Chapter

1

Development and Fabrication of IC Chips

Charles Cohn
Agere Systems
Allentown, Pennsylvania

1.1 Introduction

At the end of the nineteenth century, the consumer products of that time included simple electrical circuits for lighting, heating, telephones, and telegraph. But the invention of radios and the need for electrical components that could rectify and amplify signals spurred the development of vacuum tubes. Vacuum tubes were found in products such as radios, televisions, communication equipment, and in early computers. Their use lasted until the late 1960s, when the development of semiconductor devices ushered in a new era in electronics. The semiconductor, containing an array of complex transistors and other components on a single IC chip, provided improved reliability and reduced power, size, and weight, and it made possible today's sophisticated electronic products.

This chapter, which is subdivided into five sections, presents a simplified approach to the understanding of the fundamentals of semiconductors, IC development, and IC chip fabrication. The topics cover

- Atomic structure
- Vacuum tubes
- Semiconductor theory
- Fundamentals of integrated circuits
- IC chip fabrication

1.2 Atomic Structure

All matter, whether solid, liquid, or gas, is composed of one or more of the 109 presently recognized elements referenced in the periodic table (Fig. 1.1). Of these, 91 elements occur naturally, and the rest are either man-made or are by-products of other elements. An element is composed of molecules, which are divisible into even smaller particles called atoms. The atomic structure for each element is unique and defines the element's properties.

Materials can be categorized according to the way they conduct electricity when a voltage is applied across them. Insulators, as the name implies, do not conduct electricity, whereas conductors allow a large flow of current, depending on the voltage applied and the conductance properties of the material. Semiconductors have properties in between those of resistors and conductors, having limited current flow capabilities that depend on their atomic structure, the purity of the material, and temperature.

The structure of an atom, as was first proposed by Neils Bohr in 1913 and later supported by extensive experimental evidence, consists of negatively charged electrons rotating in somewhat defined orbits, or energy levels, about a highly dense nucleus consisting of protons and neutrons (Fig. 1.2). The protons are positively charged, and the neutrons have no charge, or are electrically neutral. Each atom has an equal number of (+) protons and (–) electrons, but the number of neutrons may vary.

Each element in the periodic table is assigned an atomic number, which is equal to the number of protons, and therefore electrons, contained in its atom. The atomic number is shown in the upper part of the box representing the element (Fig. 1.1).

The actual weight of an atom is extremely small, which makes it very difficult to work with. As a result, a weight scale was devised that assigns weights to atoms that show their weights relative to one another. The weights assigned are based on the densest part of the atom; namely, the sum of the number of protons and neutrons in the nucleus.

The positively charged protons exert an inward force on the negatively charged electrons, which is balanced by an outward centrifugal force created by the electrons spinning in their orbits around the nucleus. Thus, the two opposing forces provide a balanced structure for the atom.

The maximum number of electrons that a given orbit or shell can support is governed by the $2n^2$ rule, where "n" is the shell number.[6] That is, shell #1 (closest to the nucleus) can hold a maximum of two electrons, shell #2 can have a maximum of 8 electrons, and so on. If the number of electrons for a given shell exceeds the maximum indicated by the $2n^2$ rule, then the extra

