewall capacitance of the doped source and drain regions

In the section on capacitors, the development of high-dielectric-constant materials for higher performance in a smaller area were discussed. These same factors apply to MOS gate materials. The drive to 0.18 μm design rules will require gate thicknesses in the 90- to 60-Å range.[5]

The thinner the gate oxide, the faster the device and the lower the threshold voltage. Current gate oxide thicknesses for production devices are in the 100- to 300-Å range. The technology roadmap projects that, as the industry enters the 0.35- and 0.18-μm design rule era, gate thicknesses will fall into the 90- to 60-Å range.[5] In this range, maintaining thickness control, pinhole-free gates, and small-area gates has driven the development of new gate materials. They are the same materials discussed in the section on capacitors. The new materials are oxide/nitride sandwiches (ONO), oxide tantalum pentoxide (OTa2O3), and amorphous silicon film stacks.

Channel lengths also affect the speed. Channel lengths have continually shrunk to the submicrometer range. In self-aligned structures, the channel length is established by the gate width. The gate doping level influences the threshold voltage by modifying the work function difference between the gate metal and surface. Ion implantation, with its ability to dope through thin oxides, is often used to set the gate doping level. Sidewall capacitance of the doped source and drain also serves to slow up the operation of the device, as enough charge must be built up to overcome the junction capacitance.

Polycide-gate MOS. MOS development was impeded by the inability of the industry to grow noncontaminated and thin oxides in the 1960s. Contamination, especially the mobile ionic variety, interferes with the field effect, making very unreliable gates. In fact, clean gate oxides and the oxide-silicon interface are so well understood and effective that gate oxide replacement is proceeding very cautiously.

Consequently, the quest for a higher-efficiency gate has led to the polycide structure (Fig. 16.22). The gate sandwich retains the thin oxide on the wafer surface topped by a layer of polysilicon. The polysilicon provides a low work function (and lower threshold) gate, and the reliable polysilicon-oxide interface is preserved. The new layer is a refractory metal silicide on top of the polysilicon. The silicide makes a low contact resistance with the polysilicon (as compared with aluminum) and reduces the overall sheet resistance of the polyside sandwich.

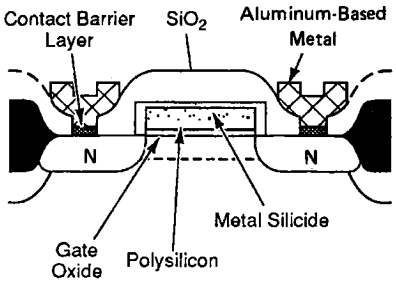


Figure 16.22 Polycide gate structure.

Salicide-gate MOS. The self-aligned process that uses the polycide-gate structure is called a *salicide* gate. Its formation is illustrated in Fig. 16.23. The process combines the best features of a polysilicon gate with self-alignment. The source and drain are lightly diffused around the polysilicon gate. Then a layer of silicon dioxide is deposited and anisotropically etched to form spacers on the side of the gate. These spacers act as ion implantation masks for a subsequent heavier doping of the source and drain. The more lightly doped "finger" under the gate is called a lightly doped drain extension (LDD) and is necessary for channel lengths of less than 2 μm.[6] After ion implantation, the refractory metal is deposited, and the silicide is formed by reaction with the underlying polysilicon layer by an alloy step. The final step is the removal of the unreacted refractory metal from the wafer surface.

The LLD process illustrated puts a lightly doped fingers in both the source and drain areas. There are processes designed to create asymmetrical LLD structures that place the finger only in the drain region.[7]

Diffused MOS (DMOS). DMOS refers to a diffused MOS structure (Fig. 16.24). The channel length is established by two diffusions through the same opening. As the second diffusion is taking place, the

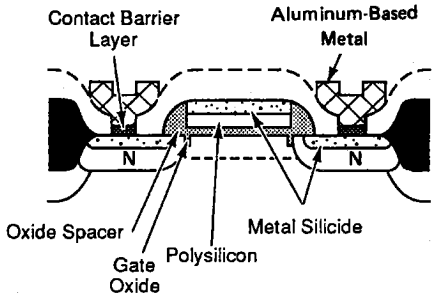


Figure 16.23 Salicide-gate structure.

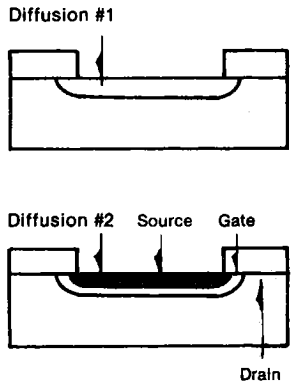


Figure 16.24 DMOS structure.

first moves laterally to the sides. The second diffusion functions as the source, and the bulk semiconducting material of the wafer functions as the drain. The difference between the two diffusion widths is the channel length of the transistor. DMOS structures feature narrow and well controlled gate widths.

Memory MOS (MMOS). MMOS is a structure that provides a more or less permanent storage of the charge in the gate region. The storage is provided by a thin layer of silicon nitride between the wafer and the gate oxide (Fig. 16.25). When the gate is charged to store data, the silicon nitride layer traps and retains it. This type of transistor is used in nonvolatile circuits where protection against memory loss is important (see Chapter 17).

Junction field-effect transistors (JFETs). A junction field-effect transistor (Fig. 16.26) is similar in construction to a MOSFET but has a junction formed under the gate. During operation, the current flows *under*

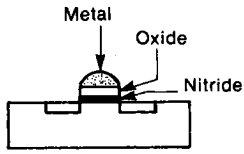


Figure 16.25 MMOS structure.

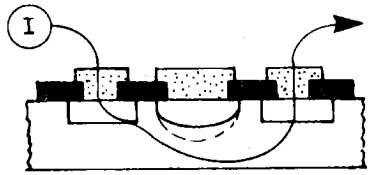


Figure 16.26 Junction field-effect transistor.

the diffused region from the source to the drain. As the gate voltage is increased, a region depleted of charge (the depletion region) spreads under the junction toward the N-type and P-type interface. The depleted region will not support current flow and has the effect of restricting the current flow as it increases in depth.

A JFET operates opposite from a MOS transistor. In the MOS version, increasing the gate voltage *increases* the current flow. In a JFET, increasing the gate voltage *decreases* the current flow. JFETs are a standard gallium arsenide device. The transistor is formed in an N-type GaAs layer that is on top of a semi-insulating GaAs wafer (Fig. 16.27).

Alternative (Scaled) Transistor Designs

A standard approach to denser circuits is *scaling*, also called a *die shrink*. Scaling starts with a proven design and reduces the dimensions. However, scaling can introduce new problems, such as increased leakage. Many of the new materials are designed to address leakage and other scaling issues. Alternative transistor structures are also being explored. Ideas include double gates (DGs) and ultra-thin body (UTB) for MOSFET devices. Schematics are illustrated in Fig. 16.28. Other ideas include vertical transistor structures and a fin structure, called a FinFet. In this structure, the gate is build vertically, and the source and drain build up on either side of the fin.

Metal semiconductor field-effect transistor (MESFET). The MESFET is the basic GaAs transistor structure (Fig. 16.29). The MESFET operates in the same manner as the JFET, but the gate metal is deposited directly onto the N-type GaAs layer.

Fuses and conductors

The formation of fuses and surface wiring conductors were detailed in Chapter 14. In very dense circuits, precious surface area is preserved

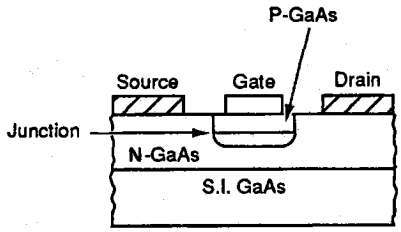


Figure 16.27 GaAs JFET.

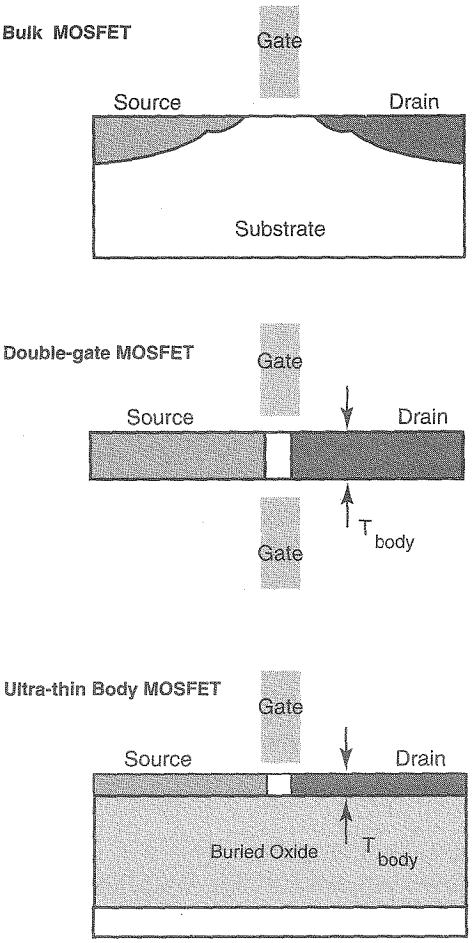


Figure 16.28 Alternate MOSFET devices. *(Courtesy of Semiconductor International.)*

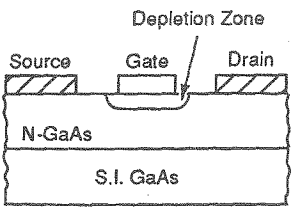


Figure 16.29 GaAs MESFET.

by the use of subsurface *underpass conductors*. These are created from heavily doped regions formed under surface wiring leads.

The issues and challenges are depositing and patterning multiple metal layers. Key technologies are planarization, step coverage, and plug filling. Development of low-contact-resistance metal film and film sandwiches will be required in the 0.5-micron design rule era.

Integrated-Circuit Formation

Integrated circuits contain all the components described in the previous sections. The components are formed in specific sequences with the process flows designed around the transistor in the circuit. The process designer will attempt to form as many component parts as possible with each doping step.

Circuits are designated by the transistor type. A bipolar circuit means that the circuitry is based on bipolar transistors. MOS circuits are based on one of the MOS transistor structures. For the first 30 years of the semiconductor industry, the bipolar transistor and bipolar circuits were the structures of choice. Bipolar transistors had fast speeds (switching times), control of leakage currents, and a long history of process development. These qualities fit nicely into the logic, amplifying, and switching circuits that were the first offerings of the industry. These circuits handled the computational requirements of the growing computer industry. The internal memory functions of the early computers were handled by core memories. These memories were limited in capacity and slow. Much of the information needed was stored outside the computer on tape, disks, or punch cards. While bipolar memory circuits were available, they could not compete economically with core memories.

MOS transistor circuits held the promise of fast, economical, solid state memory, but early metal-gate MOS circuits suffered from high leakage currents and poor parameter control. Even so, the built-in advantages of MOS transistors drove the development of MOS memory circuits. The advantages are smaller dimensions, which allow denser circuits, and relatively faster switching speeds. Yields tend to be higher on smaller-dimensioned circuits, since a given defect density will affect fewer transistors and components.

Perhaps the biggest density factor advantage of MOS components is the smaller area required for isolation of adjacent components. The various isolation schemes used are discussed in the next sections. Another advantage is low-power operation. First, MOS transistors sit in the circuit in the "off" mode, not soaking up power or generating heat like bipolar transistors, which must be all the time to be in a "ready" state. Second, MOS transistors, being voltage-controlled devices, require a lower power to operate. CMOS circuits are an integrated circuit design that reduces power requirements even further.

An initial advantage of MOS circuits was fewer processing steps and smaller die sizes, which made for lower processing costs and higher yields. These advantages have disappeared as MOS circuits have evolved to VLSI/ULSI size with the additional steps required to fabricate CMOS circuits. In general, the faster switching speed of bi-

polar circuits has made them favored for logic circuits. MOS circuits, with their smaller component dimensions and lower power requirements, have been incorporated into memory circuits. By the 1980s, these traditional uses blurred, with CMOS technology the preferred system for most circuit designs. These topics are addressed further in Chapter 17.

Bipolar circuit formation

Junction isolation. The bipolar transistor structure and basic performance have been illustrated. However, combining transistors with the other devices to form a circuit requires additional structures. They include an isolation scheme and a low-resistance collector contact. If two transistors or other devices are fabricated next to each other, they will not work, because they are electrically shorted together (Fig. 16.30). An early challenge of bipolar IC designers was a way to *isolate* the various circuit components. This need led to the epitaxial (epi) layer bipolar structure (Fig. 16.31).

The process starts with a P-type wafer into which an N-type diffusion is made (the flow diagram does not show the oxidation and masking steps required to create the diffused layer). After the diffusion step, an N-type epitaxial layer is deposited, leaving the N-type diffused region "buried" under the epitaxial layer. The N-type region is known as a *buried layer* or as the subcollector of the transistor. Its function is to provide a lower-resistance path for the collector current as it flows out of the base region on its way to the surface collector contact.

After the deposition of the epitaxial layer, it is oxidized, and a hole is opened up on each side of the buried layer. A P-type doping step is performed deep enough to reach the P-type wafer surface. The doping step divides (*isolates*) the epitaxial layer into N-type islands, each surrounded on the sides (the doped regions) and the bottom (the P-type wafer) by P-type doped regions. Components formed on the surface of each of the islands are electrically isolated from each other (Fig. 16.32). The electrical isolation occurs because the N-P junction is "wired" into the circuit to function in the reverse bias mode; that is, no

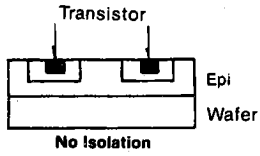


Figure 16.30 Adjacent bipolar transistors with common collectors.

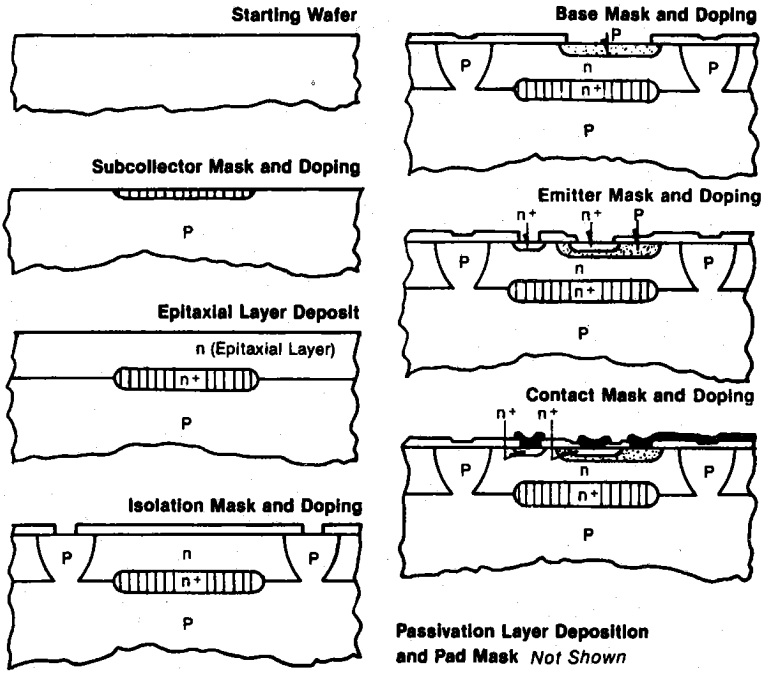
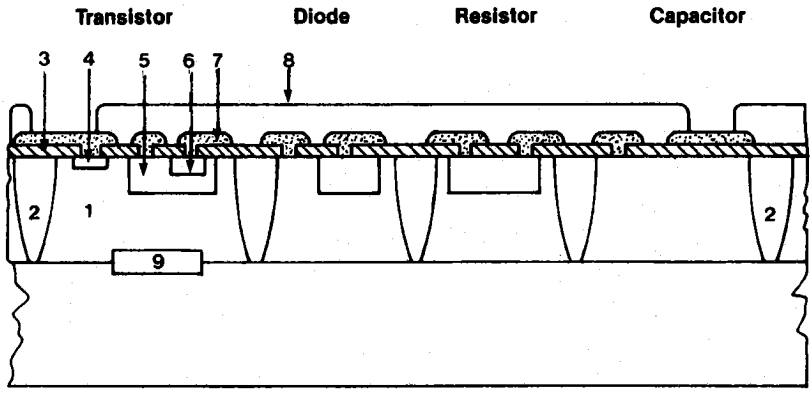


Figure 16.31 NPN bipolar process.



1. EPI and Collector
2. Isolation
3. Surface Oxide
4. Collector Contact
5. Base
6. Emitter
7. Metalization
8. Passivation Layer
9. Buried Layer

Figure 16.32 Bipolar structures.

current crosses the junction. This scheme is called *junction isolation* or *doped junction isolation.*

Dielectric isolation. In high-radiation environments, such as outer space or in the vicinity of atomic weapons, doped junctions produce holes and electrons that flow in the device and compromise the junction functions. Besides causing circuit component failure, the radiation swamps out the isolation protection of the doped regions. Dielectric isolation schemes provide the necessary electrical isolation and radiation protection.

The process (Fig. 16.33) starts with the etching of pockets or trenches in a wafer surface. The etching can be either an isotropic wet etch or an anisotropic dry etch. Isotropic etch profiles follow the orientation structure of the wafers. Dry etch processes allow the shaping of

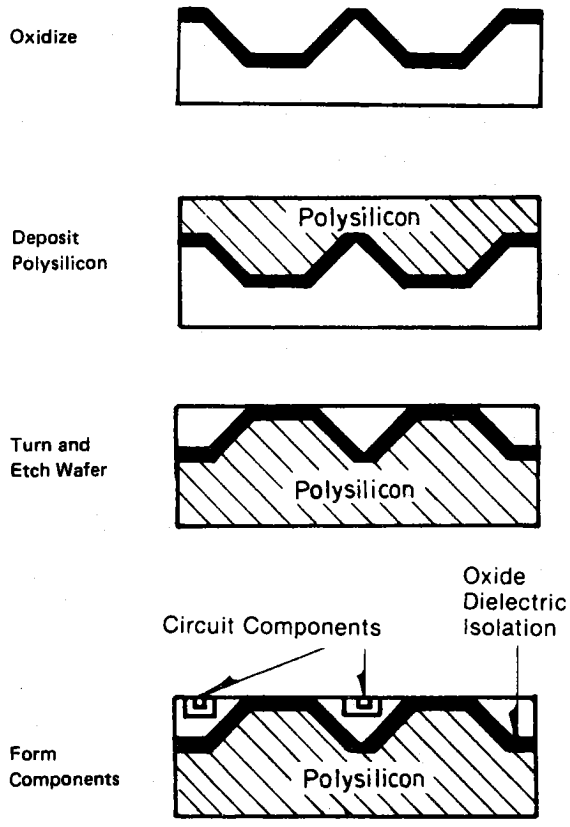


Figure 16.33 Dielectric isolation.

the trenches. One goal of the etch step is to minimize the area of the pocket on the wafer surface. Wide pockets limit the packing density of the circuit.

After etch, the pocket sides are oxidized and backfilled with deposited polysilicon. Next, the wafer is turned over, and the silicon of the wafer is lapped until the oxide layer is reached. These steps leave a wafer surface containing oxide-isolated pockets of the original single silicon material. The circuit components are fabricated in the silicon pockets, with each pocket being isolated on three sides by the layer of silicon dioxide. The dielectric property of the silicon dioxide prevents leakage currents in both normal and radiation environments.

Localized oxidation of silicon (LOCOS). Junction isolation takes up valuable surface real estate, and dielectric isolation is area consuming and requires extra processing steps. A popular alternative is LOCOS (Fig. 16.34). The process starts with a layer of silicon nitride deposited and etched on the wafer surface. Active devices will be formed in the area defined by the silicon nitride layer. In the partially recessed version, an oxidation follows. Oxygen will not penetrate the silicon nitride to cause the oxide to grow on the exposed silicon surface. The silicon for the silicon dioxide comes from the wafer surface and, because silicon dioxide is less dense than silicon, the oxide layer forms slightly above the original silicon surface. It is *partially recessed* rela-

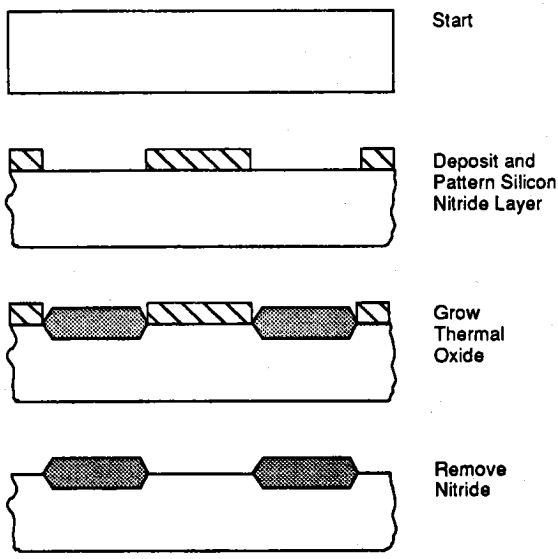


Figure 16.34 LOCOS process.

tive to the wafer surface. After the oxidation, the silicon nitride is removed, leaving the area free for the formation of circuit components. A variation of the LOCOS isolation process is illustrated in Fig. 16.35*b*. In this process, the silicon surface is etched *before* the oxidation. By calculating the proper removal amount, the subsequent oxidized layer is *fully recessed* to the original surface. Bipolar schemes using LOCOS isolation are shown in Fig. 16.35.

Collector contact. Note that, in the cross section (Fig. 16.32) of a circuit integrated bipolar transistor, there is a doped region under the collector contact. This doped region is put into the surface along with the emitter N-type doping. The emitter is usually designated N^+ to indicate that it is highly doped. The N^+ region under the collector contact is present to create a lower resistance between aluminum metallization and the silicon of the collector.

MOS integrated circuit formation

MOS LOCOS isolation. While MOS transistors are somewhat self-isolated (because they don't share common electrical parts), there is some electrical leakage between devices, especially as the spacing becomes very close. Isolation is necessary to block leakage currents. The structures are generically called *channel stops.*

LOCOS is the preferred isolation technique. However, there are several problems to overcome to make a LOCOS structure effective in advanced circuits.[8] One problem is the *bird's beak* spur that grows under the edge of the blocking silicon nitride layer (Fig. 16.36). The beak takes up real estate, effectively enlarging the circuit. At a performance level, it induces stress damage in the silicon during the oxidation step. The stress comes from the mismatch in thermal expansion properties between silicon nitride and silicon. A solution for the stress problem is

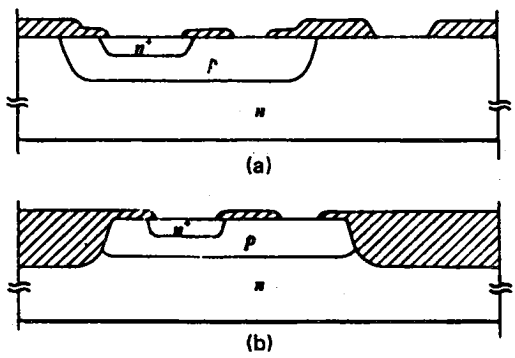


Figure 16.35 (*a*) Conventional bipolar transistor and (*b*) LOCOS isolated bipolar transistor. *(Source:* VLSI Fabrication Principles, *Ghandhi.)*

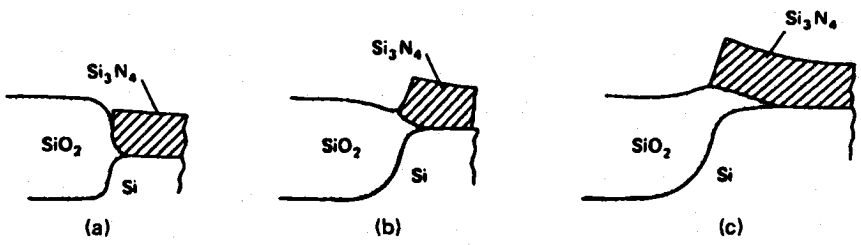


Figure 16.36 Bird's beak growth. (*a*) No pre-etch, (*b*) 1000-Å pre-etch, and (*c*) 2000-Å pre-etch. *(Source:* VLSI Fabrication Principles, *Ghandhi.)*

to grow a thin oxide layer under the silicon nitride film. This is called a *pad oxide.*

Minimizing the bird's beak and reducing the stress in the active device region has spurred a number of variations on the LOCOS process. One, called SWAMI, was developed by Hewlett Packard (Fig. 16.37).[9] The process starts the same as a standard LOCOS. After the nitride

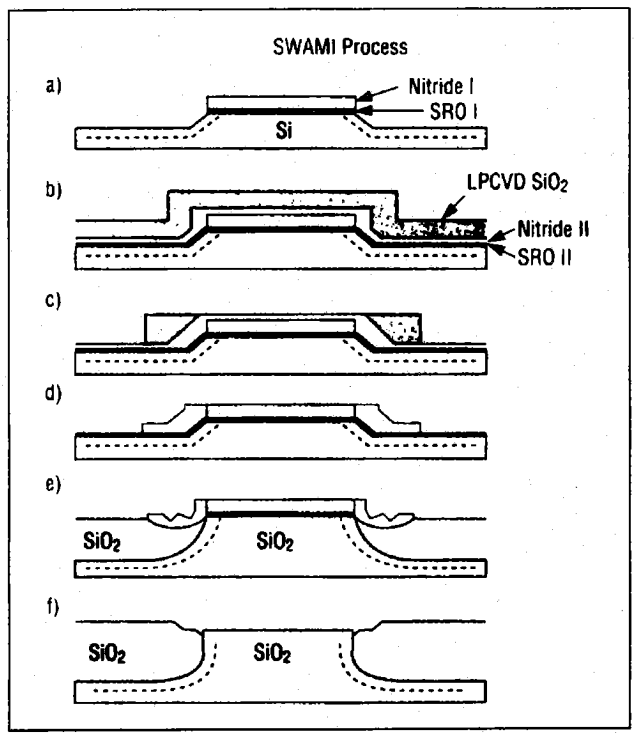


Figure 16.37 SWAMI process. *(From* SST, *June 1993, p. 97.)*

and pad oxide, grooves (or trenches) are etched with a orientation sensitive etchant. On ⟨100⟩-oriented material, the groove side walls are at a 60° angle, which reduces stress in the silicon. Next, another stress relieve oxide (SRO) layer is grown and covered by a layer of silicon nitride, which provides conformal coverage. An LPCVD silicon dioxide completes the sandwich before etching (*c*). This oxide protects the silicon nitride layer from removal. Finally, the field oxide (FOX) is grown. The length of the nitride layer governs the size of the bird's beak encroachment. Removal of the original nitride/SRO and the second nitride leaves a somewhat planarized surface for device formation. LOCOS isolation schemes usually include an ion implanted layer between active regions to provide further channel stop capabilities.

Trench isolation. Trench isolation is also used for MOS circuits (Fig. 16.38). The procedure is the same as forming trench capacitors. One version, called *shallow trench* isolation, addresses the bird's beak problem that comes with standard LOCOS isolation. In this structure, a shallow trench is etched between the components and filled with a deposited dielectric (Fig. 16.39). After sidewall oxidation and dielectric fill of oxide, a CMP step replanarizes the surface.

CMOS. Complementary MOS (CMOS) is a MOS circuit formed with both N-channel and P-channel transistors. CMOS has become the

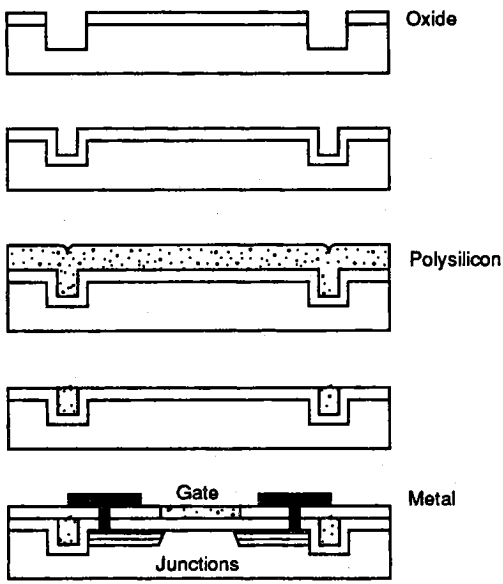


Figure 16.38 MOS trench isolation.

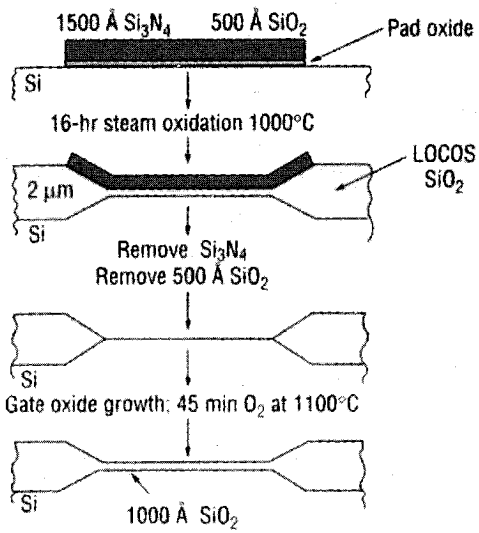


Figure 16.39 Shallow trench process. *(Reprinted from the September 1997 edition of* Solid State Technology, *copyright 1997 by PennWell Publishing Company.)*

standard circuit for many applications. It is the CMOS circuit that has made possible digital watches and hand-held calculators. It allows circuits on one chip that would require several chips using N-channel and P-channel only circuits. CMOS circuits also use lower amounts of power than comparable circuits.

CMOS structures (Fig. 16.40) are formed by first fabricating an N-channel MOS transistor in a deep P-type well formed in the wafer surface. After N-channel transistor formation, a P-channel transistor is fabricated. The transistor structures are silicon gate or other advanced structures. CMOS processing uses the most advanced techniques, because smaller, more densely packed, and higher-qual-

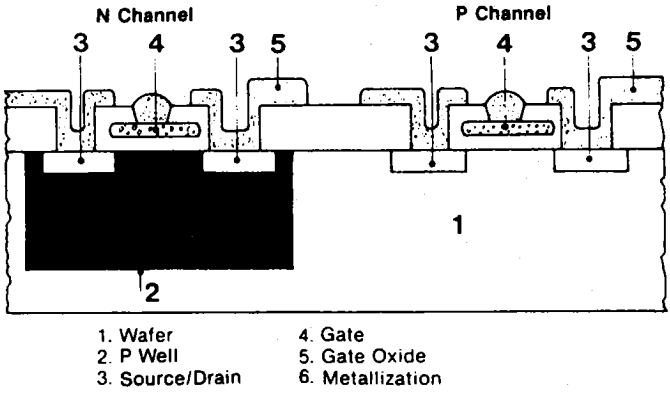


Figure 16.40 CMOS structure.

ity components all increase the advantages inherent in the CMOS design.

An isolation problem particular to CMOS structures is latch-up. Figure 16.41 shows a cross section of a portion of a CMOS chip. The side-by-side MOS transistors form a lateral bipolar (NPN) transistor. During circuit operation, the lateral bipolar transistor can function as an unwanted amplifier, increasing its output to a point where it causes the memory cell to go into a state in which it cannot be switched. This is *latch-up*. In this state, the cell cannot be addressed for its information. One solution to preventing latch-up is a low-resistivity epitaxial layer that shunts (shorts) the emitter of the bipolar transistor so that it does not "turn on," preventing the latch-up.

Another solution is LOCOS and other structures. One is a hardened-well design using a buried layer that effectively breaks up the lateral bipolar transistor (Fig. 16.42).

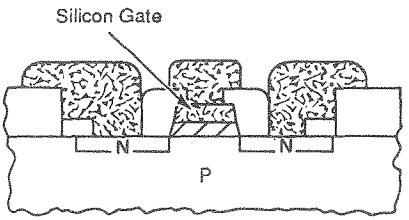


Figure 16.41 Cross section of silicon gate MOS transistor.

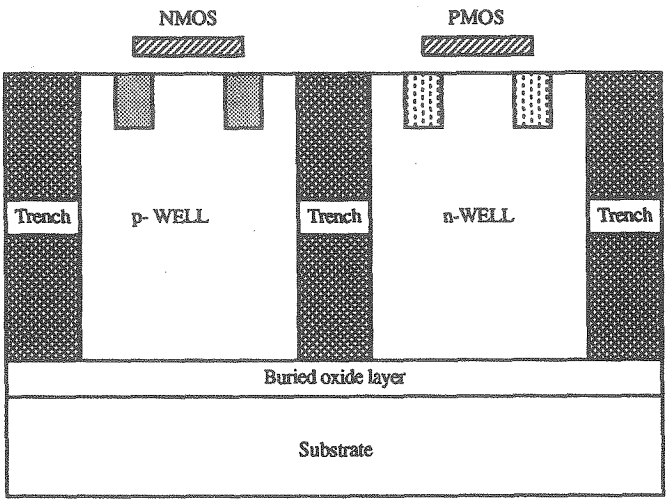


Figure 16.42 CMOS well structure silicon on insulator (SOI) with buried layer and trench isolation. *(Source:* Semiconductor International, *July 1993, p. 34.)*

Formation of the deep P-well by diffusion techniques requires a large thermal budget that contributes to lateral diffusion problems and stress growth. Ion-implanted wells have become common. Additionally, with ion implantation, the wells can be dopant graded in the vertical direction to maximize performance. Called *retrograde wells,* they are formed by high-energy implants (MeV) such that the bottom of the well has a higher dopant concentration. Implanted wells offer the advantage of fewer steps compared to conventional well formation[10] (Fig. 16.43). Also, because the N and P wells are formed in the same surface, planarity is preserved. In conventional well processing, the two well surfaces are offset and can cause depth-of-focus problems.

Bi-MOS

The unique advantages of bipolar and CMOS transistors and their respective circuits come together in bi-MOS (or bi-CMOS) circuits. The circuits (Fig. 16.44) have in them bipolar, P-channel, and N-channel transistors along with memory cells (see Chapter 17). The low-power advantage of CMOS is used in logic and memory sections of the circuit, while the high-speed performance of bipolar circuitry is used for signal processing.[11] These circuits represent great challenges to the processing area with their large die size, small component size, and large number of processing steps.

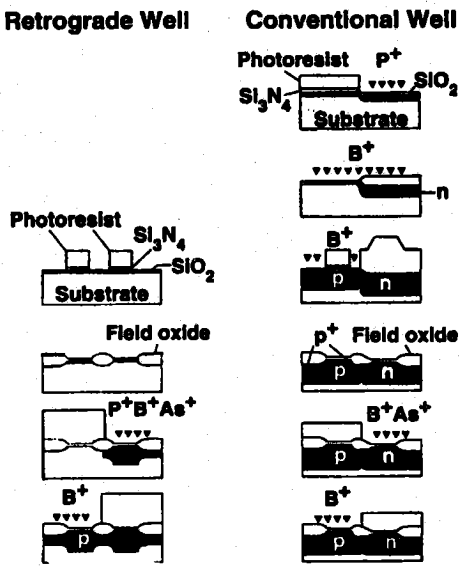


Figure 16.43 Implanted and conventional wells. *(Source:* Semiconductor International, *June 1993, p. 84.)*

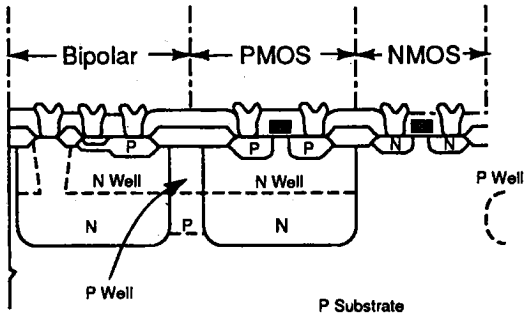


Figure 16.44 Bi-MOS structure.

Silicon on insulator (SOI) isolation

There has been renewed interest in forming CMOS and bi-CMOS circuits in thin epitaxial silicon films deposited over an insulating substrate (see Chapter 12). Elimination of a conductive substrate minimizes leakage and latch-up problems.[12]

Superconductors

Much interest has been generated by developments in superconductivity materials. Superconductivity is a phenomenon that occurs in certain materials when they are cooled close to absolute zero (–273°C). In ordinary metals, current is passed along by the flow of electrons. In an at-rest (nonconducting) state, the electrons exist in orbits around the nuclei of the atoms. To become conducting, they must gain energy to overcome the internal resistance of the particular material. The energy must be continuously supplied to maintain the current flow.

In a superconducting material, electrons exist in a "conducting state" and can support an electrical current with little or no additional energy input. The prospect of a resistanceless material has the potential of revolutionizing electronic devices.

Semiconductor researchers have investigated the superconducting effect for years. In 1962, B. D. Josephson described an effect that is now named after him (Josephson effect). When a thin oxide separates two superconducting materials, electrons will pass through the oxide with zero resistance. The structure is called a *Josephson junction.* This effect (called *tunneling*) is very complicated and requires quantum physics concepts to understand (well beyond the scope of this text!). The result is that the oxide has the functional aspects of a gate, and Josephson junctions can perform basic switching, logic, and memory functions.[13]

Microelectromechanical systems (MEMS)

Semiconductor junctions have the property that almost anything done to them produces a change in their functioning. The anything can be a physical strain, exposure to light, radiation, or heating. These properties make possible semiconductor devices that respond to physical inputs rather than electrical signals. Miniature devices or machines can be constructed using wafer fabrication techniques. These devices are called *microelectromechanical systems* (MEMS).[14] MEMS devices made with wafer fabrication techniques include motors, motion sensors (as in air bag sensors), spectrometers, and miniature scanning tunneling microscopes.

Strain gauges

A device that takes advantage of a junction reaction to physical pressure is the strain gauge. The gauge is made by back-etching through the wafer until only a very thin membrane is left. A junction and supporting circuitry are formed in the membrane. When the membrane is deflected by some force, as in a weight scale or pressurized gas line, the semiconductor circuit produces an output proportional to the deflection. The output of the circuit is correlated with the amount of pressure on the membrane and displayed on the appropriate meter.

Batteries

Thin film rechargeable batteries have been made from lithium and vanadium pentoxide layers. They could be formed on a package and provide backup for CMOS devices in the event of power failures.[15]

Light-emitting diodes

One effect when certain compound junctions are reversed-biased is the production of photons. Photons are a form of radiation that humans see as light. The devices are called *light-emitting diodes* (LED). These are the displays that are used in consumer electronics equipment and automobile displays.

The devices (Fig. 16.45) are made on gallium-arsenic-phosphide (GaAsP) wafers that are covered with thousands of diodes, wired so that they can be turned on or off individually. Groups of diodes are turned on in groups to form letters and numbers. GaAsP material produces the familiar red displays, with other colors being produced when dopants are added to the wafer.

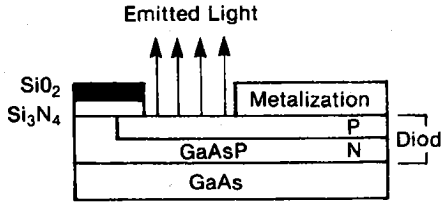


Figure 16.45 LED structure.

Optoelectronics

On the other side of LEDs are chips that react to light. One use sought for years is integrated opto/electronics connection in local area networks (LANs). In this use, the opto devices (sensitive to laser and other lights) are connected to a fiber optic network or waveguide. Like the LEDs, the receivers are mostly III–V based ICs. An on-board conventional IC processes the data and sends it on through an output LED or laser.[16]

Solar cells

Not only can semiconductor junctions emit light, they respond to it. This property is taken advantage of in the solar cell (Fig. 16.46). The cell is composed of diodes formed in a thin layer of semiconducting material such as amorphous silicon. When sunlight strikes the junction region, a current passes across it. The current is captured by on-board circuitry.

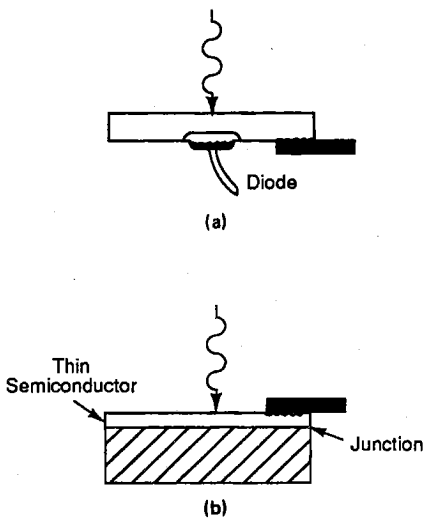


Figure 16.46 Light-sensitive semiconductor structures. (*a*) Photodiode and (*b*) solar cell.

Temperature sensing

Semiconductor junctions are temperature sensitive. Heating a device will produce more current across the junction. This effect is taken advantage of in a variety of devices, such as solid-state medical thermometers and industrial control units.

Acoustic wave devices

Acoustic wave devices (Fig. 16.47) are nonsilicon solid-state components used in microwave communication systems. They serve the function of converting microwaves into electrical impulses. A compound material such as $Be_{12}GeO_{20}$ has the property of reacting to the wave and setting up an electrical response in a solid-state circuit formed on the chip by normal semiconductor processes.

Micromachines

With semiconductor technology, it is possible to create miniature machines with moving parts. It easy to understand the definition of surface patterns. Etching underneath the surface to free up a part is more difficult and the greatest challenge of this application. Nevertheless, micromachines are being produced. For example, Texas Instruments has produced an array consisting of numerous cells, each with a floating thin layer panel suspended from a "hinge" at each corner. A thin oxide layer gives the panel the ability to refract light. On-board electronic circuitry allows control of signals to the hinges. The signal causes a physical force to tilt the layer in specific directions that creates specific colors through the refraction process. The whole array functions as a high-resolution solid-state screen that can be used in projectors. Projecting the image onto a mirror provides a larger view-

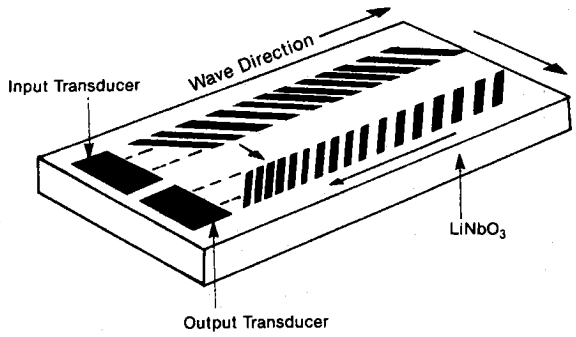


Figure 16.47 Acoustic wave device.

ing screen. Another recent micromachine is an actuator used to drive an optical shutter. Due to the light weight of the parts, speeds of up to 200,000 revolutions per minute have been made.[17] Shown below (Fig. 16.48) is a MEMS lens/laser system used as an optical bench. One intriguing potential is miniature space satellites that can be put into space with low-power rockets.

Flat panel displays

Flat panel displays (FPDs) are most often seen as laptop computer screens. Their future is the replacement of cathode ray tubes (CRTs), where they provide a thin and more rugged option. However, improvements in clarity, brightness, and cost are necessary before they become regular replacements for CRTs.[18]

The majority of FPDs are the color-active matrix liquid crystal displays (AMLCDs). Both color and monochrome LCD displays are made using microchip fabrication technology. A display is a sandwich of two glass plates with a liquid crystal material in between (Fig. 16.49). On the bottom plate is an array of simple MOS devices with on-board driver circuitry. The transistors are formed by standard process steps, deposition, patterning, and so forth. The top panel is a color filter (for

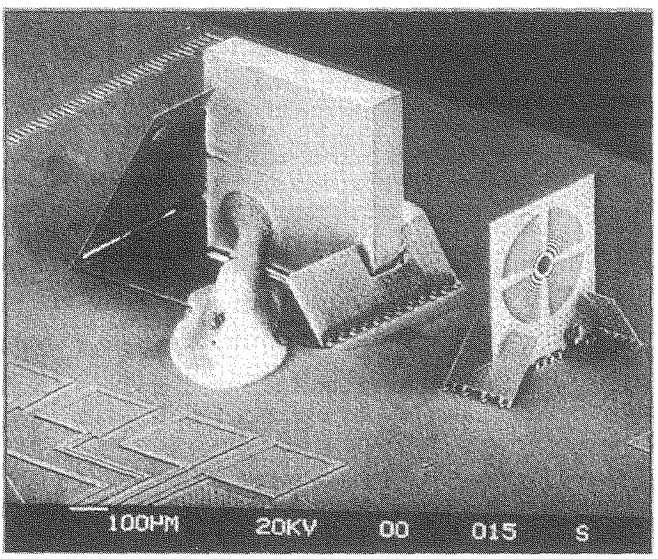


Figure 16.48 MEMS optical guide system. *(Reprinted from the September 1997 edition of* Solid State Technology, *copyright 1997 by PennWell Publishing Company.)*

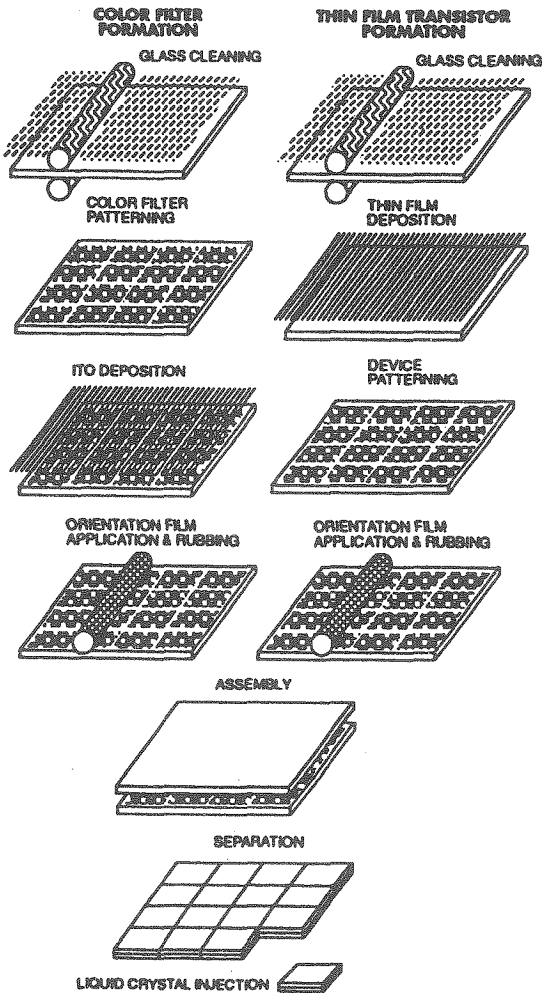


Figure 16.49 Active matrix liquid crystal display (AM-LCD) process flow. *(Source: O'Mera Associates.)*

the color display) formed by multiple photoresist steps. In between is sandwiched the liquid crystal.

The image displayed on the screen is first formed as a pattern in the MOS transistors (on/off conditions).[19] This pattern is replicated by the liquid crystal, which can change states in response to an electrical signal. Lastly, for the color displays, the top filter transmits the proper color. Black-and-white displays simply transmit the on/off condition of the LCD.

Field emission devices (FEDs)

In CRTs and FPDs, the information required to put a pattern on the screen has to come from electronic signals (such as a computer) and be transformed into a physical entity. In conventional CRTs, the entity is a stream of electrons from an electron gun. An alternative device to produce electrons is an FED (Fig. 16.50). Formed from semiconductor materials using wafer fabrication techniques, the end result is an etched microtip. When a voltage is placed between the cathode and gate, electrons are emitted. FPDs made with FEDs have the promise of brighter displays, a wider viewing angle, and improved speed.[20]

Process flows

Increased device and circuit performance has created new structures with many new "parts." A typical structure circa 2000 is shown in Fig. 16.51. Creating the structures requires more and more process steps. Figure 16.52 shows the SIA roadmap projections for the chip and process in the year 2112.[21] Keeping track and making sense of this structure and process complexity requires a good grounding in the basic electrical operations of the individual components and basic processes.

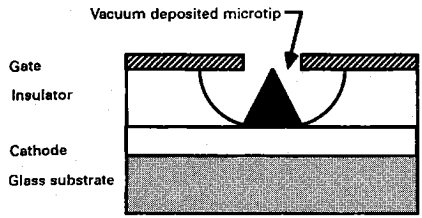


Figure 16.50 Cross section of field emission device.

Review Questions

1. Sketch and name the three parts and functions of a bipolar and MOS transistor.
2. Explain the function of a diode, resistor, and capacitor.
3. What is the operational advantage of a Schottky barrier diode?
4. Sketch and name the layers in a metal, polycide, and salicide MOS gate structure.
5. What is the structural difference between FETMOS, JFET, and MESFET gate structures?
6. Describe why component isolation is required in integrated circuits.

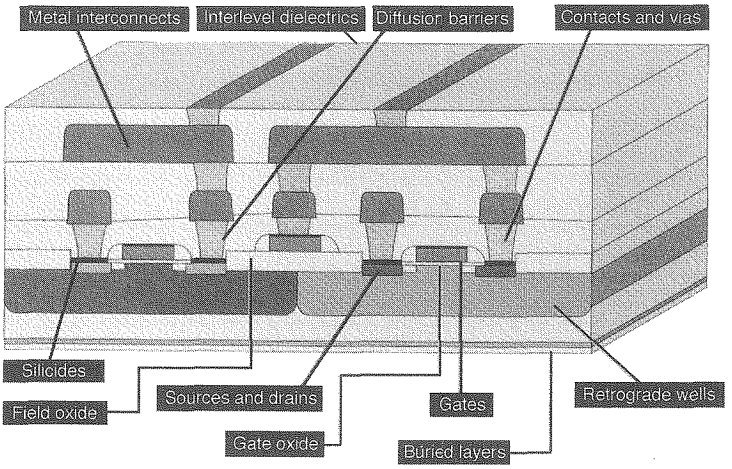


Figure 16.51 Two-metal structure. *(Courtesy of Semiconductor Services, January 1998.)*

Year of Production	2001	2006	2012
Line width (nm)	150	100	50
Memory size	1 Gb	16Gb	64 Gb
Logic Bits/cm2	380M	2.2B	17B
Chip Size-DRAM (mm2)	445	790	1580
Max wiring levels	7	7-8	9
Mask layers	23	24/26	28
Defect density-DRAM (D/m)	875	490	250
Chip conections-I/O's	1195	1970	3585
Wafer diameter (mm)	300	300	450

Figure 16.52 SIA roadmap projections. *(Semiconductor International, January 1998.)*

7. What are the differences and advantages of junction and dielectric isolation schemes?
8. Sketch the process steps used to create an LOCOS isolation scheme.
9. Name three ways that a semiconductor junction in operation can be influenced.
10. How does a CMOS structure differ from an NMOS structure?

References

1. H. R. Camenzind, *Electronic Integrated Systems Design,* Van Nostrand Reinhold, Princeton, NJ, p. 85.
2. P. Singer, "Gearing Up for Gigabits," *Semiconductor International,* November 1994, p. 34.
3. S. De Ornellas, et al., Plasma Etch of Ferroelectric Capacitors in FeRAMs and DRAMs," *Semiconductor International,* September 1997, p. 103.
4. H. R. Camenzind, *Electronic Integrated Systems Design,* Van Nostrand Reinhold, Princeton, NJ, p. 141.
5. P. Singer, "Directions in Dielectrics in CMOS and DRAMs," *Semiconductor International,* April 1994, p. 56.
6. Y. Pauleau, "Interconnect Materials for VLSI Circuits," *Solid State Technology,* April 1987, p. 157.
7. "Industry News," *Semiconductor International,* Cahners Publishing, April 1994, p. 16.
8. S. K. Ghandhi, *VLSI Fabrication Principles,* John Wiley & Sons, Inc., New York, 1994, p. 717.
9. S. Wolf, "A Review of IC Isolation Technologies—Part 8," *Solid State Technology,* PennWell Publishing, June 1993, p. 97.
10. L. Peters, "High Hopes for High Energy Ion Implantation," *Semiconductor International,* Cahners Publishing, June 1993, p. 84.
11. C. B. Yarling, "M. I. Current, Ion Implantation for the Challenges of ULSI and 200 mm Wafer Production," *Microelectronic Manufacturing and Testing,* March 1988, p. 15.
12. K. Yallup, "SOI Provide Total Dielectric Isolation," *Semiconductor International,* Cahners Publishing, July 1993, p. 134.
13. P. W. Anderson, "Electronic and Superconductors," in E. Ante'bi (Ed.), *The Electronic Epoch,* Van Nostrand Reinhold, Princeton, NJ, p. 66.
14. K. Gabriel, "Engineering Microscopic Machines," *Scientific American,* September 1995, p. 150.
15. J. Bates, "Rechargeable Thin-Film Lithium Microbatteries," *Solid State Technology,* PennWell Publishing, July 1993, p. 59.
16. P. Singer, "The Optoelectronics Industry: Has It Seen the Light?," *Semiconductor International,* Cahners Publishing, July 1993, p. 70.
17. J. Sniegowski, "Moving the World With Surface Micromachining," *Solid State Technology,* February 1996, p. 83.
18. P. Singer, "Flat Panel Display: An Interesting Test Case for the U.S.," *Semiconductor International,* Cahners Publishing, July 1994, p. 79.
19. *Ibid.*
20. K. Derbyshire, "Beyond AMLCDs: Field Emission Displays?" *Solid State Technology,* PennWell Publishing, November 1994, p. 55.
21. L. Peters, "Speeding the Transition to 0.18 μm," *Semiconductor International,* January 1998, p. 66.

Chapter

17

Introduction to Integrated Circuits

Overview

The solid-state semiconductor components wired into integrated circuits perform many different functions in electronic instruments and machines. In this chapter, the general integrated circuit families and their different functions are explained.

Objectives

Upon completion of this chapter, you should be able to:

1. Explain the concept of binary numbering.
2. List the three major integrated circuit functions.
3. Compare the basis of analog and digital logic circuits.
4. List the user and production advantages of logic gate arrays and PROM circuits.
5. Explain the two major memory circuit types.
6. Make a list of the four nonvolatile memory circuits.
7. Compare the performance and cost factors of DRAM and SRAM memory circuits.

Introduction

The primary products of the semiconductor industry are integrated circuits. Countless numbers and types of circuits can be created using the processes described in this text. A circuit catalog from a major integrated circuit (IC) producer such as National Semiconductor or Mo-

torola is as large as a New York City phone book. IBM estimates that their internal circuit catalog lists over 50,000 separate circuits!

Becoming familiar with integrated circuits is not as awesome a task as these high numbers might imply. The burden is eased by the fact that most circuits fall into three primary families: logic, memory, and microprocessors (logic and memory) (Fig. 17.1). Within each circuit family, there are a few principal designs and functions. The multiplicity of circuits comes from the many parameter variations required for specific uses.

The major functional circuit categories and their circuit designs are explained in this chapter. In the last section, we look at the future of IC circuitry from the perspective of the industry today. What the circuits will actually be like in 2010 can only be imagined, just as, in 1950, no one predicted the megabit RAM or the microprocessor.