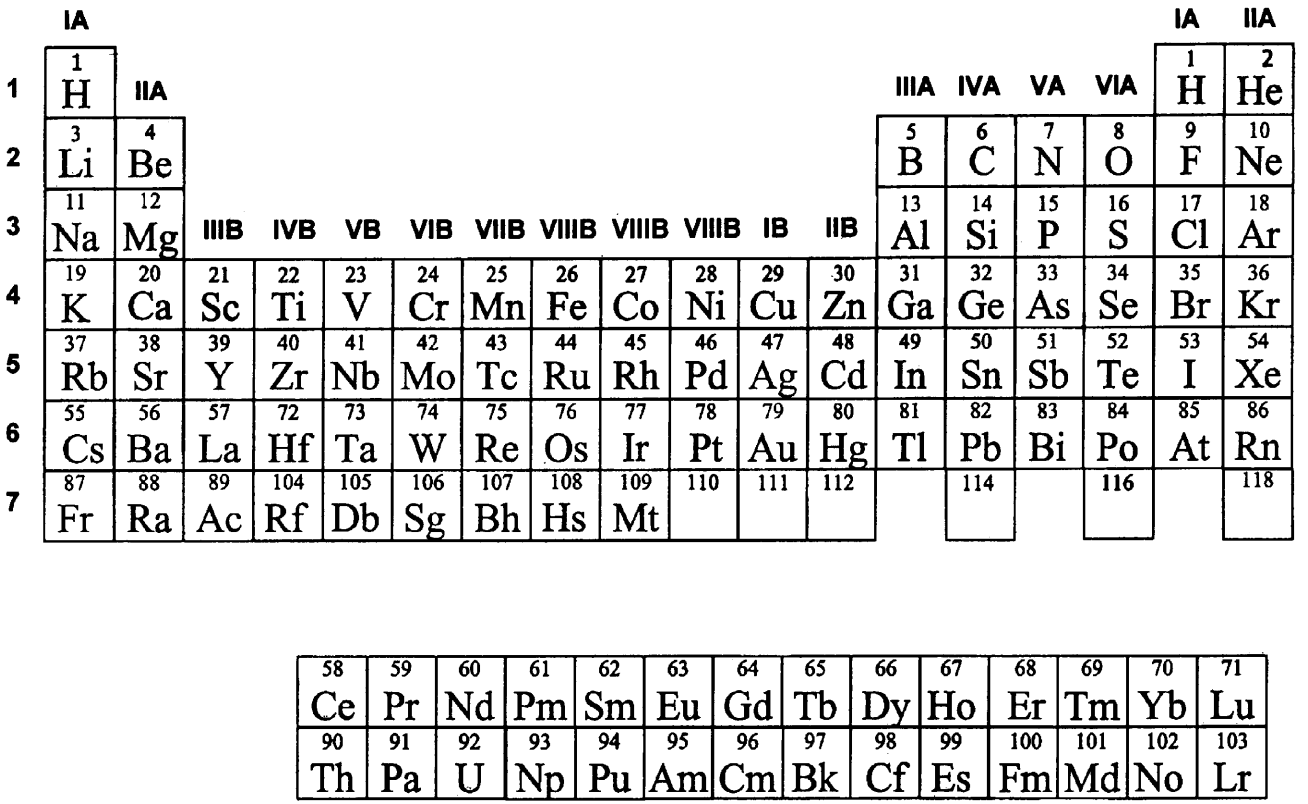

Figure 1.1 Abbreviated periodic table of the elements.

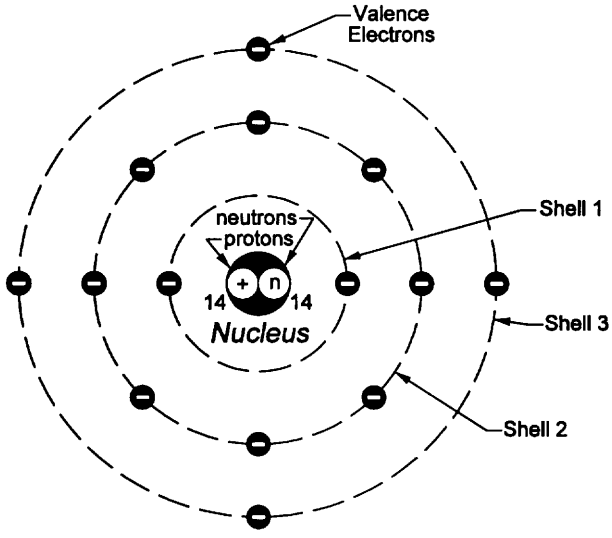

Figure 1.2 Bohr model of silicon atom.

electrons are being forced into the next higher shell. An atom is chemically stable if its outer shell is either completely filled with electrons, based on the $2n^2$ rule, or has eight electrons in it. The electrons in the outer shell are called valence electrons and, if their number is less than eight, the atom will have a tendency to interact with other atoms either by losing, acquiring, or merging its electrons with other atoms.

In the periodic table (Fig. 1.1), elements with the same number of valence electrons have similar properties and are placed in the same group. For example, elements in Group I have atoms with one electron in their outer shell. Group II shows elements that have atoms with two electrons in their outer shell, and so on. Elements on the left side of the periodic table have a tendency to lose their valence electrons to other atoms, thus becoming electropositive. The elements on the right side of the periodic table show a tendency to acquire electrons from other atoms and become electronegative.

The type of interaction occurring between atoms, as they are brought together, depends largely on the properties of the atoms themselves. The interaction may form bonds that can be classified as ionic, covalent, molecular, hydrogen bonded, or metallic. Since this chapter is concerned with semiconductors, which tend to form covalent bonds with other elements and with themselves, the emphasis will be on covalent bonding.