Circuit Basics

The question of how an integrated circuit actually works is the subject of other texts. But all circuits are based on the processing of data in binary notation. The binary system is a way of representing any number with just two digits—for example a zero and a one. It is actually an accounting system that keeps track of the place and value of the components of a number. Numbers can be expressed as the sums of numbers. For example,

$$\begin{aligned} 1 &= 1+0 \\ 3 &= 2+1 \\ 7 &= 4+2+1 \\ 10 &= 8+2 \end{aligned}$$

Another way to express numbers is as powers of their factors. Yet another way is to express numbers as powers of their roots.

The basis of binary notation is the powers of the number 2. Figure 17.2 shows the numbers 1, 2, 4, 8 expressed as powers of 2. We could also represent those numbers just as the power of 2. Thus, 1 becomes 0, 2 becomes 1, 4 becomes 2, and 8 becomes 3. Now, for the clever part, any number can be expressed as the sum of some numbers that are powers of 2. The number 25 is the sum of $16(2^4) + 8(2^3) + 1(2^0)$. In the number 25, there is one 2^4, one 2^3, zero 2^2, and one 2^0. Or

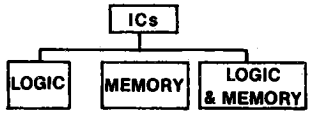


Figure 17.1 IC circuit functions.

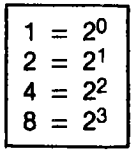


Figure 17.2 Powers of 2.

the number 25 can be represented by the code 1101, each of the digits representing the presence or absence of a particular power of 2. The chart in Fig. 17.3 lists some numbers in binary notation.

Translating numbers into binary notation is easily accomplished by establishing a grid with each column representing a power of 2. The actual number is represented by a string of zeros or ones that indicate the presence or absence of the particular powers of 2 that make up the number.

Binary notation has been known for centuries. Buckminster Fuller, in his book, *Synergistics,* has an amusing account of the use of binary coding by the ancient Phoenicians to keep track of cargo amounts. He claims that the Phoenician sailors were considered stupid because they could not count in the system of the day, when actually they were accurately keeping track of large amounts of cargo with only two "numbers"; they were counting in binary notation.

With binary notation, only two numbers are necessary—a one and a zero. In the discussion above, binary notation was represented by the numbers zero and one. In the physical world, binary numbers can be represented by any system that has two conditions. Figure 17.4 shows several different ways to code the number 7. The last row could represent binary coding by the off-on states of a transistor or memory cell.

Inside a circuit, numbers are coded, stored, and manipulated in their binary code. The numbers can be stored and manipulated because capacitors can be charged to have a charge or not have a charge, and transistors can be either on or off. The smallest piece of information in a circuit is called a "binary digit" or "bit." The binary coding system is simple. The problem of how the coded numbers could be added, subtracted, and multiplied was solved by George Boole, a nine-

Standard Power of 2	32 2^5	16 2^4	8 2^3	4 2^2	2 2^1	1 2^0
Number Standard = Binary Number						
1	0	0	0	0	0	1
7	0	0	0	1	1	1
18	0	1	0	0	1	0
33	1	0	0	0	0	1

Figure 17.3 Binary notation.

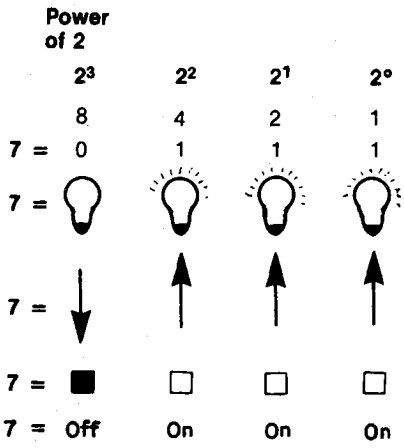


Figure 17.4 Binary representations of the number 7.

teenth-century mathematician. He developed a logic system capable of handling numbers in binary notation. Until the development of computer logic, his Boolean logic (or Boolean algebra) was an academic curiosity.

Chips and computers are designed to handle a specific size of binary number or word. An eight-bit machine manipulates numbers with eight binary bits at a time. A 32-bit machine can handle a number composed of 32 binary bits. The more bits a machine can handle at one time, the faster and more powerfully it processes data. Every eight bits is known as a byte. Thus, an eight-megabyte (8 MB) storage capacity can hold eight million bits of information.

Integrated Circuit Types

A solid-state integrated circuit is composed of a number of separate functional areas. Each chip, regardless of the circuit function, has an input and encode section where the incoming signals are "coded" into a form that the circuit can understand. The majority of the circuit area contains the circuitry required to perform the circuit function, either memory or logic. After the data is manipulated by the circuit, it goes to a decode section where it is changed back into a form that is usable by the machine's output mechanism. The circuit's output section sends the data to the outside world.

Although this is an overly simplified explanation of a circuit, it illustrates the fact that the interior of a chip is composed of definite separate functional areas. In many circuits, these areas perform the same functions as the main parts of a computer. Circuit types fall into three broad categories: logic, memory, and logic and memory (microprocessors).

Logic circuits perform a specified logical operation on the incoming data. For example, pushing the plus (+) key on a calculator instructs the logic portion of the chip to add the numbers presented to it. An on-board automobile computer goes through a logical operation to direct the signal from a sensor indicating an open door to light up the correct warning light on the dashboard.

Memory circuits are designed to store and give back data in the same form in which it is entered. Pushing the pi (π) key on a calculator activates the memory part of the circuit where the value of π (3.14) is stored. The value 3.14 is displayed on the screen. Every time that key is activated, that value is displayed.

Microprocessors are a third circuit type that combines both logic, arithmetic functions, and memory. In 1972, Intel Corporation introduced the first practical microprocessor. It was the microprocessor that allowed the design of powerful personal computers, digital watches, and one-chip calculators and the transfer of so many business machines to solid-state electronics, from phone systems to vending machines. Microprocessors can be programmed to perform many different circuit functions. To accomplish this, they contain logic and memory circuitry as well as the necessary encode, decode, input, and output sections. The microprocessor has been dubbed "a computer on a chip." While it contains all the functional areas of a computer, it is not truly a complete computer. Even simple computers require vast amounts of memory capacity that microprocessors do not have. Within many personal computers, microprocessors function as the central processing unit (CPU). Additional memory chips have to be included to make the computer of practical use.

Actually, every integrated circuit contains both logic capability and memory sections. For example, the logic circuitry of a calculator must have certain constants stored in a memory section to perform calculations. And memory circuits must have some logic functions to direct the flow of electrons and holes to the right parts of the circuit for storage.

Logic circuits

Analog-digital logic circuits. Logic circuits fall into two main categories: analog and digital (Fig. 17.5). Analog logic circuits were the earliest logic circuits developed. An analog circuit has an output that is proportional to the input. Digital circuits, on the other hand, feature a

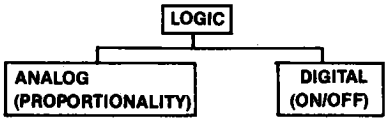


Figure 17.5 Logic circuit types.

predetermined output in response to a variety of inputs. A wall light dimmer is an analog device. Turning the control varies the voltage to the dimmer, which in turn varies the brightness of the light. A standard on-off light switch is a digital device. Only two brightness conditions are possible: on or off. Most audio circuits are of the analog type. Changing the level setting of the volume produces a proportional change in the sound coming out of the speaker.

Analog logic circuit types. Analog circuits were the first type designed in integrated form. The home computer hobby kits of the 1950s were the analog type. These simple circuits were based on Ohm's law *(R = V/I).* The circuit contains a resistance meter and a means for generating a current and measuring a voltage. The three quantities are related by Ohm's law. Any other three variables similarly related can be represented by the resistance, voltage, and current. Varying one changes the other two. The circuit thus becomes a computer for solving any equation of the form $A = B/C$.

The accuracy of analog circuits is dependent on the precision of the relationship between the input and the output. In the simple computer illustrated, accuracy is dependent on the precision of the components in the circuit, the clarity of the meters for setting the input and reading the output, and the immunity of the circuit to outside "noise." Unless the circuit contains a section to regulate incoming voltage levels, a change in the line voltage would alter the output, and hence the accuracy.

Both simple and complex analog circuits are vulnerable to variations in the incoming signal and to internal noise. Analog circuits are also dependent on precise control of the resistor values. Unfortunately, diffused resistors cannot be fabricated with a resistance variation from design value better than 3 to 5 percent, which is unacceptable for many applications. Greater resistance precision is gained by the use of matched resistor pairs, in which the effective resistance in the circuit is the difference between two resistors. This difference can be more tightly controlled than with one resistor alone.

Ion implantation also provides the analog circuit designer with a tool for producing resistors with a higher degree of control. Many analog circuits feature thin film resistors to achieve the required precision. The growth and popularity of digital circuits is based on their ability to produce a set output every time, regardless of noise or signal variations in the circuit. If a 5-V output is required, the digital circuit will deliver exactly that level.

Digital circuits, however, do not respond as fast as linear circuits. The term used in electronics is *real-time response.* In some applications,

such as airplane controls, real-time response is mandatory. Recent developments in digital circuit speed is facilitating the encroachment of digital circuitry into this traditional use of analog circuits. A major advantage of digital circuits over analog circuits is in general-purpose computers. Analog circuits are more difficult to design to respond to a general range of problems. All modern general-purpose computers are based on digital circuits.

The most popular use for analog circuits is in amplifiers. They are designed in a variety of configurations, for many different applications. All have the same basic principle—the incoming signal or pulse is amplified. Audio circuits require amplification of a weak signal from the record tone arm or other input so as to produce the level required to operate a speaker.

The real-time aspect of analog circuits also makes them real-world circuits. Wherever there is a real-world measurement, such as temperature or movement, analog circuits are used. Even when the majority of the circuitry in a system is digital, analog circuits are often part of the interface with the outside world.

Most analog amplifier circuits are of the differential operational type. These circuits produce an output voltage amplified from and proportional to the difference of two input signals. Bipolar technology is favored for these circuits, because bipolar circuits are modular electric current devices and are better suited to the applications required of analog circuits.

The output signal of an analog device can have a "one-to-one" relationship with the input signal. These circuits are called *linear*. If the input is changed, the output changes in a linear manner. So many analog circuits are of linear design that the two terms are often used interchangeably. However, there are nonlinear circuits, such as those featuring a logarithmic relationship between the input and output.

Logic circuitry is built around the logic gate. A gate controls and directs passage through a barrier. The size and design of a gate influence the amount of passage allowed. For example, a room with many "in" doors and only one "out" door is a gate. Many people can enter the room, but their exit is restricted, because only one door is provided. This example gate can also be operated in reverse, allowing people to enter through only one door and leave through many.

Electronic logic gates perform similar functions but with electrical signals. In a circuit, they perform the necessary logic operation by the dictates of Boolean logic. A discussion of their incorporation in logic design is beyond the scope of this text. The point for this text is that gates, both analog and digital, are formed in a logic circuit by wiring together various components.

Custom-semicustom logic. Using any of the logic gate approaches listed, hundreds of thousands of different logic circuits can be constructed. They vary from custom small-volume circuits to off-the-shelf standards. The bulk of logic circuits require some degree of customization. Several design and fabrication approaches are used to deliver custom and semicustom circuits to the customer at reasonable costs.

The approaches are:

1. Full custom
2. Standard circuit—custom gate pattern
3. Standard circuit—selective wiring gate
4. Programmable array logic

Full custom. A full custom-designed logic circuit is specified by the customer, who pays for design and mask-making fees along with the fabrication costs. This approach is expensive and lengthy and is not geared for experimenting with different circuits in the design stage of a project. Custom-designed circuits are not cost effective in quantities of less than 100,000.

Standard circuit—custom gate pattern. This fabrication process starts with a standard logic circuit design, but only the gates required for the particular application are formed during the fabrication process. The input, output, and other circuit sections are standard for a family of circuits.

Standard circuit—selective wiring gate arrays. This system is similar to the custom gate approach but is based on a standard circuit design for most of the fabrication process. These circuits are built with a standard number of gates. This gate section is called the *array,* and the circuit is known as a *gate array.* Working with the basic design, customers can instruct the fabrication department to wire together only the gates required to produce their custom circuit logic function.

The result is faster turnaround time than full custom processes can achieve, and moderate cost. The cost per logic function of gate arrays is higher than that of a custom circuit produced in production quantities. The larger gate section required to allow many different circuit variations results in a larger chip. This larger chip size leads to a higher manufacturing cost per chip and/or a lower yield.

The wafers receive a common process up to the contact mask. The contact mask is customized to form contacts only to the gates required

for the particular circuit. After metal deposition, only the gates with contacts are wired into the circuit. A variation of this process is to open the contacts in all the gates but use a customized metal mask that wires in only the wanted gates.

Programmable array logic (PAL). Each of the three systems described requires the chip manufacturer to do the "customizing." This requirement can result in delivery or scheduling problems and generally forces the user to buy a minimum quantity of parts. Monolithic Memories, Inc., addressed this problem with the introduction of their PAL line of circuits in 1978. PAL stands for programmable array logic. MMI applied the programmable fuse technology used in memory products to logic circuits. The result was a field programmable (custom) logic circuit.

The concept is similar to the standard gate array, but, within the circuit, a fuse connects each logic gate into the circuit. The user programs the circuit by blowing the fuses at the unneeded gates, thus removing them from the circuit.

Memory circuits

Around 1960, industry forecasters began predicting that solid state memory circuits would overtake the traditional core memory. The advantages of solid circuits were their reliability, small size, and faster speed. This prediction was made every year until the early 1970s, when solid state memories finally did surpass core memory. The major factor that prolonged the life of core memory was lower cost.

Metal-oxide semiconductor (MOS) is the favored transistor structure for memory circuits. During the 1960s, however, the cleanliness requirement for MOS processing was not reliably available. High-yield MOS processing also requires accurate alignment and clean thin gate oxides. These processes were not fully developed in those years. The resultant low process yields kept MOS memories more expensive than core memories.

With process improvements and improved silicon gate structures, complementary metal-oxide semiconductor (CMOS) is the memory technology of choice. Some bipolar memories are favored for their fast speed and switching capabilities. While logic circuits can be (and are) made in MOS technology, most MOS circuits produced are memories, with the majority incorporated into computers. They are also used in microprocessor-based products, which require auxiliary memory chips. There are two principal types of memory circuits: volatile and nonvolatile (Fig. 17.6).

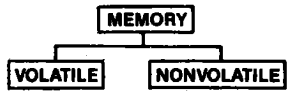


Figure 17.6 Memory circuit types.

Nonvolatile memories. A nonvolatile memory device is one that does not lose its stored information when it loses its power. An example of this is a compact disk (CD), which is an information-storage device. If power to the record player is lost, the songs are not lost from the record itself. A number of nonvolatile memory circuits are listed in Fig. 17.7.

ROM. In integrated circuits, the ROM design is the principal nonvolatile circuit. ROM stands for *read-only memory.* The sole function of this type of circuit is to give back precoded information. The information required in the circuit is specifically designed into the chip memory array section during fabrication. Once the chip is made, the stored information is a permanent part of the circuit.

Other memory types have *read and write* capability. That is, they can receive and store information from an input device (keyboard, magnetic tape, floppy disk, and so on). A phonograph record is a nonvolatile ROM device. A magnetic tape is an example of a nonvolatile device with both read and write capabilities, because information can be erased and rerecorded.

In a calculator, the constants and the rules required for the math operations are available in the ROM section. ROM circuits, like logic circuits, number in the hundreds of thousands. Although there are many standard types, the industry also produces many custom ROM circuits. The choices offered to the user in selecting a standard or custom chip are similar to those available with logic circuits. The user can buy a standard circuit, specify a variation on a standard basic circuit, design a total custom circuit, or buy a PROM, EPROM, or EEPROM.

PROM. PROM stands for *programmable read-only memory.* A PROM is the memory equivalent of a PAL. Each memory cell is connected

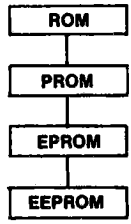


Figure 17.7 Nonvolatile memory.

into the circuit through a fuse. Users program the PROM to their own memory circuit requirements by blowing fuses at the unwanted memory cell locations. After programming, the PROM is changed to a ROM, and the information is permanently coded in the chip, where it becomes a read-only memory.

EPROM. For some applications, it is convenient to change the information stored in the ROM without having to replace the whole chip. EPROM (erasable programmable ROM) chips are designed for this use. The erasable feature is built in with the use of MMOS (memory MOS) transistors detailed in Chapter 16. The transistors can be selectively charged (or programmed), and they hold the charge for a long time—which holds the information in a nonvolatile fashion. Programming is by the mechanism of hot electron injection. When reprogramming is required, the charge in the transistors can be drained off (erasing the memory) by shining ultraviolet light on the chip. Reprogramming of the chip takes place by removing it from the circuit and putting in new memory information with an external programming machine. A typical EPROM can be reprogrammed up to ten times.

EEPROM. The next level of convenience in memory design is the ability to program and reprogram the chip while it is in a socket in the machine. This convenience is available with the EEPROM (or E2PROM), standing for *electronically erasable PROM.* Programming and erasing take place by pulses from the outside that place charges in selected memory cells or drain the charges away. Programming is by the same mechanism used for EPROMs, hot electron infection. Charge is drawn from the memory cell be a mechanism called Fowler-Nordheim tunneling. This connivance comes at the expense of a larger memory cell size and a commensurate reduction in chip density.

Flash memories. A flash memory is a form of EEPROM. It is a one-transistor cell design,[1] and like EPROMs can be programmed and reprogrammed while in their sockets in a system. Additionally, blocks, or the entire array, can be erased at one time.

Volatile memories. Semiconductor circuit and computer design involves the constant evaluation of trade-offs. In the case of memory, nonvolatile memory provides protection against power loss, but these memories are frequently slow and not very dense. More importantly, none of the circuits described above has a write capability, an essential feature in operating a computer. New information, such as a change in

pay status, must be conveniently entered into the computer and stored temporarily while the new check is being written. Memory must also be easily erasable so the computer can quickly process new information or accept a completely new program. Several memory circuit designs are used to produce fast and high-density memory circuits. Both are of the volatile type; that is, when power is lost to the chip, all the stored information is lost; information presented on a computer screen, and not saved, is eligible for loss if the power to the computer goes off.

RAM. One type of circuit used for high-density memory (Fig. 17.8) storage is random-access memory, or RAM. "Random" refers to the ability of the computer to directly retrieve any information stored in the circuit. Unlike a serial memory, the RAM design allows the chip to find the exact information asked for, wherever it is located in the computer memory. This feature allows faster information retrieval and makes the RAM the principal memory circuit in computers.

Dynamic random-access memory (DRAM). DRAMS come in two principal designs: dynamic and static (Fig. 17.8). A dynamic memory design called a DRAM, for dynamic RAM, is used in great quantities in computer memories. The memory cell design is based on only one transistor and a small capacitor (Fig. 17.9). The information is stored in it by a charge built up in the capacitor. Unfortunately, the charge drains away very rapidly. To combat this problem, the memory information must be re-inputted to the circuit on a constant basis. The term for this function is *refresh*. The refreshing of the circuit occurs many thousands of times per second. Dynamic RAMS are vulnerable to both power loss and interruption, and to problems with the refreshing circuit.

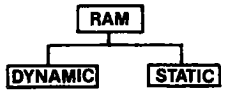


Figure 17.8 RAM cell designs.

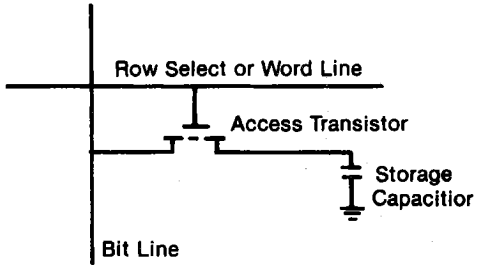


Figure 17.9 Dynamic RAM cell schematic.

The goal of DRAM design is small-cell design for high-density and closely spaced components with small and thin parts for speed. These requirements have driven DRAM design and processing to the highest levels of the technology. All the advantages available by advanced, state-of-the-art equipment and processing are applied to DRAM circuits. This fact make them the industry's leading-edge circuit.

Static random-access memory (SRAM). Static random-access RAM (SRAM) memories are based on a cell design that does not need a refresh function. Once the information is put into the chip, it will stay as long as the power remains on. This is accomplished with a cell containing several transistors and capacitors (Fig. 17.10). The information is stored as conditions of the transistors are alternately on or off. Information can be read and written with a SRAM cell much faster than with a RAM design, since transistors can be switched faster than capacitors can be charged and drained. The penalty paid for this lesser degree of volatility and speed is loss of space. The larger cell design makes static memories less dense than DRAMs.

Memory capacity is measured by the number of bits that can be stored. A 1k RAM has a capacity of 1024 bits of information; 1024 is a power of 2. A 64k RAM actually has a capacity of 65,536 bits of information. RAM capacity has expanded rapidly, with larger megabit memories (64 and higher) expected to be produced in quantities with presently identified technology. Each step upward in RAM capacity places greater pressure on wafer processing and yield improvement. The nature of the semiconductor chip business is exemplified by the 64k RAM, introduced by IBM in 1977. The chips were soon available in the merchant market, priced at over $100 per circuit. By 1985, competition and yield improvements had lowered the price to under $1 per circuit!

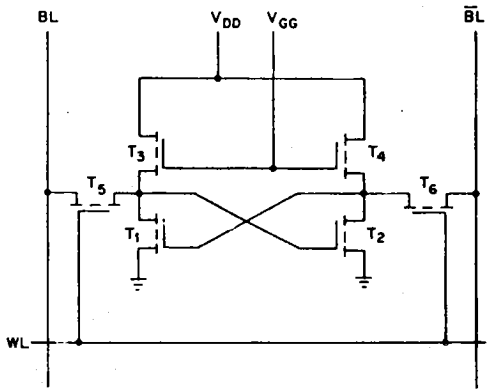


Figure 17.10 Static RAM cell schematic.

Ferroelectric memories (FERAM and DRAM). Ferroelectric materials (see Chapter 16) used in capacitors create memories that operate faster than conventional SiCMOS technology capacitors.[2] The challenge is the integration of this nonsilicon technology into a standard silicon process.

Redundancy

Redundancy is the inclusion of extra circuit components in the design. If one or more of the components don't work, others are available that do. The trade-off for redundancy is larger chip size. Also, extra circuitry is required within the main circuit to detect the functioning and nonfunctioning components and direct the selection of a functioning component. Although this approach to higher yield has been discussed for years, it hasn't yet become a mainstay of circuit design. This is due to the problem of locating the working and nonworking components and wiring the working ones into the circuit.

The Next Generation

Since the 1950s, the semiconductor industry has maintained a rather constant rate of product evolution. In general, there has been a new product generation every three years.[3] Within each generation, the density of memory chips increased four times and logic chips two to three times. Every two generations (six years), the feature size decreased by a factor of two, with chip area and package pin count increasing by a factor of two.

Predicting the future is always difficult, but there are identified end points for chip circuits as we now know them. The SIA technology roadmap[4] profiles development to the 50-nanometer feature size level by the year 2012. At that size, chip densities of 16G (billion) bits for memory and 20M (million) gates for logic. Wafer diameters climb to 200 to 400 mm, and die sizes are in the 1000 mm^2 range (1.2 inches per side) or larger.

Reaching this level will require development of X-ray lithography and low-resistance metals, low-leakage contacts, and ultraclean materials and processes. Planarization techniques will be strained as the number of metal interconnect levels climbs to six or seven. Perhaps the major challenge is reliable, stable gate oxides in the 35-Å range. Below 35 Å, carriers can tunnel through the layer, causing leakage problems. These technical problems must be overcome in a manufacturing atmosphere that is increasingly expensive and that requires automation and extensive measurement and process control programs.

Circuits coming out of the advanced technology will be a mixture of increased performance existing circuits and new circuits. Certainly, DRAMs will continue to increase in memory capacity, and microprocessors will become more powerful and faster. New areas for chip development are speech recognition, expert systems, and the continuing "microchipization" of automobiles and most power-consuming products (control and conservation).

Review Questions

1. Name the two principal solid-state circuit types.
2. Name the two principal logical circuit types.
3. Give an example of a digital electrical device.
4. Which of the following describes a PROM circuit?

   a. Factory programmable

   b. Field programmable

   c. *a* and *b*

   d. Not programmable
5. List the three functions in a microprocessor circuit.
6. What does RAM stand for?
7. What is the difference between a DRAM and SRAM?
8. How does a serial memory device read back information? Give an example.
9. Why are DRAM circuits considered the leading-edge ULSI circuits?
10. Express the number "25" in binary code.

References

1. M. McConnell, et al., "An Experimental 4-Mb Flash EEPROM with Sector Erase," *IEEE Journal of Solid State Circuits,* vol. 26, no. 4, April 1991.
2. R. Jones, "Integration of Ferroelectric Nonvolatile Memories," *Solid State Technology,* October 1997, p. 201.
3. C. Hu, "MOSFET Scaling in the Next Decade and Beyond," *Semiconductor International,* June 1994, p. 105.
4. Semiconductor Industry Association, Palo Alto, CA, 1992.

Chapter

18

Packaging

Overview

After wafer sort, the chips on the wafer surface are completed and the electrically functioning ones identified. The chips will be incorporated into individual protective packages, mounted with other components in hybrid or multichip modules (MCMs), or connected directly onto a printed circuit board. This chapter describes the packages and processes used for chip protection and to connect the chip to electrical connections which, in turn, allow integration of the circuit into a larger electronic system.

Objectives

Upon completion of this chapter, you should be able to:

1. List the four functions of a semiconductor package.
2. List the five common parts of a package.
3. Recognize and identify the major package designs.
4. List and describe the major packaging process flows.

Introduction

After wafer sort, the chips are still part of the wafer. For use in a circuit or electronic product, they must be separated from the wafer and, in most cases, put in a protective package. They may also be mounted onto the surface of a ceramic substrate as part of a hybrid circuit, put into a large package with other chips as part of a multichip module (MCM), or be connected directly on board a printed circuit, chip-on-

board, or direct chip attach (COB or DCA) (Fig. 18.1). All three options share some common processes. The packaging process, in addition to protecting the chip, provides an electrical connection system allowing the chip to be integrated into an electronic system, and it provides environmental protection and heat dissipation. This series of processes is known variously as *packaging, assembly,* or the *back-end process.* In the packaging process, the chips are called *dies* or *dice.*

Over the years, semiconductor packaging has lagged wafer fabrication in process sophistication and manufacturing demands. The advent of the VLSI/ULSI era in chip density has forced a radical upgrading of chip packaging technology and production automation.

Higher-density chips require more input connections (I) and more output connections (O). These are referred to as the I/O count or simply the pin count. The SIA *Technology Roadmap* (ITRS) projects pin counts increasing into the 3000 range by 2007 (Fig. 18.2). The IRTS lists pin count, cost, chip size, thickness, and temperature considerations as the primary physical drivers of packaging technology. As solid-state circuits have found more applications, the need for special package designs has increased. Higher pin counts have led to the adoption of bump/flip chip technology. Size and speed considerations have driven the use of chip scale packages in consumer products, such as cell phones and hand-held products. The harsh environments of space, automotive use, and military applications require special packages, processing, and testing to ensure high reliability in the field. These packages, processes, and tests are referred to as *hi-rel.* The other chips and packages are referred to as *commercial* parts.