Covalent bonds occur when two or more atoms jointly share each other's valence electrons. If the outer shell is partially filled with electrons, the atom will be attracted to other atoms also having a deficiency of electrons, so sharing each other's valence electrons will result in a more stable condition. As an example, two chlorine atoms will attract and share each other's single electron to form a stable covalent bond with eight electrons in each shell (Fig. 1.3).

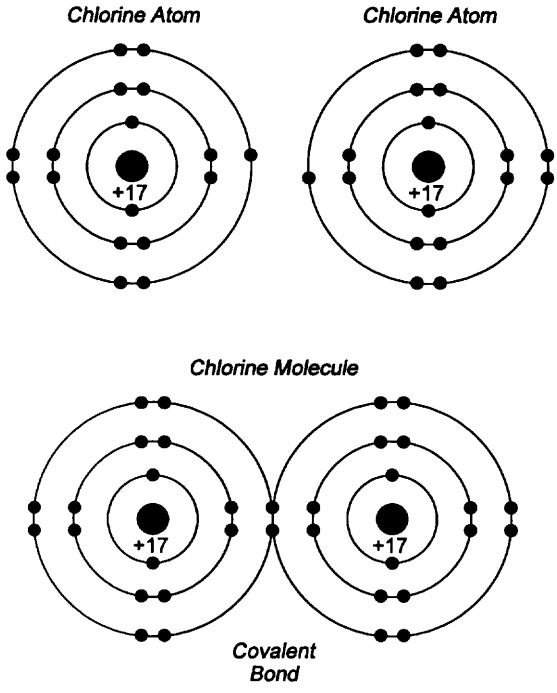

Figure 1.3 A chlorine molecule forms a covalent bond.

1.3 Vacuum Tubes

Modern electronics can trace its roots to the first electronic devices called vacuum tubes. Although, today, solid state devices have totally replaced the vacuum tube, the fundamental principle as to its usage remains relatively unchanged. For more than 40 years, until the late 1960s, the most important part in a consumer electronics product was the vacuum tube. It is with this historical perspective in mind that this section is presented so that readers will not lose sight of where it all started.

The vacuum tube got its start in 1883, when Edison was developing the incandescent lamp. To correct the premature burnout of the red-hot filament in light bulbs, Edison tried a number of experiments, one of which was to place a metal plate sealed inside a bulb and connect it to a battery and ammeter, as shown in Fig. 1.4. Edison observed that, when the filament was hot and the plate was positively (+) charged by the battery, the ammeter indicated a current flow through the vacuum, across the gap between the filament and the plate. When the charge on the plate was reversed to negative (–), the current flow stopped. As interesting as this phenomena was, it did not improve the life of Edison's lamps and, as a result, he lost interest in this experiment and went on to other bulb modifications that proved more successful. For about 20 years, Edison's vacuum tube experiment remained a scientific curiosity. In 1903, as radios were coming into use, J. A. Fleming, in England, found just

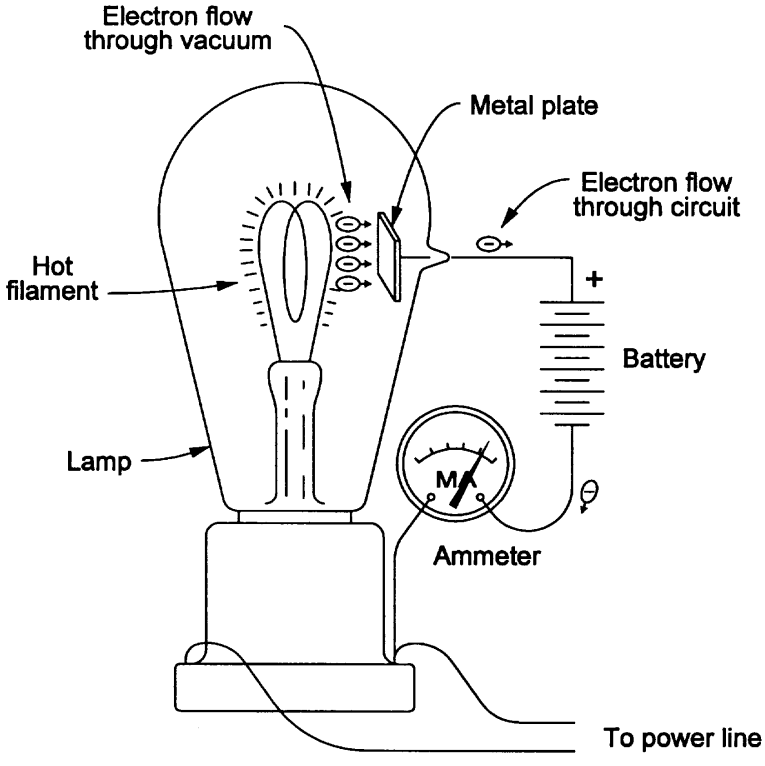

Figure 1.4 Edison's vacuum tube.