No longer is packaging the stepchild of the semiconductor industry. Many feel that, eventually, packaging will be the limiting factor on the

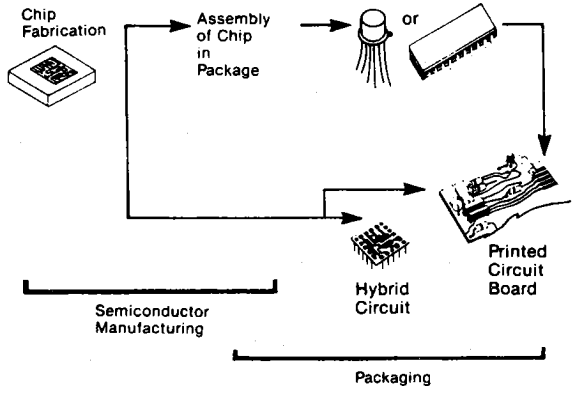


Figure 18.1 Chip packaging options.

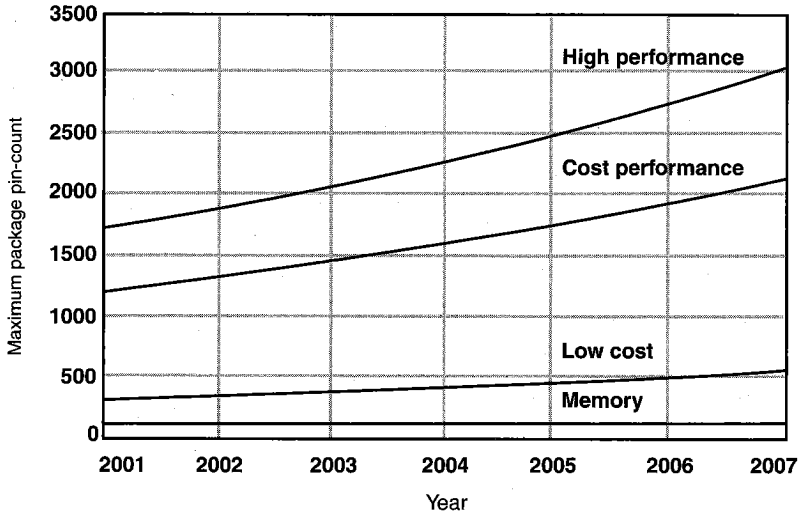


Figure 18.2 Pin-count predictions. (*Courtesy of* Semiconductor International.)

growth of chip size. For the time being, however, much effort is going into new package designs, new material development, and faster and more reliable packaging processes.

Chip Characteristics

Throughout this text, many characteristics of discrete devices and integrated circuits have been mentioned. Several of them have a direct bearing on package design and the packaging processes (Fig. 18.3). The chip density (integration level) determines the number of connections required, with higher-density chips having larger surface areas and more bonding pads. The trend to larger chips has resulted in the need for thicker and larger-diameter wafers. These factors have caused changes in die separation processes, package design, and the need for wafer thinning.

In previous chapters, it was pointed out that the functioning of the chip components (transistors, diodes, capacitors, resistors, and fuses)

- Integration Level
- Wafer Thickness
- Dimensions
- Environmental Sensitivity
- Physical Vulnerability
- Heat Generation
- Heat Sensitivity

Figure 18.3 Chip characterizations affecting the packaging process.

can be altered by various contaminants. Primary among them are chemicals such as sodium and chlorine. Additionally, other chemicals can attack the chip layers, and environmental factors (e.g., particulates, humidity, and static) can ruin chips or change their performance. Other concerns are the influence of light and radiation impinging on the chip surface. Some chips are extremely light or radiation sensitive. This factor is considered in the selection of package materials and processing. A dominant chip characteristic is the extreme vulnerability of its surface to physical abuse. The surface components are only a small distance down into the wafer surface, and the surface wiring can be as thin as 1 micron.

These environmental and physical concerns are addressed in two ways. First is the passivation layer deposited near the end of the fabrication process. This may be a hard layer based on silicon dioxide or silicon nitride. Often, passivation layers are doped with boron, phosphorus, or both to increase their protective properties. Alternatively, it may be a soft layer such as a polyimide (Fig. 18.4). The second method of protecting the chip is provided by a package.

Another chip characteristic of importance to the package design and material is heat generation. Chips used in high-power circuits and highly integrated circuits can generate enough heat to actually damage themselves and the circuit. Package design includes heat dissipation factors. Heat is also an important parameter in packaging processes, with packaging process temperatures limited to 450°C. Above this temperature, the aluminum and silicon contacts on the chip can form an alloy in the wafer surface that causes electrical shorts.

Package Functions and Design

There are four basic functions performed by a semiconductor package. They are to provide:

1. A substantial lead system
2. Physical protection

- Silox
- Vapox
- Pyrox
- Glassivation Layer
- PSG
- BSG
- PBSG

Figure 18.4 Passivation layer names.

3. Environmental protection
4. Heat dissipation

Substantial lead system

The primary function of the package is to allow connection of the chip to a circuit board or directly to an electronic product. This connection cannot be made directly, due to the thin and fragile metal system used to interconnect the components on the chip surface. The metal leads are typically less than 1.5 μm thick and often only 1 μm wide. The thinnest wires available are 0.7 to 1.0 mils in diameter, which is many times larger than the surface wiring. Solder balls used in bump connection technology are about 100 μm in diameter. This difference in wiring sizes is the reason why the chip wiring terminates in the larger bonding pads.

Even though the wires are larger, at 1 mil in diameter, they too are very fragile. This fragility is overcome by a more substantial electrical lead system that serves as the connection of the chip to the outside world by traditional leads, pins, or the balls used in grid array packages (Fig. 18.5). The lead system is an integral part of the package.

Physical protection

The second function of the package is the physical protection of the chip from breakage, particulate contamination, and abuse. Physical protection needs vary from low, as in the case of consumer products, to very stringent, as in the case of automobile circuits, space rockets, and military uses. The protection function is accomplished by securing the chip to a die-attachment area and surrounding the chip, wire bonds, and inner package leads with an appropriate enclosure. The size and eventual use of the chip dictate the choice of materials for the enclosure and the design and size of the package.

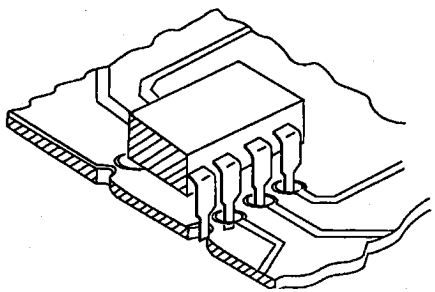


Figure 18.5 DIP through-hole assembly.

Environmental protection

Environmental protection of the chip from chemicals, moisture, and gases that may interfere with the chip functioning is provided by the package enclosure.

Heat dissipation

Every semiconductor chip generates some heat during operation. Some generate large quantities. The package enclosure materials serve to draw the heat away from most chips. Indeed, one of the factors in choosing a package material is its thermal dissipation property. The chips that generate large quantities of heat require additional consideration in the package design. This consideration will influence the size of the package and will often require the addition of metal heat-dissipating fins or blocks on the package.

Overview of Packaging Operations

In this section, a simple overview of the primary packaging process is presented.

Cleanliness and static control

The chips are vulnerable to contamination during their entire lifetime. While assembly areas are not generally required to maintain the same cleanliness levels as fabrication areas (see Chapter 4), cleanliness is important. The table in Fig. 18.6 lists common static-control practices used in packaging areas. High-rel areas, particularly, demand higher cleanliness levels. In fact, many companies are finding that halfway contamination control programs are doomed to failure. Consequently, more assembly areas are practicing very stringent controls, especially for people-generated particles and chemicals.

An environmental danger that is most serious in packaging areas is static. In the fabrication cleanrooms, static is controlled primarily to prevent the attraction of particles to the wafer surface. This is also a concern in a packaging area. But the greatest concern is *electrostatic*

- HEPA Filters/VLF Air
- Smocks, Hats, Shoe Covers
- Finger Cots or Gloves
- Filtered Chemicals
- Sticky Floor Mats
- Static Control

Figure 18.6 Contamination control practices.

discharge, or *ESD.* Static charge can build up to levels of tens of thousands of volts. If this voltage is suddenly discharged onto a chip surface, it can easily destroy a portion of the circuit. MOS gate structures are particularly vulnerable to ESD.

Every packaging area making high-density chips should have an active antistatic program (Fig. 18.7). Those assembling military parts will need to have one as a condition of getting the contract. The antistatic program is based on operators wearing grounding straps and nonstatic smocks; the use of antistatic carrier materials; moving work by lifting rather than sliding; and grounded equipment, work surfaces, and floor mats. Static is also reduced by the placement of ionizers in nitrogen and air blow-off guns (Fig. 18.8) and in the path of air coming out of HEPA filters.

Basic processes

In wafer fabrication, the wafers pass many times in and out of four basic operations (layering, patterning, doping, and heat processing). In packaging as well, there are also several basic operations (Fig. 18.9). However, packaging is a once-through process. Each of the major processes is required only once. As in fabrication, the exact order of the operations is determined by the package type and other factors. Each process may or may not be used in a particular process, and each is

- Wrist Ground Straps
- Nonstatic Garments
- Antistatic Materials
- Grounded Equipment
- Grounded Work Surfaces
- Grounded Floors or Floor Mats

Figure 18.7 Static control practices.

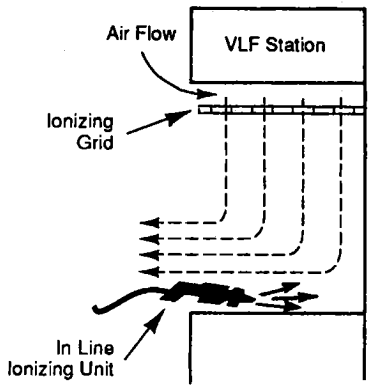


Figure 18.8 Static control techniques.

• Backside Preparation	• Preseal Inspection
• Die Separation	• Packaging Sealing
• Die Pick	• Plating
• Die Attach	• Lead Trim
• Inspection	• Marking
• Bonding	• Final Tests

Figure 18.9 Basic packaging operations.

customized to the particular chip and package requirements. There are three primary methods of connecting the chip to the package: wire bonding, bump (ball) technology, and tape automated bonding (TAB). The process flow(s) based on wire bonding are presented first, followed by the variations required of the bump technique and TAB. The details of the various operations and their optional processes are explained in more detail in the remainder of the chapter.

Backside preparation. At the end of the fabrication process, some wafers have to be thinned (wafer thinning) to fit in the package and/or to remove backside damage or junctions. Wafers whose die are going to be attached to the package by a gold-silicon eutectic will receive a deposited layer of gold (backside gold).

Die separation. The wafer is separated into individual die by sawing or scribe-and-break techniques.

Die pick and place. The functioning die identified at wafer sort are picked from the separated wafer and placed in carriers (Fig. 18.10).

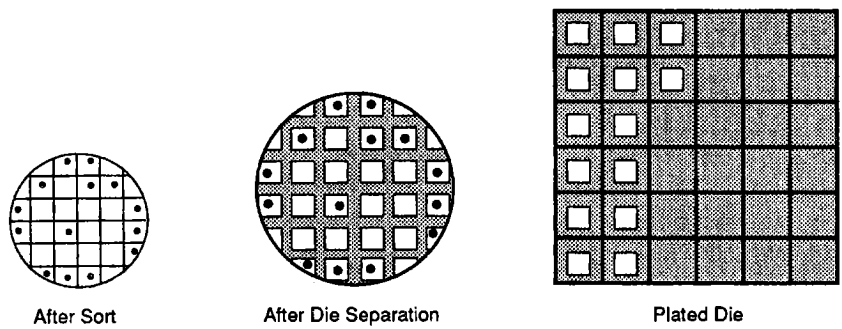


Figure 18.10 Die separation to plate.

Inspection. The dies are optically inspected for edge integrity, contamination, and cosmetic defects.

Die attach. Each die is attached into the die-attach area of the package by either a gold-silicon eutectic layer or an epoxy adhesive material (Fig. 18.11).

Lead bonding. Thin wires are bonded between the chip bonding pads and the inner leads of the package. Bonding is also done by bumps of metal on the bonding pads or by the TAB technique.

Preseal inspection. The bonded die is optically inspected in the package. Criteria checked for are the alignment of the die in the package, proper bond placement, contamination, die-attach quality, and bonding quality.

Plating, trimming, and marking. The outer package leads are plated with an additional layer of conductive metal to improve their solderability into a printed circuit board. Near the end of the process, each package with outer leads will go through a trimming operation to separate the leads from supports. The marking operation is performed to code important information (Fig. 18.12) on the outside of the package enclosure.

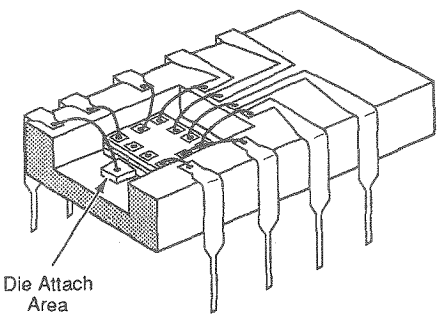


Figure 18.11 Die attach area of package.

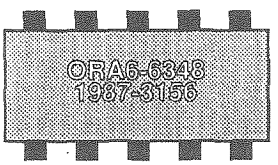


Figure 18.12 Marked package.

Final test. A series of final tests, including electrical and environmental, are performed on each packaged chip to ensure quality. Some packages will receive a "burn-in" test to detect early failures.

Common package parts

The four functions of a chip package are accomplished through the use of a wide variety of package designs. However, most packages have five common parts. They are as described below.

Die-attachment area. In the center of every package is an area where the chip is securely attached into the package. This *die-attachment area* may have an electrical connection that serves to connect the back of the chip to the rest of the lead system. A major requirement for this area is absolute flatness so as to intimately support the chip in the package (Fig. 18.11).

Inner and outer leads. The metal lead system chip is continuous from the die-attach cavity to the printed circuit board or electronic product. The system inner connections are called the *inner leads, bonding lead tips,* or *bond fingers.* The inner leads are generally the narrowest portion of the lead system. The leads become progressively wider, finally ending outside the package. These portions of the lead system are called the *outer leads* (Fig. 18.13). Most of the lead systems are composed of one continuous piece of metal. One exception is the side-brazed package. In this package construction method, the outer leads are brazed onto the interior leads. Two different alloys are used for the outer lead system—either an iron-nickel alloy or a copper alloy. The iron-nickel alloy is desirable for its strength and stability, while copper is used for its electrical and heat-conduction properties.

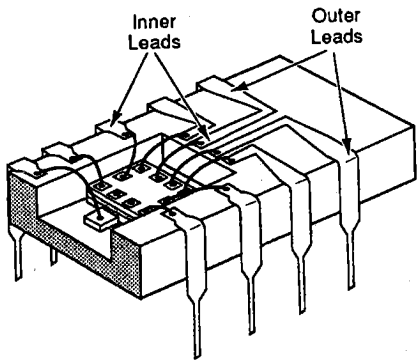


Figure 18.13 Inner and outer leads.

Chip/package connection. The chip is electrically connected to the lead system of the package with bonding wires, balls, or other on-chip connectors (Fig. 18.14).

Enclosures. The die-attach area, bonding wires, and inner and outer leads constitute the electrical parts of the package. The other part is the *enclosure* or *body.* This is the part that provides the protection and heat-dissipation functions. These functions are achieved by several different techniques and package designs as described in the sealing section. The completeness of the seal falls into two categories: hermetic and nonhermetic (Fig. 18.15).

Hermetic sealing results in a package that is impervious to the penetration of moisture and other gases. Hermetic seals are required for chips operating in harsh and demanding environments such as in rockets and space satellites. Metal and ceramic enclosures are the preferred materials for hermetically sealed packages.

Nonhermetically sealed packages are adequate for most consumer applications such as computers and entertainment systems. This sealing system provides good and adequate environmental protection of the chip, except in the most demanding situations. A better term for this type of enclosure sealing method would be *less hermetic.* Nonhermetic packages are composed of epoxy resins or polyimide materials and are generally referred to as *plastic packages.*

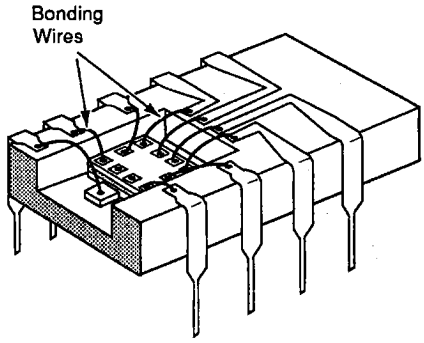


Figure 18.14 Bonding wires.

Hermetic
- Metal
- Ceramic

Nonhermetic
- Epoxy Resins
- Polyimides

Figure 18.15 Package sealing designations.

Packaging Processes

Prepackaging wafer preparation

After the final passivation layer and an alloy step in wafer fabrication, the circuits are complete. However, one or two additional processes may be performed on the wafer before transfer to packaging. These steps (wafer thinning and backside gold) are optional, depending on the wafer thickness and the particular circuit design.

Wafer thinning. The trend to thicker wafers presents several problems in the packaging process. Thicker wafers require the more expensive complete saw-through method at die separation. While sawing produces a higher-quality die edge, the process is more expensive in time and consumption of diamond-tipped saws. Thicker die also require deeper die attach cavities, resulting in a more expensive package. Both of these undesirable results are avoided by thinning the wafers before die separation.

Another situation requiring wafer thinning is electrical in nature. If the wafer backs are not protected as the wafers go through the dopant operations in fabrication, the dopants will form electrical junctions in the wafer back, which may interfere with good conduction in the back contact that is required for the circuit to operate correctly. These junctions may require physical removal by wafer thinning.

The thinning step generally takes place between wafer sort and die separation. Wafers are reduced to a thickness of 0.2 to 0.5 mm.[1] Thinning is done by the same processes (mechanical grinding and chemical-mechanical polishing-CMP) used to grind wafers in the wafer preparation stage (see Chapter 3). An alternate method is to protect the front of the wafers and chemically etch material from the back.

Wafer thinning is a worrisome process. In back grinding, there is the concern of scratching the front of the wafer and of wafer breakage. Since the wafer must be held down on the grinder or polishing surface, the front of the wafer must be protected and, once thinned, wafers are easier to break. In back etching, there is a similar need to protect the front of the wafer from the etchant. The protection can be provided by spinning a thick layer of photoresist on the front side. Other methods include attaching adhesive-backed polymer sheets cut to the wafer diameter. Stresses induced in the wafer by grinding/polishing processes must be controlled to prevent wafer and die warping. Wafer warping interferes with the die separation process (broken and cracked die). Die warping creates die attach problems in the packaging process.[2]

Backside gold. Another optional wafer process is adding a layer of backside gold. A layer of gold is required on wafers that are going to be

attached to the package by eutectic techniques (see the "Die attach" section). The gold is usually applied in the fabrication area (after back grinding) by evaporation or sputtering.

Die separation

The chip-packaging process starts with the separation of the wafer into individual dies. The two methods of die separation are scribing and sawing (Fig. 18.16).

Scribing. *Scribing,* or *diamond scribing,* was the first production die separation technique developed in the industry. It requires dragging a diamond-tipped scribe through the center of the scribe lines and separating the die by flexing the wafer. Scribing becomes unreliable in wafers over 10 mils thick.

Sawing. The advent of thicker wafers has led to the development of sawing as the preferred die separation method. A saw consists of a wafer table with rotation capability, a manual or automatic vision system for orienting the scribe lines, and a diamond-impregnated round saw. Two techniques are used. Both start with the passage of the diamond saw over the scribe lines. For thinner wafers, the saw is lowered into the wafer surface to create a trench about one-third of the way through the wafer. The separation of the wafer into die is completed by the stress and roller technique used in the scribing method. The second sawing method is to separate the die by a complete saw-through of the wafer.

Often, the wafers for complete saw-through are first mounted on a flexible plastic film. The film holds the die in place after the sawing operation and aids the die pick operation. Sawing is the preferred die separation method due to the cleaner die edges and the fewer cracks and chips left on the sides of the die (Fig. 18.17).

Die pick and place

After sawing, the separated die are transferred to a station for selection of the functioning die (non-inked). In the manual method, an op-

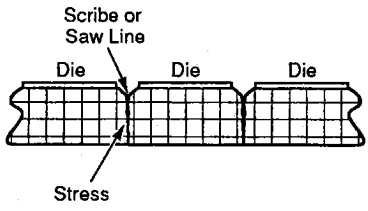


Figure 18.16 Scribe and saw separation.

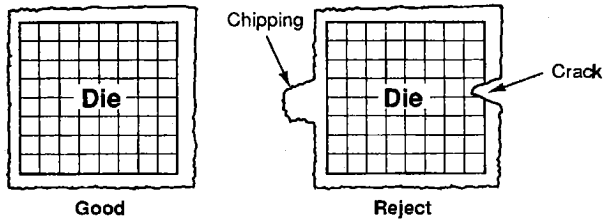


Figure 18.17 Die-separation results.

erator will pick up each of the non-inked dies with a vacuum wand and place it in a sectioned plate. Wafers that come to the station on the flexible film are first placed on a frame that stretches the film. The stretching separates the die, which aids the die pick operation.

In the automated version of this operation, a memory tape or disk that has the locations of the good die (from wafer sort) is loaded into the tool. A vacuum wand picks up good die and automatically places them in a sectioned plate for transfer to the next operation.

Die inspection

Before being committed to the rest of the process, the die are given an optical inspection. Of primary interest is the quality of the die edge, which should be free of chips and cracks. This inspection also sorts out surface irregularities, such as scratches and contamination. Inspection may be manual with microscopes or automated with a vision system. At this step, the die is ready to go into a package.

Die attach

Die attachment has several goals: to create a strong physical bond between the chip and the package, to provide either an electrical conducting or insulating contact between the die and the package, and to serve as a medium to transfer heat from the chip to the package.

A requirement is the permanency of the die-attachment bond. The bond should not loosen or deteriorate during subsequent processing steps or when the package is in use in an electronic product. This requirement is especially important for packages that will be subjected to high physical forces, such as in rockets. Additionally, the die attach materials should be contaminant-free and should not outgas during subsequent process heating steps. Lastly, the process itself should be productive and economical.

Eutectic die attach. There are two principal methods of die attach: *eutectic* and *epoxy* adhesives. The eutectic method is named for the phe-

nomenon that takes place when two materials melt together (alloy) at a much lower temperature than either of them separately. For die attach, the two eutectic materials are gold and silicon (Fig. 18.18). Gold melts at 1063°C, while silicon melts at 1415°C. When the two are mixed together, they start alloying at about 380°C. Gold is plated onto the die-attach area and alloys with the bottom of the silicon die when heated.

The gold for the die-attach layer is actually a sandwich. The bottom of the die-attach area is deposited or plated with a layer of gold. Sometimes, a preformed piece of metal composed of a gold and silicon mixture is placed in the die-attach area. When heated, these two layers, along with a thin layer of silicon from the wafer back, form a thin alloy layer. This layer is the actual bond forming the die-package attachment.

Eutectic die attach requires four actions. First is the heating of the package until the gold-silicon forms a liquid. Second is the placement of the chip on the die-attach area. Third is an abrasive action, called *scrubbing*, that forces the die and package surfaces together. It is this action, in the presence of the heat, that forms the gold-silicon eutectic layer. The fourth and final action is the cooling of the system, which completes the physical and electrical attachment of the chip and package.

Eutectic die attach can be performed manually or by an automated machine that performs the four actions. Gold-silicon eutectic die attach is favored for high-reliability devices and circuits for its strong bonds, heat dissipation properties, thermal stability, and lack of impurities.

Epoxy die attach. The alternate die-attach process uses thick liquid epoxy adhesives. These adhesives can form an insulating barrier between the chip and package or become electrically and heat conductive with the addition of metals such as gold or silver. Polyimide may also be used as an adhesive. Popular adhesives are silver-filled epoxy for copper lead frames and silver-filled polyimide for Alloy 22 metal frames.[3]

Conducting
- Gold/Silicon Eutectic
- Metal-Filled Epoxy
- Conducting Polyimide

Nonconducting
- Epoxy Adhesive
- Insulating Polyimide

Figure 18.18 Die-attach material matrix.

The epoxy process starts with the deposit of the epoxy adhesive in the die-attach area by dispensing the adhesive with a needle or screen printing it into the die-attach area. The die, held by a vacuum wand, is positioned in the center of the die-attach area. The second action is to force the die into the epoxy to form a thin uniform layer under the die. The final action is a curing step in an oven at an elevated temperature that sets the epoxy bond.

Epoxy die attach is favored for its economy and ease of processing, in that the package does not have to be heated on a stage. This factor makes the automation of the process easier. When compared to gold-silicon eutectic die attach, epoxy has the disadvantage of potential decomposition at the high temperatures of bonding and sealing operations. Epoxy die-attach films also do not have the bonding power of the eutectics.

Regardless of the attachment method used, there are several marks of a successful die attach. One is the proper and consistent alignment of the die in the die-attach area. Proper placement pays off in faster and higher-yield automatic bonding. Another desired result is a solid, uniform, and void-free contact over the entire area of the chip. This is necessary for mechanical strength and good thermal conduction. One evidence of a uniform bond is a continuous joint or "fillet" between the die edge and the package. The final mark of a good die-attach process is a die-attach area free of flakes or lumps that can come loose during use and cause a malfunction.

Die-to-package bonding

Once the die and package are attached, they go to the bonding process. This is perhaps the most critical of all the assembly operations. Three techniques provide the critical chip/package connection: wire bonding, bump/flip-chip, and TAB. In wire bonding, up to hundreds of wires must be perfectly bonded from the bonding pads to the package inner leads. In bump/flip chip, the bonding pad/package connection is a solder ball. Tape automated bonding (TAB) system is a process that bonds the lead frame leads directly to the die bonding pads in one step.

Wire bonding

The wire bonding procedure is simple in concept. A thin (0.7 to 1.0 mil) wire is first bonded to the chip bonding pad and spanned to the inner lead of the package lead frame. The third action is to bond the wire to the inner lead. Last, the wire is clipped and the entire process repeated at the next bonding pad. While simple in concept and proce-

dure, wire bonding is critical because of the precise wire placement and electrical contact requirements. In addition to accurate placement, each and every wire must make good electrical contact at both ends, span between the pad and inner lead in a prescribed loop without kinks, and be at a safe distance from neighboring wires. Wire loops in conventional packages are 8 to 12 mils, while those in ultra-thin packages are 4 to 5 mils.[4] Distances between adjacent wires are referred to as the *pitch* of the bonding.

Wire bonding is done with either gold or aluminum wires. Both are highly conductive and ductile enough to withstand deformation during the bonding steps and still remain strong and reliable. Each has its advantages and disadvantages, and each is bonded by different methods.

Gold wire bonding. Gold has several pluses as a bonding wire material. It is the best known room-temperature conductor and is an excellent heat conductor. It is resistant to oxidation and corrosion, which translates into an ability to be melted to form a strong bond with the aluminum bonding pads without oxidizing during the process. Two methods are used for gold bonding. They are *thermocompression* and *thermosonic.*

Thermocompression bonding (also known as *TC bonding*) starts with the positioning of the package on the bonding chuck and the heating of the chip and package to between 300 and 350°C. Chips that are going to be enclosed in an epoxy molded package are processed through die attach and bonded with the chip on the lead frame only. The bonding wire is fed out of a thin tube called a *capillary* (Fig. 18.19). An instantaneous electrical spark or small hydrogen flame melts the tip of the wire into a ball and positions the wire over the first bonding pad. The capillary moves downward, forcing the melted ball onto the center of the bonding pad. The effect of the heat (thermal) and the downward pressure (compression) forms a strong alloy bond between the two materials. This type of bonding is often called *ball bonding.* After the ball bond is complete, the capillary feeds

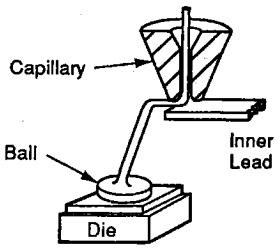


Figure 18.19 Gold ball bonding.

out more wire as it travels to the inner lead. At the inner lead, the capillary again travels downward to where the gold wire is forced by the heat and pressure to melt onto the gold-plated inner lead. The spark or flame then severs the wire, forming the ball for the next pad bond. This procedure is repeated until every pad and its corresponding inner pad are connected.

Thermosonic, gold ball bonding follows the same steps as thermocompression bonding. However, it can take place at a lower temperature. This benefit is provided by a pulse of ultrasonic energy that is sent through the capillary into the wire. This additional energy is sufficient to provide the heat and friction to form a strong alloy bond.

The majority of production gold wire bonding is done on automatic machines that use sophisticated techniques to locate the pads and span the wire to the correct inner lead. The fastest bonding machines can perform thousands of bonds per hour. There are two major drawbacks to the use of gold bonding wires. First is the expense of the gold. Second is an undesirable alloy that can form between the gold and aluminum. This alloy can severely reduce the conduction ability of the bond. It forms a purplish color and is known as the "purple plague."

Aluminum wire bonding. Aluminum wire, while not having the conduction and corrosion-resistance properties of gold, is still an important bonding wire material. A primary advantage is its lower cost. The second advantage is that the bond with the aluminum bonding pad is a monometal system and thus less susceptible to corrosion. Also, aluminum bonding can take place at lower temperatures than gold bonding, making it more compatible with the use of epoxy die-attach adhesives.

The bonding of the aluminum follows the same major steps as gold wire bonding. However, the method of forming the bond is different. No ball is formed. Instead, after the aluminum wire is positioned over the bonding pad, a wedge (Fig. 18.20) forces the wire onto the pad as a pulse of ultrasonic energy is sent down the wedge to form the bond. Af-

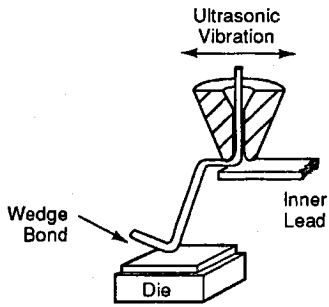


Figure 18.20 Aluminum wedge bonding.

ter the bond is formed, the wire is spanned to the inner lead where another ultrasonic-assisted wedge bond is formed. This type of bonding is known variously as *ultrasonic* or *wedge bonding.*

After this bond, the wire is cut. At this point in the process, a major difference between the bonding of the two materials occurs. In gold bonding, the capillary moves freely from pad to inner lead, to pad, and so forth, with the package in a fixed position. In aluminum wire bonding, the package must be repositioned for every single bonding step. The repositioning is necessary to line up the pad and inner lead along the direction of travel of the wedge and wire. This requirement places an additional difficulty on the designers of automatic aluminum bonding machines. Nevertheless, most production aluminum bonding is done on high-speed machines.

Bump/flip-chip bonding

Wire bonding presents several problems. There are electrical resistances associated with each bond. There are minimum height limits imposed by the required wire loops. There is the chance of electrical performance problems or shorting if the wires come to close to each other. Plus, the wires require an individual bonding step at both the chip bonding pad and at the package lead. Perhaps the biggest problem results from the increasing number of connections (pin count) needed to operate larger circuits. Chip designers simply run out of space to locate the required number of connections around the periphery of the chip. These issues are addressed by replacing wires with a deposited metal *bump* on each bonding pad. The bumps are also called *balls*, as in naming packages using bump/flip-chip processes as ball grid arrays (BGAs). This bonding method allows chip design with bonding pads located both along the edge of the die as well as in the interior of the die (Fig. 18.21). These locations place the bump closer to the chip circuitry, increasing signal processing speed. Connection to the package is made when the chip is flipped over and the bump soldered to a corresponding package inner lead (Fig. 18.22) on a package or printed circuit board. IBM calls their version of this technology controlled collapse chip connection (C4).[5]

This process leaves the die suspended above the package surface. Physical stresses and strains are absorbed by the soft solder bump. Additional stress tolerance is provided by filling the gap with an epoxy filling, called and underfill.

Bump connection technology starts in the wafer fabrication process. Wafers are processed through the usual metallization, passivation, and bonding pad patterning processes. The last patterning process leaves an opening in the passivation layer over the bonding pads.

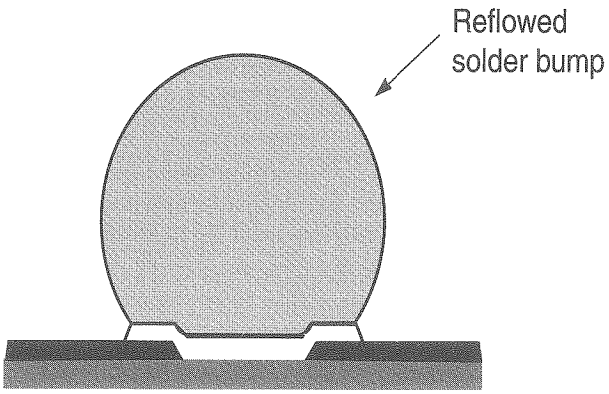


Figure 18.21 Reflowed solder bump. [5]

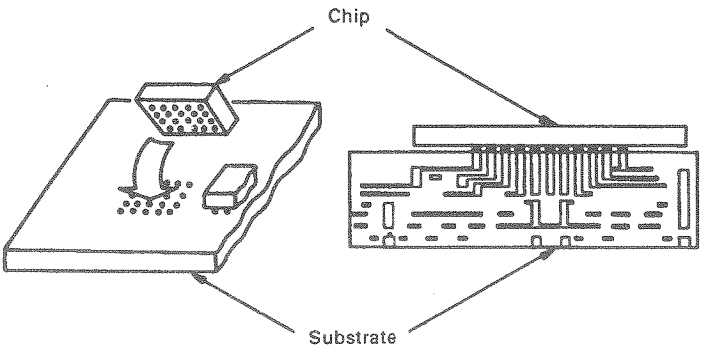


Figure 18.22 Flip-chip joining. *(Source:* Microelectronics Handbook, *by Tummala and Rymaszewski.)*

A number of process flows are available to form the solder bumps on the bonding pad. The process described below is an example.

Sputter deposit intermetal stack. Lead/tin solder balls are the preferred "bump" material. However, an intermetal layer (or stack) is required between the bonding pad and the solder ball to prevent the lead from diffusing into the aluminum pad and to aid adhesion of the solder ball onto the pad. Various metal stacks are used, including chrome-copper-gold (Cr-Cu-Au), titanium-nickel (Ti-Ni), and plain copper (see "Alternative process").

Patterning step of bump location. A patterning step covers the die surface with resist, leaving openings over the bonding pads and sur-

rounding dielectric. The resist layer is thick enough to accommodate enough solder to form a ball of sufficient volume to provide structural support and lower electrical resistance between the chip and package or substrate.

Deposit of intermediate layer stack. The intermetal materials are evaporated or sputtered through the openings on the pad.

Deposit lead/tin solder. Deposit of the lead/tin solder can be by electroplating or evaporation. If electroplating is used, a seed layer is deposited before the electroplating. Lead/tin is used to lower the melting point of the solder.

Remove resist. The resist is removed, leaving a bump of lead/tin connected to the bonding pad.

Alternative Process

Figure 18.23 shows an alternative process. It starts with a blanket sputter deposition of a Ti/Ni or copper. After a patterning step over the bonding pad, there is another deposit of Ni, followed plating of the solder plug. The system is completed to with removal of the resist and etching away the intermetal materials left on the surface. The end result is similar to the first process—a ball of lead/tin solder positioned on the bonding pad.

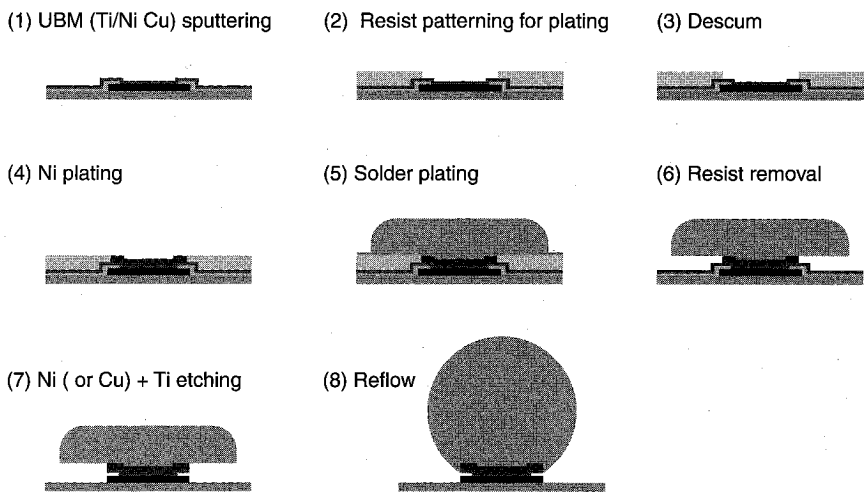


Figure 18.23 Alternative bump process. *(Courtesy of Future Fab International.)*

Transfer to Packaging Area

Once these wafer fabrication processes are complete, the wafer is transferred to the packaging operation.

Reflow

A heating of the wafer in the 350°C range in hydrogen enhances the electrical contact to the intermetal stack. Also, surface tension pulls the solder deposit into a spherical shape, much like surface tension forms spherical soap bubbles.

Die separation and die pick and place

These two steps are the same as in the wire bonding process. Typically, the die are affixed to a flexible tape on a reel that allows automatic handling in the subsequent processes.

Alignment of die to package

Packages used in flip-chip systems are usually ceramic or organic. *Organic* is the term used for packages based on the same materials used for circuit boards. Sometimes they are referred to as "plastic" packages. The first step is flipping the die upside down and aligning it to the package such that the bumps on the die are positioned over the corresponding leads on the package.

Attachment to package (or substrate)

In the presence of a solder flux, the die/package combination is heated in an oven to melt the solder ball to the package lead.

Deflux

A cleaning process removes excess flux from the surface.

Underfillment

An epoxy is introduced under the corners of the die soldered to the package. When heated, the epoxy is drawn by capillary action underneath the entire chip. After a curing step it provides a "cushion" to absorb additional stresses on the system. The package designs using the bump/flip-chip technique are discussed later on in this chapter.

Encapsulation

Typically, a molding compound is deposited over the upside-down chip to provide environmental and physical protection. The combination of the substrate, die, and protective top layer form the functions of a tradition package.

Tape automated bonding. Tape automated bonding (TAB) is used to wire up chips when extreme package thinness is required, such as in credit-card-size radios. The TAB process starts with formation of the lead system on a flexible strip of tape. Various methods are used to form the lead system. The metal for the system is deposited on the tape by sputtering or evaporation. Formation of the lead system is either by mechanical stamping or patterning techniques similar to the fabrication patterning process. The result is a continuous tape containing many individual lead systems. For the bonding operation (Fig. 18.24), the chip is positioned on a chuck, and the tape is moved by sprockets until one of these lead systems is positioned exactly over the chip. In this position, the inner leads of the system should be positioned over the bonding pads of the chip.

The bond is completed with a tool known as a *thermode.* The thermode is faced with a flat diamond surface and is heated. The thermode is moved downward, first contacting the inner leads. It continues downward with enough pressure to force the inner leads onto the bonding pads. The heat and the pressure are regulated to cause a physical and electrical bond between the two. Large chips require a larger TAB bonding area. For these chips, the thermode is faced with a synthetic diamond.

TAB bonding is also used in conjunction with package bonding. The advantages of TAB are speed, in that all of the bonds to the chip are made in one action, and the ease of automation offered by the tape and sprocket system.

Preseal inspection

After bonding, a number of steps are required to complete the packaging operation. The packaging steps described follow the traditional wire bonding and individual package process flow. Most of them have to be performed at some point in the bump/flip-chip and tape auto-

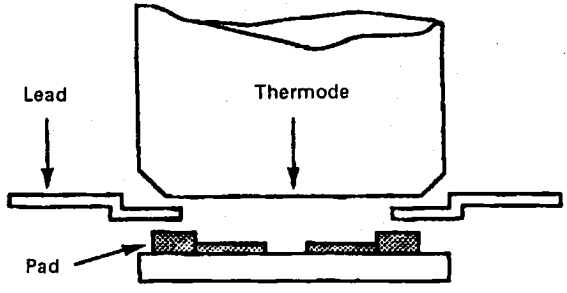


Figure 18.24 Tape automated bonding. *(Source:* Microelectronics Handbook, *by Tummala and Rymaszewski.)*

mated bonding processes. An important step in the traditional chip-packaging process is the preseal or precap inspection, sometimes called *third optical inspection,* which takes place after the bonding step. The inspection is performed to provide feedback on the quality of the operations already performed. It also is performed to reject packaged chips that may represent reliability hazards when the chip is in operation in the field.

While there are many levels of inspection criteria, they fall into two main categories: commercial and high-reliability. The commercial inspections are given to chips and packages destined for use in commercial systems and are derived from the internal quality levels of the producing company in conjunction with its experience and customer specifications. The high-reliability specifications are derived from a set of government standards identified as "Mil-Standard 883." Commercial-level inspections screen the bonded chips for die-attach quality, correct placement of the bonds on the bonding pads and inner leads, the shape and quality of the ball or wedge bond, and the general condition of the chip surface in regard to contamination, scratches, etc. Mil-Standard 883 covers the same general issues as the commercial inspections but to more stringent requirements. In particular, this standard also specifies criteria for the chip surface, including pattern alignment, critical dimensions, and surface irregularities, such as small scratches, voids, and small defects. These criteria serve to reject bonded chips that may malfunction in the rigorous conditions encountered in space and military operation.

Sealing techniques

After the bonded chip passes the optical inspection, it is ready for sealing in a protective enclosure. Several methods are used to achieve the enclosure of the chip. The method chosen depends on whether the seal must be hermetic or nonhermetic and which type of package is to be used. The principal sealing methods use welded seals, soldered seals, glass-sealed lids, CERDIP package construction, molded epoxy enclosures, and a sealing layer (see "Blob top") directly on a die bonded directly to a flip-chip package or printed circuit board (Fig. 18.25).

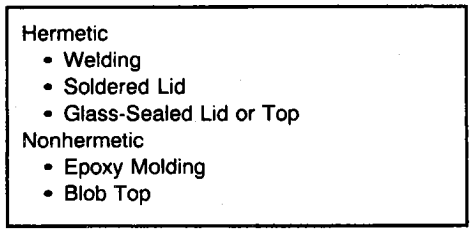


Figure 18.25 Package sealing methods.

Metal can. If the package is a metal can type, a hermetic seal is achieved by welding the flanged lid to a matching flange on the base of the package.

Premade packages. Premade ceramic packages are sealed by one of two methods, metal or ceramic lids (Fig. 18.26). Packages made for metal lids have a ring of gold around the top of the die-attach cavity, called the *seal ring.* Placed on top of the seal ring area is a performed piece of gold-tin solder. The metal lid is clamped in position over the seal ring and placed on the belt of a conveyor furnace. As the clamped package passes through the furnace, the lid and package are soldered together to form a hermetic seal. The sealing takes place at a temperature range of 320 to 360°C in a pure nitrogen atmosphere. If the package is to receive a ceramic lid, the procedure is similar. The part of the ceramic lid that contacts the base, outside the die attach area, is coated with a layer of low-melting-point glass. The hermetic seal is completed as the package passes through a conveyor furnace. This sealing takes place at a temperature of about 400°C in an atmosphere of clean, dry air.

CERDIP packages. A completed CERDIP package results in a hermetic seal around the chip and bonding wires. This seal is accomplished with glass, similar to the ceramic seal on the premade packages.

In the case of the CERDIP package, the inner metal lead system is buried in a glass layer. The ceramic top of the package system has a cavity (Fig. 18.27). The underside of the lid, outside the cavity area, is coated with a layer of low-melting-point glass. The lid is placed over the base and clamped. The seal is accomplished as the assembly is passed through a conveyor furnace or placed in an oven. In the furnace or oven, the glass melts, fusing the base and top together. The CERDIP glass sealing system is used to seal DIP and flat packs. These latter packages are known as *Cerpacks* and *Cerflats.*

Molded epoxy enclosures. The fourth major method of enclosure, *epoxy molding,* produces the plastic package (Fig. 18.28). The resultant

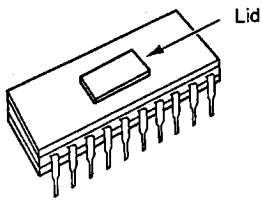


Figure 18.26 Premade ceramic package.

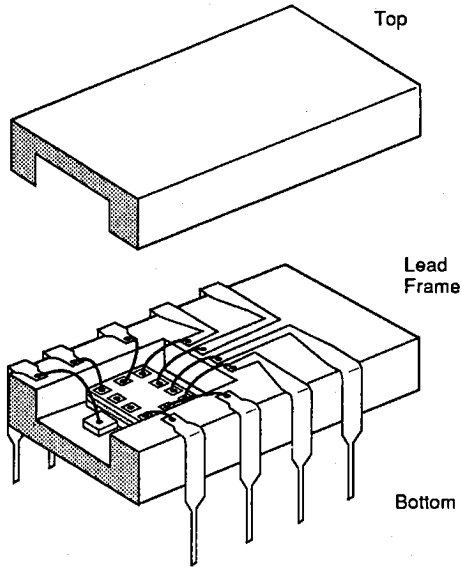


Figure 18.27 CERDIP parts.

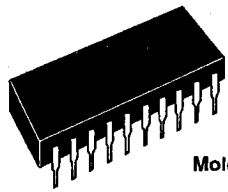


Figure 18.28 Molded CERDIP.

seal, while protecting the die from moisture and contamination, is not classified as hermetic. However, there exists a considerable amount of research into the development of improved epoxy materials to create better enclosures. The major advantages of epoxy molded enclosures are weight, low material cost, and manufacturing efficiency.

This sealing method follows a different process flow. The die is attached and bonded to a lead frame containing a number of lead systems (Fig. 18.29). After the preseal inspection, the lead frames are transferred to the molding area. The frames are placed on a mold mounted in a transfer molding machine. The molding machine is in turn charged with pellets of the epoxy material, which have been previously softened by a radio frequency heater. Inside, the pellets are forced by a ram into a liquid state. The ram then forces the liquid around the die on the lead frames, forming an individual package around each lead frame. After the epoxy sets in the mold, the frames are removed and placed in an oven for final curing.



Figure 18.29 Lead frame.

Lead plating

An important feature of the completed package is the finish on the package leads. Most package leads are coated with lead-tin solder, tin plate, or gold plate. The plating serves several important functions.

First is the solderability of the leads into a circuit board. The additional metal finish improves the lead solderability, resulting in a more reliable electrical connection of the package and the printed circuit.

The second benefit of the lead finish is that it protects the leads from oxidation or corrosion during periods of storage prior to mounting on the circuit board. The third benefit of lead plating is the protection of the leads from corrosive agents in the packaging and printed-circuit board mounting processes. These agents include solder flux, corrosive cleaners, and even tap water. The plating continues to protect the leads during their lifetime of use.

Electrolytic plating. Plated layers, such as gold and tin, are applied by electrolytic procedures. The packages are mounted on racks with each lead connected to an electric potential. The racks are placed in a tank containing a plating solution. Next, a small current is passed between the packages and an electrode in the tank. The current causes the particular metal in the solution to plate onto the leads.

Tin-lead solder. Tin-lead solder layers are applied either by dipping the packages into a pot of the molten solder or by a wave soldering technique. This latter technique offers the advantage of good control of

the layer thickness and provides a shorter exposure of the package to the molten solder.

Plating process flows

Metal cans, side-brazed DIP packs, and pin grid arrays have their leads plated before starting into the packaging process. CERDIP and plastic packages go through the plating process after the sealing steps.

Lead trimming

One of the last steps in the package assembly process is trimming away excess material from the leads. The outer leads of DIP and flat-pack packages are made with a tie-bar (Fig. 18.30). This bar keeps the leads from becoming bent during the packaging process. At the end of the process, the package goes through a simple trimming machine that simultaneously trims away the tie-bar and trims the leads to the proper length.

Plastic package lead frames have an additional piece of material. It is a bridge of metal close to the package body that functions as a dam to prevent the liquid epoxy material from running into the lead area (Fig. 18.31). The dam is cut away from the frame with a series of precise cutting tools. After the dam is cut away, the packages move to another station on the cutter where the frame is separated into individual packages. If the package is a surface-mount type, the leads will be bent into the required shape.

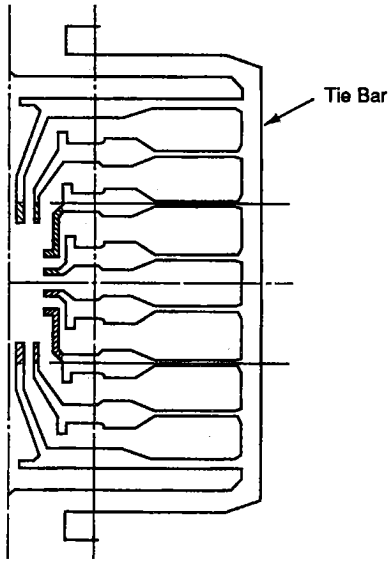


Figure 18.30 Tie-bar.

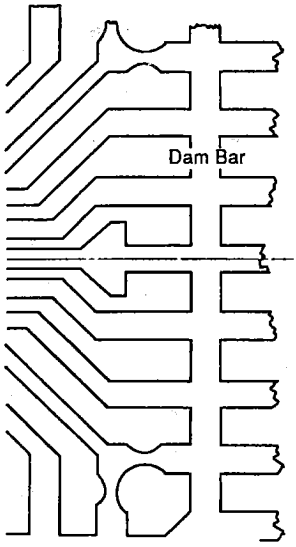


Figure 18.31 Lead frame dam.

Deflashing

Plastic packages receive an additional process, called *deflashing,* which is required to remove excess molding material from the package enclosure. Deflashing is done by either dipping the packages in a chemical bath followed by a rinse or by a physical abrasion process. Physical deflashing is done in a machine similar to a sandblaster that uses plastic beads as the abrasive.

Package marking

Once completed, a package must be identified with key information. Typical information coded on the package is the product type, device specifications, date, lot number, and where it was made. The main methods of marking are ink printing and laser inscription. Ink printing has the advantage of good adherence to all of the package materials.

The composition of the ink is chosen for permanence in the eventual operating environment of the device. The ink is applied by an offset printer followed by a curing step. Curing is done by oven drying, room-temperature air drying, or by ultraviolet light. Laser printing is especially well suited to plastic packages. The mark is permanently inscribed in the package surface and can provide good contrast on the dark packages. Additionally, laser marking is fast and noncontaminating, since no foreign material is added to the package surface, and no curing step is required. A drawback to laser marking is the difficulty

of changing the mark if a wrong code was used or the status of the device is changed. Regardless of the marking method, all marks must meet the requirements of legibility (especially on smaller packages) and permanence when exposed to harsh environments.

Final testing

At the conclusion of the packaging process, the completed package is put through a series of environmental, electrical, and reliability tests. These tests vary in type and specifications, depending on the customer and use of the packaged devices. The tests may be performed on all of the packages in a lot or on selected samples.

Environmental tests

Environmental tests are performed to weed out leaking and defective packages. Defects detected are loose chips, contaminants and particles in the die-attach cavity, and faulty bonding. This testing series starts with a stabilization bake to drive off any volatile substances in the package. A typical bake is at 150°C for 24 hours.

The first environmental test is *temperature cycling.* The packages are loaded into a chamber and cycled between two temperature extremes. The number of cycles may reach several hundred. The high and low temperatures of this test vary with the device use. Commercial parts receive a narrower temperature range than hi-rel parts. The hi-rel cycle range is –25 to 125°C. During the cycling, any weakness in the seal, die attachment, or bonding will be aggravated and detected in later electrical tests.

A second environmental test is *constant acceleration.* In this test, the packages are accelerated in a centrifuge (Fig. 18.32) that creates a force as high as 30,000 times the pull of gravity on the Earth (30,000 *g*'s). During the acceleration, loose particles in the package, poorly attached dies, and weakly attached bonds are stressed so that they will be detected at the final electrical tests.

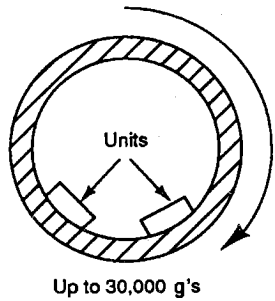


Figure 18.32 Acceleration test.

Leaks in the package enclosure are detected by two tests. *Gross leak* testing (Fig. 18.33) is conducted by submerging the packages in a hot liquid. The heated liquid raises the temperature of the package and forces trapped gases in the cavity to escape. The escaping gases are observable as bubbles rising in the liquid. The chamber has a transparent side, allowing an operator to observe the bubbles. Smaller (or fine) leaks are detected by using tracer gases. For this check, helium is pumped under pressure into a chamber containing the packages. If the package has small leaks, the gas will be pumped into the package cavity. The gas is detected as it escapes through the small leaks by a machine known as a mass spectrometer, which can identify the escaping gas. An alternate fine leak test uses the radioactive gas krypton-85. It too is pumped through any leaks into the package under pressure. Detection of any krypton-85 in the package is by a device similar to a Geiger counter.

Electrical testing

The purpose of the wafer fabrication and packaging processes is to provide to the customer a specific semiconductor device that performs to specific parameters. Thus, one of the last steps is an electrical test of the completed unit to verify that it performs to specifications. The tests are similar to the wafer sort tests. The overall objective is to verify that the good chips identified at wafer sort have not been compromised by the packaging process.

First, there is a series of parametric tests. These electrical tests check the general performance of the device or circuit and ensure that it meets certain input and output voltage, capacitance, and current specifications. The second part of the final test is called the *functional test.* This test actually exercises the specific chip functioning. Logic chips are put through logic tests, and memory chips are exercised in their data storage and retrieval capabilities. The equipment used to conduct the final test is electrically similar to that used in the wafer-sort operation. The electrical tests are performed by a computer-con-

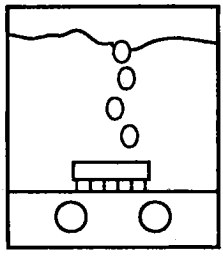


Figure 18.33 Gross leak bubbles.

trolled tester that directs the sequences and levels of the parametric and functional tests. The packages are connected to the tester through sockets; the socket unit is known as the *test head.* The packages are inserted into the test head manually or by an automatic unit known as a *handler* (Fig. 18.34). This handler may be mechanical or robotic, depending on the speed and complexity of the operation.

Burn-in tests

The last of the tests is the optional burn-in test(s). The reason it is optional is that, although it is required for all high-reliability device lots, it may or may not be performed on commercial devices. The test requires the insertion of the packages in sockets and mounting in a chamber with temperature-cycling capability. During the test, the circuits are temperature cycled while under an electrical bias.

The burn-in test is intended to stress the electrical interconnection of the chip and package and drive any contaminants in the body of the chip into the active circuitry, thus causing failure. This test is based on data that indicate that chips prone to these types of failures actually malfunction in the early part of their lifetime. By conducting burn-in tests, the early failures are detected. The devices passing the test are statistically more reliable.

Package Process Flows

It should be obvious that there is no universal packaging process flow. The package construction technique and lead-plating requirements dictate the flow. The table in Fig. 18.35 illustrates typical flows for three of the more common package types. As reported bump/flip-chip bonding techniques follow a completely different process flow.

Package/Bare Die Strategies

Integrating a chip into an electronic product or system requires an electrical subsystem between the chip and the system. Two main ap-

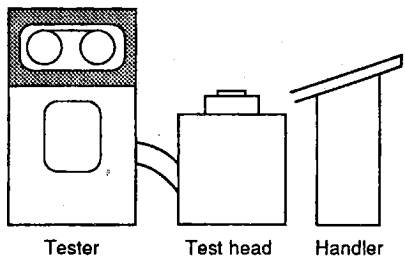


Figure 18.34 Final test.

Process Step	Premade Ceramic	CERDIP	Epoxy Molded
Die Separation	X	X	X
Pick and Plate	X	X	X
Die Attach	X	X	X
Wire Bonding	X	X	X
Preseal Inspect	X	X	X
Lid Sealing	X	X	
Epoxy Molding	X		
Deflashing	X		
Lead Plating and Trim		X	X
Marking	X	X	X
Environmental Test	X	X	X
Final Electric Test	X	X	X
Burn-In	X	X	X

Figure 18.35 Table of packaging process flows.

proaches have emerged: individual packaging and bare die techniques (Fig. 18.36). There are two principal bare die techniques: wire bonding and bump/flip-chip. The wire bonding technique starts with adhesion of the die, face up, on the printed circuit board (PCB) or substrate. Each bonding pad is wire bonded to a corresponding electrical lead on the board. The bump/flip chip technique uses the same technology described to connect die directly to a printed circuit board surface.

Package/PCB connections

On the package side, there are four primary techniques used to connect the package to a printed circuit board (PCB). They are through-hole, surface mount, TAB, and ball (bump) technology. Through-hole connections feature straight pins on the package, which are inserted into corresponding holes in the printed circuit board (Fig. 18.37). A newer method is *surface mount,* also referred to as *SMD.* This method

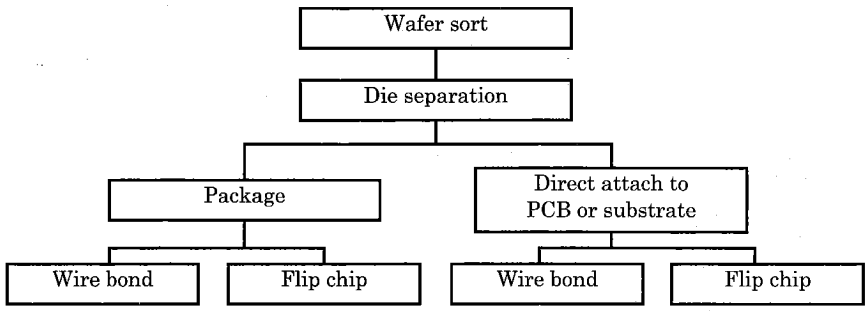


Figure 18.36 Post-die-separation options.

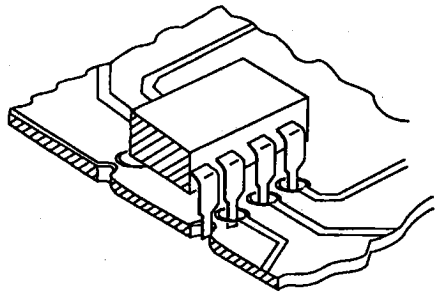


Figure 18.37 DIP through-hole assembly.

features packages that have their leads bent into a J shape or bent outward to allow direct soldering to the board surface (Fig. 18.38). Some SMD packages do not have leads; rather, they terminate in metal traces hugging the package body. They are known as *leadless packs*. For inclusion on a circuit board, they are inserted into chip carriers, which in turn have the leads that attach to the printed circuit board. Last, there is the bump technology used to connect die to packages, adapted to connecting packages to a PCB.

Tape automated bonding (TAB) has two uses. One is bonding the chip bonding pads directly to the lead frame (see section on bonding). TAB is also a technique for bonding the outer leads directly to the PCB.

Package Design

Up to the early 1970s, most chips ended up in either a metal package, known as a "can," or in the familiar dual in-line package (DIP). The trends in chip size and integration levels and new electronic products with special packaging requirements (smart cards) have driven the development of new packages and strategies.

Metal cans

Metal cans are cylindrically shaped packages with an array of leads extending through the base (see Fig. 18.39). The chip is attached to

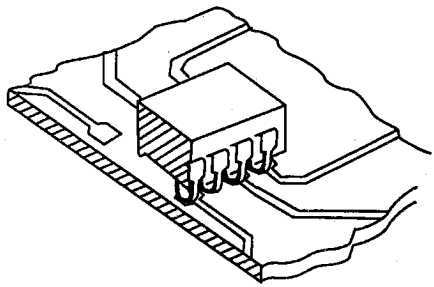


Figure 18.38 Surface-mount device.

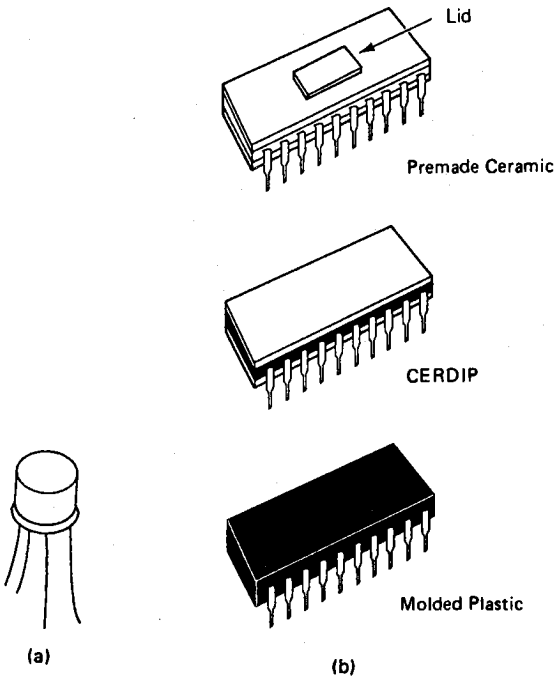


Figure 18.39 (*a*) Metal can and (*b*) DIP packages.

the base and wire bonded to posts that are connected to the leads. The lid and base have matching flanges that are welded together to create a hermetic seal. These packages are designated by numbers, with the T0-3 and T0-5 being the most common. Metal cans are used to package discrete devices and small-scale integrated circuits.

The DIP is probably the most familiar package design. It features a thick sturdy body with two rows of outer leads coming out of the side and bending downward. DIPs are constructed by three different techniques (Fig. 18.39). Chips designed for high-reliability use will be packaged in a premade ceramic DIP. The package is formed with a solid body of ceramic with the leads buried in the ceramic. The die attachment area is a cavity recessed into the body. The hermetic seal is completed by a soldered metal lid or a glass-sealed ceramic lid.

Another approach to the DIP is the CERDIP, which stands for ceramic DIP. This type of package is composed of a bottom ceramic base with the lead frame held firm in a glass layer. The chip is attached to the base and wire bonded to the lead frame. The hermetic seal is completed with a ceramic top glass sealed to the base. CERDIP construction is used for a number of package types. The vast majority of DIPs

are made by the epoxy molding technique. In this technique, the chip is attached to a lead frame and then wire bonded. After bonding, the frame is placed in a molding machine and the package is formed around the chip, wire bonds, and inner leads.

Pin grid arrays

Larger chips, with more leads, have outgrown the DIP configuration. The pin grid array is a package designed for larger chips. It features a premade "sandwich" with the outer leads coming out of the bottom of the package in the form of pins (Fig. 18.40). The chip is attached in a cavity that is formed in either the top or bottom of the body, usually using bump/flip-chip technology. Connections to the chip cover the entire chip area, unlike most chips with connections restricted to bonding pads around the periphery of the chip. Ceramic PGAs are hermetically sealed with a soldered metal lid.

Ball grid arrays (BGAs) or flip-chip ball grid arrays (FCBGAs)

BGAs have the same body shape as PGAs. Instead of pins on the package bottom, there is a series of solder bumps (balls) used to connect the package to the PCB (Fig. 18.40). This is essentially the same technology used to connect die to packages.

The effect is to lower the package profile and weight as well as providing higher pin counts by using the whole chip surface for bonding pads. Balls (or bumps) also bring the aspect absorbing stresses created from the thermal expansion differences between the package and the PCB.

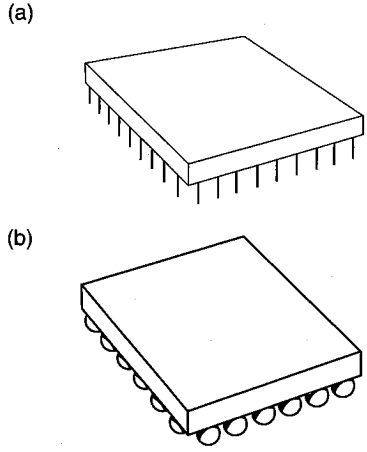


Figure 18.40 (*a*) Pin grid array. and (*b*) ball grid array.

Quad packages

While pin grid array packs are a convenient design for larger chips, their ceramic construction is expensive compared to molded epoxy packages. This consideration led to the development of the "quad" package. A quad (short for quadrant) pack is constructed by the epoxy molded technique but has leads coming out of all four sides of the package (Fig. 18.41).

Thin packages

New products such as smart cards require thin packages. Several techniques are used to make thinner packages. They are called flat packs (FPs), thin small outline packages (TSOPs), small outline IC (SOIC), or ultra-thin packages (UTPs). Flat packs are constructed by the same techniques used to form DIP packages. These packages are designed with flatter height profiles and have their leads bent out to the side of the package (Fig. 18.42). Ultra-thin packages have total body heights in the 1 mm range. There are also quad flat packs (QFPs).

Chip-scale packages

In the world of ICs, the perfect package is no package. It is recognized that any package brings with it electrical resistance, weight, the opportunity to degrade the circuit performance, and cost. Overall, the smaller the package, the cheaper the cost of packaging and the benefit of higher densities. Chip-scale packages meet this need. They are simply packages with dimensions within 1.2 times the die size.[6] The

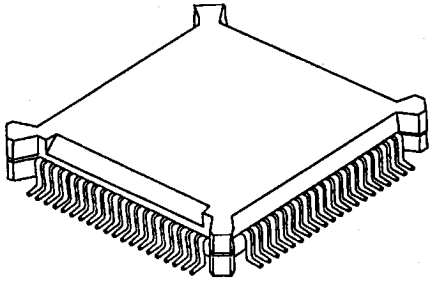


Figure 18.41 Quad package.

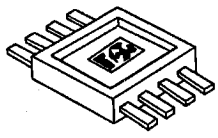


Figure 18.42 Flat pack.

challenges are to provide adequate mechanical and environmental protection for easy connection to printed circuit boards.

General design approaches favor flip-chip technology with ball grid arrays and blob top protection. The march to smaller packages and more reliable electrical connections has lead to the micro-ball grid array, or μBGA.

Lead on chip (LOC)

The LOC package, intended for large die, has the bonding pads arranged down the middle of the chip (Fig. 18.43). Package leads sit on a cushion over the chip surface.

Bare die techniques and blob top

Increased reliability with higher density and faster circuits presents an ongoing goal and challenge. Reliability is addressed in hybrid circuits. Speed and higher density comes with the elimination of the individual die package. Fewer links in the chain reduce resistances and (in some cases) shorten the length carriers have to travel, increasing speed. This strategy, called *bare die strategy,* is used in hybrid, multichip modules, and chip on board technology.

The most direct use of bare die is to bond them directly to the printed circuit board. Bonding techniques include all the ones used to bond chips into packages. Protecting the chip after connection to the board is by *blob top protection.* The protection is accomplished by covering the chip and bonds with a blob of epoxy resin material (Fig. 18.44).

The material is similar in properties to that used to mold the plastic packages. Blob top coverings are used with TAB and other packaging schemes.

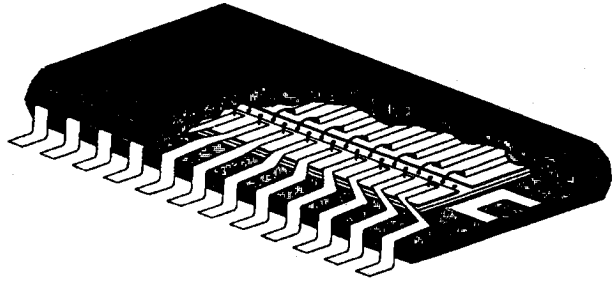


Figure 18.43 Lead on chip (LOC). *(Source:* Semiconductor International, *May 1993.)*

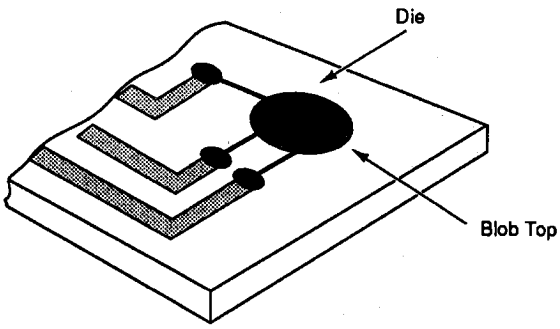


Figure 18.44 Blob top.

Hybrid circuits

Hybrid circuitry is an old technology, long favored for use in military and harsh environments. Hybrid circuits consist of a substrate on which standard electrical and semiconductor devices are mounted. The devices are connected by conductive or resistive thick-film paths on the surface of an insulating substrate. These paths are formed by silk screening inks containing a proper filler or from thin films evaporated or sputtered onto the surface and patterned by photolithography techniques. The term *hybrid* refers to the mix of solid state and conventional passive (resistors, capacitors) electrical devices present in the same circuit. Most hybrid substrates are ceramic. High-performance hybrids may have AlN, SiC, or diamond substrates used alone or in combinations.[7]

Hybrid circuits offer the advantages of structural rigidity and low leakage between devices due to the insulating property of the ceramic. They can provide a circuit with a mix of CMOS, bipolar and other components and offer functions not available in ASIC circuits.[8]

On the downside, they generally have a much lower density than integrated circuits and have a higher cost.

Multichip modules (MCMs)

Mounting individual chip packages on a PCB presents several problems. A chip package is several times the area of the chip taking up space on the board. Circuit resistance is increased by the individual resistances of all the package pins, and the electrical path lengths are multiplied by the number of chips and package leads. Each of these problems is reduced by mounting several chips on the same substrate. The technology is similar to hybrid circuits, but thick film screened components are not, or rarely, used.

Three types of MCMs have evolved. MCM-L (laminated) is similar to advanced laminated printed circuit boards, using copper foil conductors on plastic dielectrics. MCM-C (ceramic) are more like hybrid circuits. The substrate is a cofired ceramic with thick film conductors. MCM-D (deposited) uses ceramic, metal, or silicon substrates use deposited thin metal conductors.

The known good die (KGD) problem

In the individual package process, a final test assures quality of the completed product. If the chip has gone bad or the process was faulty, the entire chip and package is discarded. But the cost is much higher when putting bare die into hybrid, MCM, and printed circuit boards. These vehicles are more expensive to build and carry other expensive chips or components.

One option is to rely on the results of the wafer-sort test to certify die performance. Unfortunately, wafer sort does not include environmental tests or long-term reliability tests. However, performing a final test on a bare die is difficult if not impossible.

Package Type/Technology Summary

There are thousands of individual package types, and there is no uniform system of identifying them. Some are named by their design (DIP, flat packs, and so on), some are named by their construction technique (molded, CERDIP, and so forth), and others are named by their use, such as SMDs. When trying to understand a package type, keep in mind the three considerations: design type, construction technique, and use.

Into the future, wire-bonded/back-connected chip techniques will be overwhelmed by pin counts numbering in the hundreds. While exact packaging techniques are hard to project, the feeling is that flip-chip, ball grid arrays (BGAs), and chip scale packages (CSPs) will be the package designs of future.[9]

Review Questions

1. Name the four functions of a chip package.
2. What is the function of marking dies with ink dots?
3. Name the five parts of a package.
4. Define a hermetic package.
5. Give an example of a hermetic and nonhermetic package.

6. List, in order, the six major steps in the packaging process.
7. Describe the process and material used for eutectic die attachment.
8. Describe the principal thermosonic and ultrasonic bonding techniques.
9. What is a DIP package and how is it formed?
10. Name two methods of package marking.

References

1. H. Hinzen and B. Ripper, "Precision Grinding of Semiconductor Wafers," *Solid State Technology,* PennWell Publishing, August 1993, p. 53.
2. F. I. Blech and D. Dang, "Silicon Wafer Deformation after Backside Grinding," *Solid State Technology,* PennWell Publishing, August 1994, p. 74.
3. L. Plummer, "Packaging Trends," *Semiconductor International,* Cahners Publishing, January 1993, p. 33.
4. R. Iscoff, "Ultrathin Packages: Are They Ahead of Their Time?" *Semiconductor International,* Cahners Publishing, May 1994, p. 50.
5. R. Tummala and E. Rymaszewski, *Microelectronics Packaging Handbook,* Van Nostrand Reinhold, New York, 1989.
6. T. DiStefano and J. Fjelstad, "Chip-scale Packaging Meets Future Design Needs," *Solid State Technology,* April 1996, p. 82.
7. N. Nguyen, "Using Advanced Substrate Materials with Hybrid Packaging Techniques for Ultrahigh-Power ICs," *Solid State Technology,* PennWell Publishing, February 1993, p. 59.
8. R. Iscoff, "Will Hybrid Circuits Survive?" *Semiconductor International,* Cahners Publishing, October 1993, p. 57.
9. J. Baliga, "Package Styles Drive Advancements in Die Bonding," *Semiconductor International,* June 1997, p. 101.

Glossary*

acceptor An impurity that causes semiconducting materials to accept valence electrons, thereby leaving "holes" in the valence band. The holes act like carriers of positive charge, referred to as P type.

aligner (align and expose) A process tool used to align wafers and masks or reticles and expose the photoresist with a UV or other radiation source.

alignment Refers to the positioning of a mask or reticle with respect to the wafer.

alignment marks Targets on the mask and wafer used for correct alignment.

alloy (1) A compound composed of two metals. (2) In semiconductor processing, the alloy step causes the interdiffusion of the semiconductor and the material on top of it, forming on ohmic contact between them.

aluminum (Al) The metal most often used in semiconductor technology to form the interconnects between devices on a chip. It can be applied by evaporation or sputtering.

amorphous Materials with no definite arrangement of atoms, e.g., plastics are amorphous.

angstrom A unit of length, an angstrom (Å) is one ten-thousandth of a micron (10^{-4} μm) or 100,000,000 Å = 1 cm.

anistropic An etch process that exhibits little or no undercutting.

anneal A high-temperature processing step (usually the last one), designed to minimize stress in the crystal structure of the wafer.

antimony (Sb) A Group V element that is an N-type dopant in silicon. It is often used as the dopant for the buried layer.

arsenic A Group V element that is an N-type dopant in silicon.

assembly The series of operations after fabrication in which the wafer is separated into individual chips and mounted and connected to a package.

atmospheric oxidation A process of oxidation of silicon carried out at atmospheric pressure. The equipment used for thermal oxidation is the same as that used for thermal diffusion. It is composed of four subassemblies: a reactant source cabinet, a reaction chamber, a heating source, and a wafer holder.

---

*Source: Glossary terms are abstracted from Beverly Griggs, Anne Miller, and Peter Van Zant, *Semiconductor Terminology—Graphic Glossary of Terms,* Semiconductor Services, Redwood City, CA, 1989.

atomic force microscope (AFM) A microscope for profiling wafer surfaces by plotting the output of a spring-balanced probe moved over the surface.

atomic number A number assigned to each element equal to the number of protons (therefore the number of electrons) in the atom.

atomic particles The parts of an atom: electrons, protons, and neutrons.

base (1) The control portion of an NPN or PNP junction transistor. (2) The P-type diffusion done using boron that forms the base of NPN transistors, the emitter and collector of lateral PNP transistors, and resistors.

binary notation A way of representing any number using a power of 2 (the digits 0 and 1).

bipolar transistor A transistor consisting of an emitter, base, and collector, whose action depends on the injection of minority carriers from the base by the collector. Sometimes called NPN or PNP transistor to emphasize its layered structure.

bi-MOS A circuit containing both bipolar and MOS transistors.

boat (1) Pieces of quartz or metal joined together to form a supporting structure for wafers during high-temperature processing steps. (2) A Teflon® or plastic assemblage used to hold wafers during wet processing steps.

boat puller A mechanical arrangement to push a boat loaded with wafers into a furnace and/or withdraw it at a fixed speed.

BOE See *buffered oxide etch.*

bonding pads Electrical terminals on the chip (generally around the periphery) used for connection to the package electrical system.

boron (B) The P-type dopant commonly used for the isolation and base diffusion in standard bipolar integrated circuit processing.

boron trichloride (BCl_3) A gas that is often used as a source of boron for doping silicon.

bubbler An apparatus in which a carrier gas is "bubbled" through a heated liquid, causing portions of the liquid to be transported with the gas, e.g., a carrier gas (nitrogen or oxygen) is bubbled through deionized water at 98 to 99°C on its way to the oxidation tube.

buffered oxide etch A mix of hydrogen fluoride (HF) and ammonium fluoride (NH_4F) used to allow oxide etching to occur at a slow, controlled rate.

bump connection technology A structure of metal bumps formed on the bonding pads, allowing connection of the chip to a package when the chip is flipped over.

buried layer The N+ diffusion in the P-type substrate done just prior to growing the epitaxial layer. The buried layer provides a low-resistance path for current flowing in a device. Common buried layer dopants are antimony and arsenic.

can A metal package used for connecting a chip to a printed circuit board with from three to five leads.

capacitor A discrete device that stores electrical charge on two conductors separated by a dielectric.

capacitance Electrical charge storage capability.

capacitance-voltage plot (C/V plot) A plot that provides information on the amount of mobile ionic contamination present in the oxide.

carrier gas An inert gas that will transport atoms or molecules of a desired substance to a reaction chamber.

centistokes Units used to measure viscosity; centipoise divided by density.

channel A thin region of a semiconductor that supports conduction. A channel may occur at a surface or in the bulk, essential for the operation of MOSFETs and SIGFETs. In cases where channels are not part of the circuit design, their presence may indicate contamination problems or incomplete isolation.

channeling A phenomenon in which an ion beam will penetrate into the crystal planes of the wafer. Preventing channeling is accomplished by cutting the wafer "off orientation." The effect is to tilt the crystal planes relative to the beam direction.

charge carrier A carrier of electrical charge within the crystal of a solid-state device, such as an electron or hole.

chemical etching Selective removal of material by means of liquid reactants. The precision of the etch is controlled by the temperature of the etchant, the time of immersion and the composition of the acid etchant.

chemical mechanical polishing (CMP) A wafer flattening and polishing process that combines chemical removal with mechanical buffing. Used for polishing/flattening wafers after crystal growing and wafer planarization during the wafer fabrication process.

chemical vapor deposition See *CVD.*

chip Die or device, one of the individual integrated circuits or discrete devices on a wafer.

chrome A metal often used in mask fabrication to form the layer in which the circuit pattern is generated.

circuit board See *printed circuit board.*

circuit layout The calculation of the physical device dimensions required to produce the required electrical parameters. Vertical dimensions determine CVD and doping thickness specifications. Horizontal dimensions determine the wafer pattern dimensions and are the basis for a scaled drawing of the finished circuit (composite drawing).

class number Number of contaminant particles in a cubic foot of air.

cleanroom An area in which semiconductor device fabrication takes place. The cleanliness of the room is highly controlled so as to limit the number of contaminants to which the semiconductor is exposed.

clear field mask A mask on which the pattern is defined by the opaque areas.

cluster tool Several process stations or tools served by one loading–unloading chamber and wafer-transport system.

CMOS (complementary field-effect transistor) N- and P-channel MOS transistors on the same chip.

collector Along with the emitter and base, one of the three regions of the bipolar type of transistor.

collimated light Light in which the rays are parallel; used for gross visual inspection of surfaces.

composite drawing A scaled drawing of the finished circuit.

conductivity The ability of materials to conduct electricity (measured in siemens for conductance or ohms for resistance).

conductor A material that has low resistivity and high conductivity.

contact The regions of exposed silicon that are covered during the metallization process to provide electrical access to the devices.

contact aligner An aligner tool that clamps the wafer and mask into a tight contact before the resist exposure cycle.

contact mask The step at which holes are put into the wafer layers to allow the metal layer to reach down to the doped silicon substrate.

contamination A general term used to describe unwanted material that adversely affects the physical or electrical characteristics of a semiconductor wafer.

copper (Cu) Metal used to connect semiconductor devices on a chip surface. Generally used with a dual-damascene patterning process.

critical dimensions (CDs) The widths of the lines and spaces of critical circuit patterns as well as the area of contacts.

cryogenic pump A vacuum pump that can produce a vacuum to the 10^{-10} torr range, the same level as the vacuum of space. It does not require forepumps or cold traps and is faster than other types of vacuum pumps.

crystal A material in which the atoms are arranged in structured groups called *unit cells*.

crystal defects Vacancies and dislocations in a crystal that influence the electrical performance of a circuit.

crystal orientation The orientation of the primary crystal plane, expressed as Miller indices.

crystal planes The planes in the semiconductor crystal structure along which the die must be aligned so as to prevent "ragged" die edges when the wafer is separated into individual die.

CUM yield See *fabrication yield.*

current A measure of the number of charged particles passing a given point per unit time.

curve tracer A piece of electrical test equipment that displays the characteristics of a device visually on a screen.

CVD (chemical vapor deposition) A method for depositing some of the layers that function as dielectrics, conductors, or semiconductors. A chemical containing atoms of the material to be deposited reacts with another chemical, liberating the desired material, which deposits on the wafer while by-products of the reaction are removed from the reaction chamber.

Czochralski crystal grower A type of crystal grower that uses a seed to pull a crystal from a crucible of molten material.

dark field mask A mask on which the pattern is defined by the clear portion of the mask.

deep ultraviolet (DUV) A light wavelength often used to expose photoresist; it has the advantage of an ability to produce smaller image widths.

defect density The density of defects per square centimeter on a chip.

dehydration baking A heating process by which wafer surfaces are restored to a hydrophobic condition by baking. Surface water is evaporated from the wafer at elevated temperatures.

deionized (DI) water Process water that is free of dissolved ions. Specification levels are generally 15 to 18 mΩ of resistance.

depletion layer The region in a semiconductor where essentially all charge carriers have been swept out by the electric field that exists there.

deposition Process in which layers are formed as the result of a chemical reaction in which the desired layer material is formed and coats the wafer surface.

design rule The minimum feature size of a circuit.

develop inspection The first inspection in the photomasking process consisting of measurement of critical dimensions and inspection for defects. It is done after development or development and hard bake if an automatic baking system is used.

development A photoresist processing step in which photoresist is removed from areas defined by the masking and exposure step of wafer fabrication.

developer Chemical used to remove areas defined in the masking and exposure step of wafer fabrication.

device A single-function component such as a transistor, resistor, or capacitor.

DI water Deionized water; purity of this water is measured by its resistivity, with the standard being 18 MΩ.

diborane (B_2H_6) A gas that is often used as a source of boron for doping silicon.

die One unit on a wafer separated by scribe lines; after all of the wafer fabrication steps are completed, die are separated by sawing; the separated units are referred to as *chips*.

die bonding Assembly step in which individual chips are attached to the package with conductive adhesives or metal alloys.

die sort See *wafer sort*.

dielectric A material that conducts no current when it has a voltage across it. Two dielectrics encountered in semiconductor processing are silicon dioxide and silicon nitride.

diffusion A process used in semiconductor production that introduces minute amounts of impurities (dopants) into a substrate material such as silicon or germanium and permits the impurity to spread into the substrate. The process is very dependent on temperature and time.

diffusivity The rate of movement or diffusion of dopants in a semiconductor.

diode Device that enables current flow in one direction but not in the other.

DIP (dual in-line package) A rectangular circuit package, with leads coming out of the long sides and bent down to fit into a socket.

discrete device A circuit having a single electrical function. Discrete devices include capacitors, resistors, transistors, diodes, and fuses.

dislocation A discontinuity in the crystal lattice, a type of crystal defect.

DMOS (diffused MOS) A transistor structure that features a narrow (channel length) separation between the source and drain. The channel length is created by two sequential diffusions through the same hole.

donor An impurity that can make a semiconductor N-type by donating extra "free" electrons; electrons carry a negative charge.

dopant An element that alters the conductivity of a semiconductor by contributing either a hole or electron to the conduction process. For silicon, the dopants are found in Groups III and V of the periodic table.

dopant deposition The first step in the diffusion process, in which the dopant atoms are diffused into the wafer surface.

doping The introduction of an impurity (dopant) into the crystal lattice of a semiconductor to modify its electronic properties—for example, adding boron to silicon to make the material P type.

DRAM (dynamic random access memory) Memory device for the storage of digital information. The information is stored in a "volatile" state.

drain Along with the source and gate, one of the three regions of a unipolar or field-effect transistor (FET).

drive-in Stage in diffusion where the dopant is driven deeper into the wafer.

dry etch See *plasma etch.*

dry ox The growth of silicon dioxide using oxygen and hydrogen, which form water vapor at process temperatures, rather than using water vapor directly.

dry oxide Thermal silicon dioxide grown using oxygen.

dual damascene A patterning process that first defines the required pattern in a trench in the top wafer surface followed by overfilling with a conductive metal. The overfill is removed, usually by a chemical-mechanical-polishing process, leaving the metal pattern within the trench.

e-beam (electron beam) An exposure source that allows direct image formation without a mask. An e-beam can be deflected by electrostatic plates and therefore directed to precise locations, resulting in the generation of submicron-size patterns.

e-beam aligner An aligner tool that exposes the resist-coated wafer surface by steering (writing) an electron beam across the wafer surface.

e-beam evaporation (electron beam evaporation) Phase change that uses the energy of a focused electron beam to provide the required energy to change solid metal or alloys from solid to gas.

e-beam exposure system A machine in which the image pattern is stored in a computer memory and used to control the electrostatic plates that, in turn, direct the e-beam, resulting in the generation of patterns without the use of reticles or photomasks.

edge die The incomplete die located on the edge of the wafer.

EEPROM (electrically erasable PROM) A memory circuit with the capability of data erasure and acceptance of new information by the application of an electrical pulse.

electromigration The diffusion of electrons in electric fields set up in the lead while the circuit is in operation. It occurs in aluminum and is exhibited as a field failure, not as a process defect. The metal thins and eventually separates completely, causing an opening in the circuit.

electron A charged particle revolving around the nucleus of an atom. It can form bonds with electrons from other atoms or be lost, making the atom an ion.

ellipsometer An instrument that uses laser light sources to measure thin film thickness.

emitter (1) The region of a transistor that serves as the source or input end for carriers. (2) The N-type diffusion usually done using phosphorus, which forms the emitter of NPN transistors, the base contact of PNP transistors, the N+ contact of NPN transistors, and low-value resistors.

epitaxial (Greek for "arranged upon.") The growth of a single-crystal semiconductor film upon a single-crystal substrate. The epitaxial layer has the same crystallographic characteristics as the substrate material.

epoxy package See *molded package.*

EPROM (erasable PROM) A memory circuit with the capability of data erasure and acceptance of new information.

etch A process for removing material in a specified area through a wet or dry chemical reaction or by physical removal, such as by sputter etch.

evaporation A process step that uses heat to change a material (usually a metal or metal alloy) from its solid state to a gaseous state with the result of the source being deposited on wafers. Both electron beam and filament evaporation are common in semiconductor processing.

exposure Method of defining patterns by the interaction of light or another form of energy with photoresist that is sensitive to the energy source.

fabrication Integrated circuit manufacturing processes.

fabrication yield The percentage of wafers arriving at wafer sort compared with the number started into the process.

feature size The minimum width of pattern openings or spaces in a device.

FET (field-effect transistor) A transistor consisting of a source, gate, and drain, whose action depends on the flow of majority carriers past the gate from the source to drain. The flow is controlled by the transverse electric field under the gate. See *unipolar transistor.*

field oxide The region on an electrical device where the oxide serves the function of a dielectric.

final test The final assembly step in which the packaged die is put through its last electrical test.

flash memory An EPROM or EEPROM with the capability of block erasure of data in the memory array.

flat zone The highly temperature-controlled region of a tube furnace.

flip-chip joining A chip or package connection process where "bumps" of connecting metal are formed on the chip surface and the chip is "flipped" over for soldering to the package.

four-point probe A piece of electrical test equipment used to determine the sheet resistivity of a wafer.

furnace A piece of equipment containing a resistance-heated element and a temperature controller. It is used to maintain a region of constant temperature with a controlled atmosphere for the processing of semiconductor devices.

fuse A circuit component that can be blown to allow a desired memory cell or logic gate to be programmed.

gallium arsenide (GaAs) The most common of compound semiconductor materials. It has the advantage of producing higher-speed devices than those produced using silicon as a substrate.

gate Along with the source and drain, one of the three regions of the unipolar or field-effect transistor (FET or MOS).

gate array Type of integrated circuit made up of an arrangement of interconnected gates used to provide custom functions.

gate oxide (gate ox) The thin oxide that causes the induction of charge, creating a channel between source and drain regions of a MOS transistor.

germanium Semiconducting material used in the manufacture of crystal diodes and of early transistors.

HEPA filter (high-efficiency particulate attenuator) A filter constructed of fragile fibers in an accordion-folded design that allows a larger filtering area at an air velocity low enough for operator comfort. This filter permits a filtering efficiency of 99.99 percent.

hexamethyldisilizane (HMDS) Primer used to promote photoresist adhesion.

high-pressure oxidation Oxidation carried out at high pressure (10 to 20 atm) to reduce the amount of heat or time required. The reaction chamber for this process must be constructed of stainless steel to safely contain the pressure.

hole (1) The absence of a valence electron in a semiconductor crystal. Motion of a hole is equivalent to motion of a positive charge. (2) A "hole" in a surface layer created by the photomasking process.

hybrid integrated circuit A structure consisting of an assembly of one or more semiconductor devices and a thin-film integrated circuit on a single substrate, usually of ceramic.

hydrofluoric acid (HF) An acid used to etch silicon dioxide; often diluted or buffered before it is used.

hydrogen (H_2) A gas used in semiconductor processing primarily as a carrier gas for high-temperature reaction steps such as epitaxial silicon growth.

hydrophilic Affinity toward water (water-loving); a hydrophilic surface is one that will allow water to spread across it in large puddles.

hydrophobic Aversion to water; a hydrophobic surface is one that will not support large pools of water. The water is pulled into droplets on the surface. These surfaces often are termed "dewetted."

hydroscopic Attracts and absorbs water.

integrated circuit A circuit in which many elements are fabricated and interconnected on a single chip of semiconductor material, as opposed to a "nonintegrated" circuit, in which the transistors, diodes, resistors, and so on, are fabricated separately and then assembled.

integration level The range of total component count in a die. Varies from SSI (small-scale integration, less than 50 components) to ULSI (ultra-large-scale integration, over 1,000,000 components).

intrinsic semiconductor An element or compound that has four electrons in its outer ring (i.e., elements from Group IV of the periodic table or compounds of Group III and V).

ion An atom that has either gained or lost electrons, making it a charged particle (either negative or positive).

ion beam milling A dry etching method that uses an ion beam. Argon atoms are ionized and accelerated toward a wafer. The exposed areas are removed through a sputtering action.

ion implantation Introduction of selected impurities (dopants) by means of high-voltage ion bombardment to achieve desired electronic properties in de- fined areas.

interconnect See *lead.*

isolation diffusion Diffusion step resulting in P-N junctions surrounding the areas to be separated.

isotropic etching Refers to the etching of the photoresist both downward and to the side.

JFET (junction field-effect transistor) Device in which voltage is applied to a terminal to control current between the source and drain regions.

junction The interface at which the conductivity type of a material changes from P type to N type or vice versa.

killer defect A defect that causes the failure of a device or circuit.

lateral diffusion The diffusion of dopants from side to side every time the wafer is heated near the diffusion temperature range.

layering A process by which thin layers of different materials are grown on, or added to, the wafer surface.

lead A metal strip on the wafer surface.

LED (light-emitting diode) A semiconductor device in which the energy of minority carriers in combining with holes is converted to light. Usually, but not necessarily, constructed as a P-N junction device.

lift-off process A material removal process in which the material is deposited into a hole in a photoresist layer and the pattern defined where the photoresist layer is removed (lifted off) the surface.

light field mask See *clear field mask.*

lithography Process of pattern transfer; when light is utilized, it is termed photolithography; when patterns are small enough to be measured in microns, it is referred to as microlithography.

local oxidation of silicon (LOCOS) A MOS surface isolation scheme in which silicon dioxide is grown around islands of silicon nitride, which is in turn removed to leave oxide-free areas for the formation of the circuit components.

low-pressure CVD (LPCVD) A CVD system and process performed in a low-pressure range.

LSI (large-scale integration) Refers to chips with between 5000 and 100,000 components each.

majority carrier The mobile charge carrier (hole or electron) that predominates in a semiconductor material—for example, electrons in an N-type region.

mask A glass plate covered with an array of patterns used in the photomasking process. Each pattern consists of opaque and clear areas that, respectively, prevent or allow light through. Masks are aligned with existing patterns on silicon wafers and used to expose photoresist. Mask patterns may be formed in emulsion, chrome, iron oxide, silicon, or a number of other opaque materials.

masking See *patterning*.

memory The storing of data or information.

metal mask The step at which an island of conducting material is left on the wafer surface.

metalorganic CVD (MOCVD) A VPE process that uses halides and metalorganic sources.

micrometer One-millionth of a meter (10^{-6} m); abbreviation is μm.

micron Same as micrometer.

microchip See *chip*.

Miller indices A numerical system of three numbers used to identify the orientation of planes in a crystal.

mini-environment An environment that maintains wafer cleanliness by storing, transporting, and loading or unloading wafers in small, clean enclosures.

minority carrier The nonpredominant mobile charge carrier in a semiconductor—for example, electrons in a P-type region.

MMOS (memory MOS) A nonvolatile memory device structure that enables information to be retained during power shutdown.

mobile ionic contaminant Wafer contaminants that are electrically charged particles, which can cause electrical failures in a wafer or device.

molded package A package of epoxy or other polymer material that is molded around the chip and package lead system.

molecular beam epitaxy (MBE) An evaporative deposition process capable of extreme control of the deposition process.

molecule Smallest quantity of a substance that retains the properties of that substance.

monochromatic light Light of a single wavelength.

MOSFET A field-effect transistor containing a metal gate over thermal oxide over silicon.

MSI (medium-scale integration) Refers to chips with between 50 and 5000 components each.

multichip module (MCM) A package containing two or more IC chips connected by a thin-film metal system.

multilayer resist process An image-resolution process that uses two or more layers of photoresist.

nanometer (nm) length unit = 1×10^{-9} meter.

nanotechnology Processes and materials used to build semiconductor devices and other structures with nanometer dimensions.

negative resist Photoresist that remains in areas that were not protected from exposure by the opaque regions of a mask while being removed by the developer in regions that were protected. A negative image of a mask remains following the develop process. A "clear" or "light" field mask is most often used with negative resist.

nitric acid (HNO_3) A strong acid often used to clean silicon wafers or etch materials.

nitridation The formation of silicon nitride by the high-temperature exposure of a silicon surface to nitrogen.

nitrogen (N_2) A gas that seldom reacts with other materials. It is often used as a carrier gas for chemicals in semiconductor processing.

NMOS N-channel MOS; type of MOSFET in which the channel is negative during conduction.

nonvolatile memory circuit A memory circuit that retains its data when power to the chip is lost.

N-type A semiconductor material in which the majority of carriers are electrons and therefore negative. N-type dopants in silicon are Group V elements, in which the fifth outer electron is free to conduct current.

NPN transistor A transistor that has a base of P-type silicon sandwiched between an emitter and a collector of N-type silicon.

Ohm's law A relationship between resistance, voltage, and current; $R = V/I$.

oil diffusion pump A type of high-vacuum pump that uses evaporated hot oil particles to "push" chamber particles out of the system.

overall yield The percentage of functioning packaged chips from a wafer related to the number of die mapped onto the wafer. Overall yield is the product of fabrication yield, sort yield, and assembly yields.

oxidation The growth of oxide on silicon when exposed to oxygen. This process is highly temperature dependent.

oxidation reaction chamber A chamber in which oxidation takes place. Quartz or silicon carbide tubing is used to make an oxidation reaction chamber due to their thermal resistance and purity.

oxide See *silicon dioxide.*

oxide etching An etching process that uses acid—usually hydrofluoric acid (HF). The acid must be buffered for the reaction to proceed at a rate slow enough to be controlled. Buffered oxide etch (BOE) is often used.

oxygen (O_2) Gas used to combine with silicon to form silicon dioxide.

package Protective container for semiconductor chip having electrical leads.

packaging yield The percentage of packaged die passing the final tests compared to the number of good die that entered the packaging process.

passivation Sealing layer added at the end of the fabrication process to prevent deterioration of electronic properties through chemical action, corrosion, or handling during the packaging processes. The passivation layer, usually silicon dioxide or silicon nitride, protects against moisture or contamination.

patterning A process in which the pattern in a reticle or photomask is transferred to a wafer resulting in the identification of areas to be doped or selectively removed.

pellicle A thin film of an optical-grade polymer that is stretched on a frame and secured to a mask or reticle. This solves the problem of airborne dirt collecting on the mask and acting as an opaque spot. During the exposure, the dirt is held out of the focal plane and does not "print" onto the wafer.

phosphine (PH_3) A gas that is often used as a source of phosphorus for doping silicon.

phosphorus (P) The N-type dopant commonly used for the sinker and emitter diffusions in standard bipolar integrated circuit technology.

phosphorus oxychloride ($POCl_3$) A liquid that is often used as a source of phosphorus for doping silicon.

photomasking See *patterning.*

photoplate A coated mask blank before imaging.

photoresist The light-sensitive film spun onto wafers and "exposed" using high-intensity light through a mask. The exposed (or unexposed, depending on its polarity) photoresist is dissolved with developers, leaving a pattern of photoresist that allows etching to take place in some areas while preventing it in others.

pinhole A small, undesired hole in the photoresist or in the opaque region of a mask or reticle.

pin grid array (PGA) A large chip package with many leads coming out of the entire bottom surface of the package.

planar structure A flat-surfaced device structure fabricated by diffusion and oxide masking, with the junctions terminating in a single plane.

plasma High-energy gas made up of ionized particles.

plasma-enhanced CVD (PECVD) A CVD system and process using plasma energy to drive the deposition.

plasma etch A dry-etch process using reactive gases energized by a plasma field.

plastic package See *molded package.*

plug (via plug) A metal (generally a refractory metal) deposited in a connecting via hole between the conducting layers of a multilayer metal system.

PMOS (P-channel MOS) Type of MOSFET in which the channel is positive due to conduction achieved by holes.

PNP Semiconductor crystal structure consisting of an N-type region sandwiched between two P-type regions, as commonly used in bipolar transistors.

polycide MOS gate A MOS gate structure composed of a sandwich of silicon dioxide topped with a layer of polysilicon that is covered with a refractory metal silicide.

polycrystalline silicon (poly) Silicon composed of many crystal unit cells randomly arranged.

polymer A complex organic chemical compound made up of repeating units.

positive resist Photoresist that is removed in areas that were not protected from exposure by the opaque regions of a mask while remaining after develop in regions that were protected from exposure. A positive image of the mask remains following the develop process. A "dark field" mask is used most often with positive resist.

post exposure bake (PEB) A baking step performed after resist exposure to reduce standing wave effects.

predeposition (predep) The process step during which a controlled amount of a dopant is introduced into the crystal structure of a semiconductor.

primer chemical A chemical that enhances the adhesion of a desired layer (in semiconductor technology, the layer is usually photoresist).

programmable read-only memory See *PROM.*

projection alignment An exposure system in which the image on the mask is projected onto the wafer. This results in little mask or photoresist damage and has about the same productivity as the contact method. For LSI and VLSI, projection alignment is standard.

projection aligner An aligner tool that projects the mask or reticle image over a distance onto the wafer.

PROM (programmable read-only memory) A technology in which fuses are used in every memory cell, and selected fuses are blown so as to program the chip with user-specified information.

proximity aligner An aligner tool that holds the wafer and mask or reticle a small distance apart during the resist exposure cycle.

P-type Semiconductor material in which the majority carriers are holes and therefore positive. P-type dopants in silicon are Group III-A elements.

quartz Commercial name for silicon dioxide formed into glass products. Because of its high temperature resistance, quartz is used in many processing steps in integrated circuit fabrication.

RAM (random access memory) Device that temporarily stores digital information.

rapid thermal oxidation (RTO) An oxidation process performed in a rapid thermal processing (RTP) tool.

rapid thermal processing (RTP) A single-wafer processing tool that uses high-intensity lights or other sources to heat and cool the wafer in milliseconds.

RCA clean A multiple-step process to clean wafers before oxidation; named after RCA, the company that developed the procedure.

reactive ion etching (RIE) An etching process that combines plasma and ion beam removal of the surface layer. The etchant gas enters the reaction chamber and is ionized. The individual molecules accelerate to the wafer surface. At the surface, the top layer removal is achieved by the physical and chemical removal of the material.

reactor (1) A piece of equipment used for the deposition of a layer of material used in semiconductor processing. Common types of reactors are epitaxial reactors, vapox reactors, and nitride reactors. (2) See *plasma etcher.*

resistivity A measure of the resistance to current flow in a material. A function of the attraction between the outer electrons and inner protons of a material. The more tightly bound the electrons, the greater the resistivity.

resolution capability The minimum feature size capability of a photolithography process or tool.

reticle An exposure mask with only a portion of complete die pattern.

rinse The removal of wet etchants or developers from the wafer. This process results in stopping the etching or developing processes and removing the active chemical from the surface. There are several different methods of rinsing: overflow rinsing, spray rinsing, dump rinsing, and spin-rinse dryers.

ROM (read-only memory) A memory circuit containing permanent data with no capability of accepting new information.

salicide MOS gate A polycide MOS gate structure sequenced in the process to self-align to the source or drain. See *polycide MOS gate.*

scanning electron microscope (SEM) Microscope used to magnify images as much as 50,000 times by means of scanning with an electron beam. The impinging electrons cause electrons on the surface to be ejected. The ejected electrons are collected and translated into a picture of the surface.

scribe lines Lines used to separate die on a wafer. The wafer will be sawed along the scribe lines, resulting in individual chips.

self-aligned gate A MOS structure that allows the direct alignment of the source or drain to the gate without a photoresist alignment step.

semiconductor An element such as silicon or germanium, intermediate in electrical conductivity between the conductors and the insulators, in which conduction takes place by means of holes and electrons. Common single-element semiconductors are Si (silicon) and Ge (germanium); a common compound semiconductor is GaAs (gallium arsenide).

sheet resistance A measurement of resistance with dimensions of ohms per centimeter squared that shows the number of N-type or P-type donor atoms in a semiconductor.

side diffusion see *lateral diffusion.*

silicon (Si) The Group IV element used for fabricating diodes, transistors, and integrated circuits.

silicon dioxide (SiO_2) A nonconducting layer that can be thermally grown or deposited on silicon wafers. Thermal silicon dioxide is commonly grown using either oxygen or water vapor at temperatures above 900°C.

silicon gate MOS A MOS gate structure with a layer of polysilicon on top of a thin layer of silicon dioxide.

silicon nitride (Si_3N_4) A nonconductive layer chemically deposited on wafers at temperatures between 600 and 900°C. When it is the final layer in the process, it protects devices against contamination.

single crystal Refers to substances that have all unit cells arranged in a definite and repeated fashion as opposed to polycrystalline materials, which have unit cells randomly arranged.

slope etching Controlled undercutting; an etch strategy in which the sides of the contact holes are purposely overetched so as to reduce the shadow effect of the sidewall and the resultant thinning of the film.

soft baking A heating process used to evaporate a portion of the solvents in resist. The term "soft" describes the still soft resist after baking. The solvents are evaporated to achieve two results: to avoid retention of the solvent in the resist film and to increase the surface adhesion of the resist to the wafer.

solid-state electronics Designation used to describe devices and circuits fabricated from solid materials such as semiconductors, ferrites, or thin films, as distinct from devices and circuits making use of electron tube technology.

source Along with the gate and drain, one of the three regions of a unipolar or field-effect transistor (FET).

spinning A technique in which the photoresist is spun onto the wafer, resulting in a typical photoresist layer 0.5 μm thick with an allowable thickness variation of 10 percent.

spin rinse dryer A machine that automatically rinses and dries wafers by spinning them in cassettes around a central axis.

sputtering A method of depositing a thin film of material on wafer surfaces. A target of the desired material is bombarded with radio-frequency-excited ions that knock atoms from the target; the dislodged target material deposits on the wafer surface.

spectrophotometer An analytical instrument used to collect interference measurements that are used to calculate film thickness.

spreading resistance A technique used for measuring the dopant concentration profile in a wafer.

SSI (small-scale integration) Refers to chips with between 2 and 50 components each.

standard mechanical interface (SMIF) A system that allows the mating of portable clean wafer boxes (called *pads*) to the clean microenvironment loading stations of process tools.

standing wave effects A vertical resist exposure pattern that follows standing waves set up in the resist layer by constructive interference of the exposure light reflecting off the wafer surface.

static RAM (static random access memory) Fast read-write memory cell based on transistors.

steam oxide Thermal silicon dioxide grown by bubbling a gas (usually oxygen or nitrogen) through water at 98 to 100°C.

step and repeat An operation in which the pattern on the reticle is transferred to the mask or wafer. The photoresist-coated mask blank (chrome, emulsion, or iron oxide) or wafer is placed on an *x-y* stage, and the reticle pattern is repeatedly imaged until the entire surface is filled with the reticle pattern.

step coverage The ability of new layers to evenly cover steps formed in the existing wafer layers.

stepper An aligner tool that aligns and exposes one (or a small number) of die at a time. The tool "steps" to each subsequent die on the wafer.

stripping Removal process; usually refers to photoresist.

subcollector See *buried layer.*

substrate The underlying material upon which a device, circuit, or epitaxial layer is fabricated.

sulfuric acid (H_2SO_4) A strong acid often used to clean silicon wafers and to remove photoresist.

susceptor The flat slab of material (usually graphite) upon which wafers are held during high-temperature deposition processes such as epitaxial growth or nitride deposition.

tape automated bonding (TAB) A chip-to-package connection process in which the package leads are formed on a flexible tape, and all the lead fingers are bonded to the chip in one action.

target The material to be sputtered during the sputtering process.

test die Die on a wafer that appear to have a different pattern from most others. These contain test devices created by the same processes at the regular die; however, the devices on these die are designed on a larger scale to allow in-process quality control.

TCE (trichloroethylene) A solvent used for wafer and general cleaning.

tetraethylorthosilicate (TEOS) A chemical source for the deposition of silicon dioxide.

thermal diffusion A process by which dopant atoms diffuse into the wafer surface by heating the wafer in the range of 1000°C and exposing it to vapors containing the desired dopant.

thermal oxide On silicon semiconductor devices, an oxide fabricated by exposing the silicon to oxygen at high temperatures. The resulting interface has low levels of ionic impurities and defects (surface states).

thermocouple A device to measure the temperature in a furnace of a reactor. It is made by welding two wires together at a point. Heat generates a voltage between the two materials that is proportional to the temperature.

torr Pressure unit; international standard unit replacing the English measure, millimeters of mercury (mmHg).

transistor A semiconductor device that uses a stream of charge carriers to produce active electronic effects. The name was coined from the electrical characteristic of "transfer resistance."

tube (1) See *furnace.* (2) A cylindrical piece of quartz with fittings on one or both ends. It is placed in a furnace to provide a contamination-free and controlled atmosphere.

undercutting See *isotropic etching.*

ULSI (ultra-large-scale integration) Refers to chips with more than 1,000,000 components each.

ultraviolet (UV) light A portion of the electromagnetic spectrum from 250 to 500 nm. High-pressure mercury sources emit UV light for photoresist exposure.

unipolar transistor A transistor such as an FET whose action depends on majority carriers only.

vacancy (1) A position in the crystal for an atom which is empty. (2) A type of crystal defect.

vacuum A low-pressure condition.

vapor phase epitaxy (VPE) An epitaxial deposition system that can combine several source gases to deposit compound semiconductors.

vapor priming A technique in which primer is applied in a vapor state such that the wafer never comes in contact with any possible contamination in the liquid or, in the case of HMDS, any particles of hydrolyzed HMDS.

via Vertical opening filled with conducting material used to connect circuits on various layers of a device to one another and to the semiconducting substrate; serves same purpose as "contacts."

viscosity The qualitative measure of liquid flow. Viscosity measurements are made by measuring the force required to move an object through the liquid. It is a measurement of "internal friction."

VLF hood A workstation with vertical laminar airflow to keep particulate levels low.

VLSI (very-large-scale integration) Refers to chips with between 100,000 and 1,000,000 components.

volatile memory circuit A memory circuit that loses its data when power to the chip is lost.

voltage The force applied between two points causing charged particles (and hence current) to flow.

wafer A thin, usually round, slice of a semiconductor material from which chips are made.

wafer fabrication The series of manufacturing operations in which the circuit or device is put in and on the wafer.

wafer flat Flat area(s) ground onto the wafer's edges to indicate the crystal orientation of the wafer structure and the dopant type.

wafer sort The step after wafer fabrication during which the electrical parameters of integrated circuits are tested for functionality. Probes contact the pads of the circuit to conduct the test, leading to the name "prober" for the equipment that performs electrical tests on each die site of completed wafers.

wafer sort yield The number of functioning die at wafer sort as compared to the total number of die started; typically, the lowest major yield point for integrated circuits.

wire bonding An assembly step in which thin gold or aluminum wires are attached between the die bonding pads and the lead connections in the package.

X-ray aligner An aligner tool that uses X-rays and a mask to expose resist-coated wafers.

X-ray exposure system Imaging system using X-rays as the exposure source. Due to their short wavelengths, X-rays exhibit no detrimental diffraction effects.

yield A percentage used in the semiconductor industry that indicates the amount of finished products leaving a process as compared to the amount of product entering that process.

Index

A

acceleration tube, 349
accumulative wafer-fabrication, 140–145
acids, 46–47
acoustic wave devices, 547
adiabatic cooling, 251
adiabatic expansion, 429
aerial image, 285
air quality, 97–103
airborne molecular contaminants, 95
alignment (wafer), 231–243
alkalis, 46–47
alloying, 403
alternating phase shift mask [alternating aperture phase shift mask (AAPSM)], 296–297
alternative transistor designs, 531–532
aluminum, 405
aluminum-copper alloy, 406
aluminum-silicon alloys, 405
amorphous silicon deposition, 53, 392–394
amplification, 2
anisotropic etching, 260
annealing (doping), 79, 180, 357
annular ring illumination, 298
antireflective coatings (ARCs), 301–303
atmospheric pressure CVD systems (APCVD), 372–377
atomic force microscopy (AFM), 451–452, 462–464
atomic layer deposition (ALD), 381–382
atomic number, 27–29
atomic structure, 26–28
auger electron spectroscopy (AES), 465
autodoping, 390
automated guided vehicles (AGVs), 496
automatic defect detection, 460
automatic spinners, 225–227
automation (wafer fabrication), 494–509
autoprofiling, 173

B

back end of the line (BEOL), 16, 85, 273
backside plating, 67, 416, 590–591
backstreaming, 427
bacterial contaminants (*see* contaminants)
baking, 227–231
ball bonding, 585–586
ball grid arrays (BGAs), 587, 604
ballroom layout, 102
bare die techniques, 606
barrel etchers, 266
barrel radiant-heated PECVD, 381
barren radiant-induction-heated APCVD 374–375
barrier layer, 403
barrier metals, 406–407
barrier/liner deposition, 412
bases, 46–47
batch versus single-wafer processing, 500–501
beam focus (ionization), 351
beam scanning (ionization), 351–352
Bell, Alexander Graham, 448
Bi-MOS, 543–544
binary notation, 555–557
bipolar circuit formation, 534–538
bipolar structure, 17–19
bipolar transistors, 472
bird's beak, 185, 538–539
blob top, 606
Bohr atom, 26–27, 41
book-to-bill ratio, 490

Boole, George, 555–556
borosilicate glass (BSG), 397
Bose-Einstein (wafer sort), 152–153
breakdown voltage, 471–472
bridges (mask), 146
bubbler, 130, 176, 186
Buckminster Fuller, 555
budget, 280
bump/flip-chip bonding, 587–589, 601
buried layers, 391
burn-in tests, 600

C

C/V plot, 475–477
capacitance-voltage plotting, 474
capacitors, 30, 517–519
capacitor electrodes, 415
capillary force, 122
carrier mobility, 37
carriers, 33
cavitation, 250
cellulose acetate (AC), 299–300
centipoises, 215
CERDIP packages, 593–594, 603
channeling (doping), 358
chemical cleaning solutions, 124–125
chemical contaminants, 95
chemical mechanical polishing (CMP), 42, 310–316, 414–415
chemical purity, 48
chemical vapor deposition (CVD), 74, 162, 173, 366, 369–381
chemical, contamination, 115–116
chemically amplified (CA) resist, 286
chemical-mechanical polishing (CMP), 304
chip and wafer size, 7–10
chip characteristics, 571–572
chip cost, 11
chip uses, 12
chip, 16
chip-on-board (COB), 17
chip-scale packages, 605
chlorine-added oxidation, 187
circuit density (die), 150
circuit design, 79–82
cleaning techniques, 127–128
cleanroom, 101–109, 119–120
closed-loop control-system automation, 497
clustering, 495
CMOS epitaxy, 391
cold-wall induction, 374
collimated, 234
color-active matrix liquid crystal displays (AMLCDs), 548–550
columnar poly, 393
complementary MOS (CMOS) circuits, 19–20, 540–543
composition (wafer), 467
computer-aided process planning (CAPP), 499
computer-integrated manufacturing (CIM), 498–499
concentration/depth profile, dopant, 441
conductors
  CVD, 399
  generally, 29–34, 531–532
  metallization, 405–412
  multilevel metal schemes, 403–407
  single-level metal, 402–403
conformal layer, 305–306
constant acceleration, 598
contact aligners, 237
contact holes, 402
contaminants, 92–96
contamination
  generally, 454–462, 465–467
  personnel generated, 111–113
  problems, 95
  sources, 97–98
continuous conduction-heated APCVD, 376
contrast effects (mask), 286
contrast enhancement layers (CEL), 318
contrast threshold, 318
copper (metallization), 409
critical dimension (CD), 280, 452
crusting, 273
cryogenic cleaning, 128
cryogenic pumps, 429–430
crystals
  defects, 59–60, 357
  generally, 53
  growth, 56–60
  orientation, 55–56, 62–63
  quality, 59–60
cum fab yield, 141–143, 151

CVD metallization, 425–426
Czochralski, 56–57

D

Dark-field inspection, 457
defect density (die), 10, 150
defect detection, 454–462
deflashing, 597
deflux, 590
dehydration baking, 218–219
densification, 396
density, 44–45
deposited films (CVD), 385
deposited semiconductors (CVD), 385–386
deposition
  doping, 333–338
  metallization, 416
depth of focus/field, 293–294
DESIRE process, 307–308
develop inspect, 254–257, 277
device dielectric (silicone dioxide), 162
device electrical measurements, 467–477
device processing yield, 95–96
device reliability, 95–96
devitrification, 172
diameter grinding, 62–64
diamond semiconductors, 40
diatomic molecules, 42
dichlorosilane source chemistry, 388
die
  area, 150
  attach, 582–584
  generally, 16
  sort, 87–89
  wafers, 147–148
dielectrics
  constant, 399
  generally, 30, 369
  isolation, 536–537
  wearout, 269
die-to-database system, 462
die-to-package bonding, 584
differential Hall effect (DHE), 442
differential oxidation, 169
diffraction effect, 234, 288
diffused MOS (DMOS), 529
diffusion (doping), 326–329
diffusion limited reaction, 164
diffusivity, 333–334
dimensional control, 233
diode sputtering, 423
diodes, 3, 470, 519–521
DIP packages, 603
DIP through-hole assembly, 573
direct writing, 242, 292
discrete devices, 4
dislocations, 61
dissolution inhibitor systems, 207
doping
  concentration, 355–359
  polysilicon, 414
  semiconductors, 32–37
  sources, 337–341
  silicon dioxide, 397
doping barrier (silicone dioxide), 161
double masking, 298
downstream plasma processing, 269
downstream strippers, 276
DRAM, 8, 12–13
drawback (spinners), 225
drive-in oxidation, 341–345
dry (or plasma) development, 252
dry cleaning, 128
dry etching, 264–273
dry oxidation, 165–166
dry oxygen, 185–186
drying, 133–134
dryox (*see* dry oxidation)
dual in-line package (DIP), 602
dual-damascene process, 309–310, 411
dyed resists, 318
dynamic dispense, 223–224

E

e-beam gun, 418
edge bead removal, 225
edge die (wafer), 147
electrical conduction, 29
electrical sort, 87–89
electricity, static, 108–110
electrochemical plating (ECP), 412–414
electromagnetic spectrum, 209
electromigration, 406
electron beam aligners, 242
electron beam scanning, 291–292

electron conduction, 33
electron cyclotron resonance (ECR), 267
electron flood gun, 354
electron spectroscope for chemical analysis (ESCA), 466
electron-beam projection lithography (EPL), 295
Electronic Numeric Integrator and Calculator, 2–3
Electroplating (*see* electrochemical plating)
electrostatic beam scanning, 352
elements, 26–28
elephant tube (wafer), 188
ellipsometers, 446–447
emissivity, 182
encapsulation, 590
end cropping, 62
end station (ionization), 353
energy-sensitive polymers, 204–205
engineered substrates, 39
engineered wafers, 69
ENIAC, 2–3
environmental, safety, and health (ESH), 501
epitaxial
  film, 388–389
  layer, 18, 366, 385
  process, 391
  silicon, 386
epoxy molding, 593
equipment (wafer fabrication), 491–494
equipment, cleaning, 110–111, 118
error budget, 280
error function, 334–335
etch
  definition, 319–320
  photoresist, 258–262
  profile control, 321
etchback planarization, 309
etching, 536
eutectic alloying, 406
eutectic die attach, 582
evaporation, 74
excimer lasers, 289
exponential model (wafer sort), 153–154
exposure (wafer), 231–243
exposure response curve, 210
exposure, photoresist, 285–287

F

fab floor layout, 500
factory-level automation (wafer fabrication), 497–501
Fairchild camera, 4
Faraday cup, 354
feature size, 6–7
Federal Standard 209E, 97–99
FEOL, 315
ferroelectric materials, 40
field effect transistor (FET), 17, 524–531
field emission devices (FEDs), 550
field oxide (FOX), 162, 540
filar dimension measurement, 452
film parameters (layer deposition), 367–369
first-fail basis, 257
flat panel displays (FPDs), 548
flat, grinding, 63
flip-chip ball grid arrays (FCBGAs), 604
flip-chip bonding, 587
float zone technique, 56–60
flood guns, 350
focused ion beams, 289
four-point probe (resistance), 438–441
fringes, 445
front end of the line (FEOL), 16, 85, 120, 273
functional test, 599
furnaces (oxidation), 170–173
fuses (metallization), 415–416
fuses, 531–532

G

gallium arsenide, 144, 384, 395, 545
gallium, 37–38, 58
gas source MBE or GSMBE, 383
gas, contamination, 116–117
gate array, 560
gate oxide integrity (GOI), 121, 442
gate width, 6–7
Gauss, Karl Friedrich, 502–504
germanium, 37–39

gettering, 67, 397
global planarization, 310–311
gravity, specific, 44
green fabs, 501
grinding (wafer), 62–68
gross leak testing, 599
Grove, Robert, 420
growth defects, 61

H

hard bake, 211, 252–254
heteroepitaxial, 386
heterostructures/heterojunctions, 39
hexamethyldisilazane (HMDS), 220–221
high-aspect-ratio patterns, 368
high-density plasma CVD (HDPCVD), 381
high-k dielectrics, 30, 369, 399
high-pressure oxidation, 184
high-pressure water cleaning, 123
hi-rel, 570
histograms, 502
hole conduction, 33–35
hole flow, 36
homoepitaxial, 386
hoods, 100–101
horizontal conduction-convection-heated LPCVD, 378
horizontal conduction-heated APCVD, 376
horizontal laminar flow (HLF), 101
horizontal vertical-flow PECVD, 379
horizontal-tube induction-heated APCVD, 373–374
Horni transistor, 4–5
hot plates, 228–230
hybrid circuits, 607
hydrido-organic siloxane polymers (HOSP), 410
hydrogen ions, 46
hydrogen reduction, 53
hydrophilic, 218–219
hydrophobic, 218–219
hydroxyl ion, 165

I

image reversal (planarization), 317
image shearing dimension measurement, 452
immersion, 249
index of refraction, 216
induction (silicone dioxide), 162
ingots, 56
inorganic acids, 47
inorganic residue, 124
insulated field effect transistor (IFET), 19
insulators and dielectrics (CVD), 395
integrated circuits, 4
integrated device manufacturers (IDMs), 14
integrated image processing, 254–258
integrated-circuit formation, 533–537
integration level, 7–8
interconnects (*see* leads)
interdielectric layers (IDL), 315
intermediate metal dielectric (IMD), 409–410
intermetallic dielectric layer (IDL or IMD), 403
International Roadmap for Semiconductor Technology (IRST), 7, 10, 21, 22, 48, 52, 95, 145, 284, 410, 483, 570
intrinsic semiconductors, 31
inventory control (wafer fabrication), 506
inversion voltage (*see* threshold voltage)
ion beam etching, 271
ion implant
  doping, 326–329
  evaluation, 359
  implantation, 78–79, 345–355, 361–362
  masks, 354–355
ion milling (*see* sputtering)
ion pumps, 42, 430
ionization chamber, 347
ISO 9000, 508
ISO Global Cleanroom Standards (ISO 14644–2), 119–120
isotropic etching, 260

J

Josephson effect, 544
junction
  depth, 448–449

doping, 326–332
field-effect transistors (JFETS), 530
isolation, 534
transistor, 17–19
jungle, 176
just-in-time (JIT) inventory control, 507–508

K

kerf, 64
Kilby circuit, 4–5
killer defects, 92–93, 145, 151, 153, 462
kinematic viscosity, 215
known good die (KGD) problem, 608

L

laminar gas flow, 178
latch-up (CMOS epitaxy), 391
latent image, 246
lateral diffusion, 330
lattice, 54
layer thickness measurements (wafers), 443–445
lead on chip (LOC), 606
lead plating, 595
lead trimming, 596
leadless packs, 602
leads, 10, 402, 578
leakage current, 471
lift-off process (etching), 319320
light-sensitive polymers, 204–205
line edge roughness (LER), 286
line organization (wafer fabrication), 508–509
line width measurements (*see* critical dimensions)
linear stage (oxidation), 163
liquid encapsulated Czochralski (LEC), 58
lithography, 8, 243, 285–286
loading, wafer, 188–190
local oxidation of silicon (LOCOS), 185, 398, 537
logic circuits, 557–560
low-k dielectric materials, 30, 369, 399, 409–410
low-pressure chemical vapor deposition (LPCVD), 377–381, 414, 425–426

M

magnetron sputtering, 424
manometer, 45
manual spinners, 225
Marangoni drying, 134
mask defects, 145–146
masking, 82–83, 198–202, 278–280
mass analyzing/ion selection, 348
mass flow meter, 176–177
material safety data sheet, 48
matter, 42–45
maximum solid solubility (doping), 333–334
mechanical scanning, 353
melt, 57–58
memory circuits, 557, 560–566
memory MOS (MMOS), 530
mercury (Hg) spectrum, 210
metal cans, 603
metal film uses, 415–416
metal lines (*see* leads)
metal removal (MR), 198
metal semiconductor field-effect transistor (MESFET), 531–532
metallic ions, 94
metallization process, 402
metalorganic CVD (MOCVD), 373, 384–385
metal-oxide semiconductor (MOS)
structure, 7, 30
gate electrodes, 415
generally, 17, 185, 191, 533–534
LOCOS isolation, 538
MOSFET gate, 525, 532
silicone dioxide, 162
transistors, 473
metrology, 436
microelectromechanical systems (MEMS), 545
micro-environments, 103–104
microlithography, 198
microloading, 270
micromachines, 547
microprocessors, 557
microwave baking, 230
Miller indices, 55–56
mini-environments, 103–104
mini-fabs, 483
mobile ionic contaminants (MICs), 94

molecular beam epitaxy (MBE), 373, 383
momentum (ionization), 349
momentum transfer, 271, 420
monocrystalline, 54–55
Moore's law, 7, 243, 482
moving-arm dispense, 224
moving-belt infrared ovens, 230
MSI/bipolar integrated circuit, 87
multichip module (MCM), 16–17, 607
multilayer resist/surface imaging, 304
multimeter, 438
Murphy (wafer sort), 152–153

N

nanoanalysis era, 436
nanotechnology, 22–23
negative binomial (wafer), 152–153
negative resist development, 247
neutral beam trap (ionization), 351
next generation lithography (NGL), 285
nitridation, thermal, 194–195
nitrocellulose (NC), 299–300
Novolak, 205
Noyce, 4–5, 6
N-P junction, 329–330
NPN transistors, 523
N-type dopants, 328–332
N-type junction, 17
nucleation (layer deposition), 370
numerical aperture (NA), lens, 293–294

O

off-axis illumination, 294
Ohm's law, 438
ohmic, 405
oil diffusion pumps, 428–429
open-loop control, 183
optical critical dimension (OCD), 453
optical proximity corrected/correction (OPC), 297
optical resists, 204–207
optoelectronics, 546
organic acids, 47
organic residue, 124
out-diffusion, 390
ovens, convection, 228
overlay budget, 236, 280
overshoot, 173
oxidation
  automated, 190–191
  cycles, 191–193
  generally, 165–167
  process (wafer), 143
  wafer, 69
oxide impurities, 168
oxide removal (OR), 126, 198
oxide rupture (BVox), 442
oxygen passivation (OP), 117
ozonated chemistries, 127

P

packaging
  bare die strategies, 600
  design, 602
  functions and design, 572–578
  generally, 89
  marking, 597
  operations, 574–578
  parts, 578–590
  process flows, 600
  process, 590–597
  wafer, 69
pad mask, 84
pad oxide, 539
paddle, 188
pancake induction-heated APCVD, 375–376
parabolic stage (oxidation), 164
parallel downflow rinser, 130
parametric testing, 468
Pareto chart, 504–505
particle contaminants, 92–93
particulate removal, 122–123
passivation layer, 84
pattern shift, 389
patterning, 74, 198
pellicles, 298–300
periodic table, 27–28
Perkin Elmer light projection, 239
phase shift masks (PSMs), 296–297
phase shifting, 296–297
phenol-formaldehyde novolak resin structure, 205
phenolic organic strippers, 274
phosphorus silicate glass (PSG), 397

phosphorus-doped oxides, 168
photoacoustic, 448
photolithography, 198
photomasking, 198–202, 217–221
photoresist
  application (spinning), 221
  chemistry, 202–204
  development 246–252
  generally, 300
  performance, 207–214
  physical properties, 214–217
  process advances, 304
  removal, 121
photosolubilization, 205
physical measurement methods (wafers), 443
physical vapor deposition (PVD), 74, 366, 420
picked wafers, 188
pile-up
  dopants, 168
  drive-in oxidation, 343
pin grid arrays
pinholes (resist), 211
planar plasma etching, 267–268
planar technology, 4
planarity, 314–315
planarization, 11, 67, 303–304
planes, 18, 55–56
planetary wafer, 420
plasma descum, 251
plasma enhanced chemical vapor deposition (PECVD), 373, 379
plating process flows, 596
plug/plug filling, 408
PNP transistors, 523
point defects, 61, 145
point-of-use filters, 259
poisson (wafer sort), 152–153
polarization, 446–447
polishing
  pads, 313
  rates, 314
  wafers, 67–68
polycide-gate MOS, 528
polycrystalline, 54–55
polyimide planarization layers, 308–309
polymerized resist, 247–248
polymethylmethacrylate (PMMA), 305–306
polysilicon
deposition, 392
generally, 53
oxidation, 168
positive resists, 212–214, 247
post-CMP clean, 315
post-exposure bake (PEB), 241, 303
post-ion implant and plasma etch stripping, 276
postoxidation, 193–194
prebake (*see* hard bake)
pre-etch bake (*see* hard bake)
preseal inspection, 591–592
pressure, 45–46
priming, 220–221
printed circuit board (PCB), 601–602
process chemicals, 40–41
process control techniques (wafer), 144
process defects (wafer), 145
proportional band controllers, 173
proximity aligners, 237
proximity effect, 286–287
P-type dopants, 328–332
P-type junction, 17
puddle development, 251
pumps, vacuum, 426–431
pyrogenic steam oxidation, 165
Pyrox, 395

Q

quad packages, 605
quality control (ASQC), 508
quality control and certification—ISO 9000, 508
quartz and contamination, 118
quartz, 172–173

R

radiation hardened, 38
radiation, 227
rail guided vehicle (RGV), 496
ramp furnaces, 180
ramping, temperature, 174
rapid thermal oxidation (RTO), 182–184

rapid thermal processing (RTP), 180–182
Rayleigh constant, 293
RC constant, 409–410
RCA clean, 126–127
reaction chamber (oxidation), 172
reactive ion etching (RIE), 272
reduction steppers, 239
redundancy, 566
reflectance, 453
reflectometry (*see* spectrophotometers)
reflow (planarization), 317
refraction, 216
refractory metals, 407–409
registration capability, 233
reoxidation (reox) (*see* drive-in oxidation)
resist coverage, 223
resist light scattering, 300
resist stripping, 273
resistance (wafer electrical measurements), 437
resistivity, 30–33
resistors, 31, 469–470, 514–517
resists, optical, 204–207
resolution capability, 207, 233
reticle defects, 145–146
reticle, 82–83
retrograde wells, 361, 543
rinsing, 129–131
roughing pumps, 426
rupture voltage (*see* oxide rupture)

S

safe delivery system (SDS), 347
salicide-gate MOS, 529
same-type doping, 330
saw, wafer, 65
scaled transistor designs (*see* alternative transistor designs)
SCALPEL, 295–296
scanning auger microanalysis (SAM), 466
scanning capacitance microscopy (SCM), 451–452
scanning electron microscope (SEM), 450, 458
scanning projection aligners, 238
scatterometry, 463–465
scavenger (furnace), 175
Schottky barrier bipolar transistors, 524
scratch layer, 84
scumming, 251
seal ring, 593
sealing techniques, 592
secondary ion mass spectrometry (SIMS), 442, 450–451
seed deposition (metallization), 412
seeds (wafer sort), 152–153
selective epitaxial silicon, 391
selectivity (etching), 261
self-aligned silicon gate (SAG) structure, 321
self-aligned structure, 320
semiconducting compounds, 37
Semiconductor Equipment and Materials International (SEMI), 499, 508
semiconductor history, 1–4
semiconductor production materials, 14–16, 37
semiconductor silicon preparation, 52
semiconductor-device formation, 513–517
Semiconductor Industry Association (SIA) Roadmap [*see International Roadmap for Semiconductor Technology (IRST)*]
shape metrology, 453
sheet resistance, 439
side diffusion, 330, 345
silane source chemistry, 388
silicon dioxide (SiO2), 18, 395
silicon gate MOS, 526–528
silicon nitride, 397–398
silicon on insulator (SOI) isolation, 394–395, 544
silicon on sapphire (SOS), 394–395
silicon planar processing, 18
silicon tetrachloride source chemistry, 386–388
silicon wafers, 53
silicon, 37–39
silicone dioxide layer uses, 160–163
silicon-gate MOS technology, 83–87, 414
Silox, 395
silylation/DESIRE process, 307–308

single crystalline, 54–55
single-chamber planar PECVD, 380
single-wafer versus batch processing, 500–501
slip, 61–62, 389
slurry, 313
snow cleaning, 128
soft bake, 211, 219, 227–231
soft-contact machines, 237
solar cells, 546
solid state discrete devices, 4, 14
solids content (resist), 214
solid state integrated circuit, 557–560
solid state thermal diffusion, 330, 333
solvents, 46–47, 276
sonic-assisted cleaning/rinsing, 132
source cabinet (furnace), 175
space charge forces, 352
spectral response characteristic, 209
spectrophotometers, 446
spin priming, 220
spinning, 220–226
spot defects, 145
spray development, 250
spreading resistance probe (SRP), 450–451
sputter deposition/sputtering (PVD), 419–425
sputter etching, 271
sputtering, 74, 588
stack thickness (wafer), 467
stacking faults, 389
standard mechanical interface (SMIF), 103–104
standing-wave effect, 303
static dispense spin process, 222–223
static, 108–110
station yields, 141
statistical process control (SPC), 501–509
steam oxidation, 165–166
step and scan aligners, 241
step-and-repeat reduction system, 239
steppers, 238
strain gauges, 545
stripping, chemical, 274–276
stylus, 447–448
subcollectors, 391
subject contrast, 286–287
substrate reduction, 425
substrates, 69
subsurface reflectivity, 300
suckback (spinners), 225
superconductors, 544–550
surface concentration, 334–335
surface dielectric (silicone dioxide), 161–162
surface passivation (silicone dioxide), 160
surface profileometers (*see* stylus)
surface roughness, 120
surface tension, 134, 215
surface topography, 294
SWAMI, 539
switching, 2
synchrotron radiation, 291

T

tape automated bonding (TAB), 576, 591
target chamber (ionization), 353
TC bonding (*see* thermocompression bonding)
teardrop transistor, 4–5
temperature control (oxidation), 173
temperature cycling, 598
temperature scales, 43–44
temperature sensing, 547
test wafers, 436
testing, 598–600
tetraethyl orthosilicate (TEOS), 396
tetramethylammonium hydroxide (TMAH), 248
thermal budget, 181
thermal diffusion, 78–79
thermal flow (masking), 211
thermal nitridation, 194–195
thermal oxidation mechanisms, 160, 163–180
thermocompression bonding, 585
thermode, 591
thermosonic, 585
thin packages, 605
third optical inspection, 592
threshold voltage, 361, 475
tie-bar, 596

time of flight secondary ion mass spectrometry (TOF–SIMS), 467
top surface imaged (TSI), 308
Torricelli, 45
transfer resistor (*see* transistor)
transistor operational analogy, 521
transistor, 3, 521–532
transmission electron microscope (TEM), 460
transport-limited reaction, 164
trapping, 67
trench isolation, 540
triboelectric charging, 109
trichlorosilane, 53
trilevel resist process, 307
tube furnaces, 178–179
tungsten, 425
tunnel/bay concept, 102
tunneling, 544
turbomolecular pumps, 431
twinning, 61

U

ULSI/VLSI patterning, 284
ultrahigh vacuum CVD (UHV/CVD), 379
ultraviolet and visible spectrum, 209
undercutting, 260
underfillment, 590
underpass conductors, 532
unit cell, 53
unpolymerized resist, 247–248
unyielded die cost, 488–490

V

vacuum
  baking, 231
  evaporation, 417–419
  generally, 45–46
  pumps, 426–431
  tube, 2
van der Pauw structure, 359–360
van der Waals force, 122
vapor density, 45
vapor phase decomposition/atomic absorption spectroscopy (VPD–AAS), 467
vapor phase epitaxy (VPE), 373, 382
vapor priming, 220–222
Vapox, 395
vertical laminar flow (VLF) station, 100
vias, 403
virtual leak, 423
viscosity, 214, 217
VLSI circuit fabrication, 417
VLSI/ULSI patterning, 284

W

wafer
  assembly and test of, 155
  breakage, 144
  charging, 350
  doping, 75–76, 167
  electrical measurements, 437–442
  fabrication, 14, 72–83, 484–491
  generally, 16, 52–53
  isolation technology (WIT), 103–104
  layering, 74–76, 77
  marking, 65–66
  orientation, 167
  overall process yields, 155–156
  patterning, 74–76, 78
  planes, 73
  preparation, 62
  priming, 219–220
  quality, 59–60
  scrubbers, 123
  size, 22, 147–150, 590
  slicing, 64
  sort, 16, 72, 87–89
  throughput, 233
  warping, 144, 174
wafer-cleaning station, 187
wafer-delivery automation, 496
wafer-loading, 187, 494–495
wafers in process (WIP), 506–507
wafer-sort yield factor, 146–149, 152–154;
water, contamination, 113–115
wet development methods, 249
wet etching, 261–264
wet oxidation, 165
wet process/cleaning, 124
wire bonding, 584–589
work function, 526

X
X-ray aligners, 241, 289–291

Y
yield limiters, 142
yield measurement point, 140–143
yielded die cost, 488–490

Z
zeta potential, 122

THE #1 BOOK IN THE INDUSTRY —
COMPLETELY REVISED AND UPDATED

A perfect introduction to the industry that drives high tech, ***Microchip Fabrication*** offers a low-math, straight-talk approach to the entire process of semiconductor processing — from raw materials through shipping the finished, packaged device. With lots of detailed illustrations and analogies to everyday life, this is the industry's most novice-friendly text!

Used for training, teaching, and vocational-technical programs, ***Microchip Fabrication*** covers every stage of semiconductor processing, from raw material preparation to packaging, testing, and traditional and state-of-the-art processes. Each chapter contains quizzes and review summaries buttressed by an extensive glossary.

By the time you've finished ***Microchip Fabrication***, you'll have a solid working knowledge of the important issues, processes, materials, and methods involved in semiconductor technology, whether on the subatomic level or in the context of large-scale industrial practices.

NEW TO THIS EDITION

- **Nanotechnology**
- **300-mm Wafer Processing**
- **Next Generation Lithography**
- **"Green" Processes and Devices**
- **New Fabrication Advances**
- **More!**

" ... An excellent reference for anyone who wishes to know more about the technical side of semiconductor manufacturing. It is well-indexed, and the chapters are self-contained enough to make it useful for catching up on one topic at a time."

— *FabTime*, on the previous edition

The McGraw-Hill Companies

Visit McGraw-Hill Education **www.books.mcgraw-hill.com**

ISBN 0-07-143241-8

9 780071 432412 56995

6 39785 50772 7

Cover design:
Mary McDonnell

Electronics Engineering