what he needed to rectify alternating radio signals into DC signals required to operate headphones. By hooking up Edison's vacuum tube to a receiving antenna, the tube worked like a diode. When the signal voltage increased in one direction, it made the plate positive (+), and the signal got through. When the signal voltage increased in the other direction of the AC cycle, applying a negative (–) charge to the plate, the signal stopped.

The vacuum tube, also called the electron tube, required a source of electrons to function. In Edison's original electron tube, the electron source, called the *cathode,* was the filament that, when heated red-hot, emitted electrons that flew off into the vacuum toward the positively charged plate, called the *anode.* The effect of heating the cathode to activate the electrons was called *thermionic*. Other electron tubes used high voltage to pull the electrons out of a cold cathode. Electronic emission also occurred by applying light energy to a photosensitive cathode. Tubes using this effect were called *photoelectronic* vacuum tubes. Although a variety of methods existed to remove electrons from the cathode, the thermionic vacuum tubes were the most widely used. The cathode was either heated by resistors within or used a separate source of power for heating. The vacuum tube consisted of a glass or metal enclosure with electrode leads brought out through the glass to metal pins molded into a plastic base (Fig. 1.5).

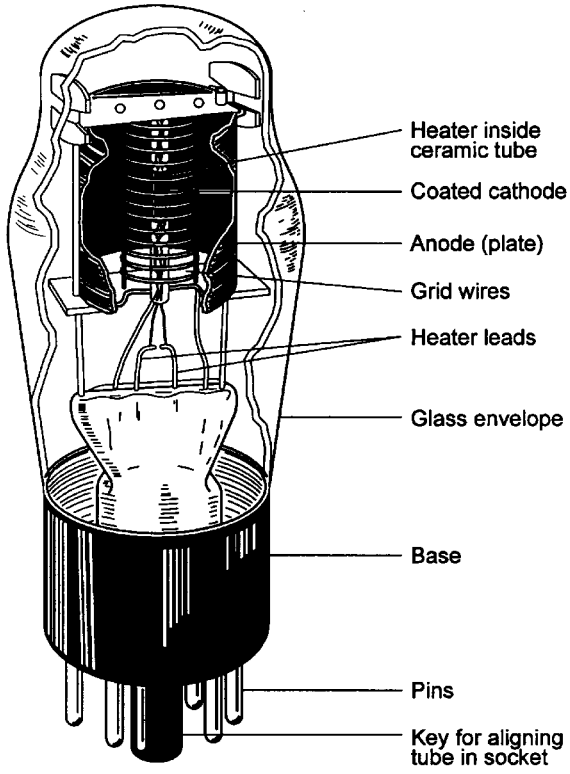

Figure 1.5 The construction of a triode vacuum tube.

When the electron tube contains two electrodes (anode and cathode), the circuit is called a *diode*. In 1906, Lee DeForest, an American inventor, introduced a grid (a fine wire mesh) in between the cathode and the anode. The addition of a third electrode expanded the application of electron tubes to other electronic functions. The grid provided a way of controlling the flow of electrons from the cathode to the plate (anode). Even though the grid had a weak positive or negative charge, its proximity to the cathode had a strong effect on the flow of electrons from cathode to plate. The open weave in the grid allowed most of the electrons to pass through and land on the stronger positively charged anode. When the grid was negatively charged, it repelled the electrons from the cathode, stopping the current flow (Fig. 1.6).

Thus, with the three electrodes (i.e., cathode, anode, and grid), it was possible to both rectify and amplify weak radio signals using one tube. The three-electrode vacuum tube was called a *triode*. Additional electrodes, such as a suppressor grid and screen grid, were also enclosed in electron tubes, making it possible to expand the functions of electron tubes.

Vacuum tubes, although widely used in the industry for a half a century, had a number of disadvantages, among them that they were bulky, generated a lot of heat, and were subject to frequent replacement because they would burn out. With the advent of solid state devices, which had none of the disad-

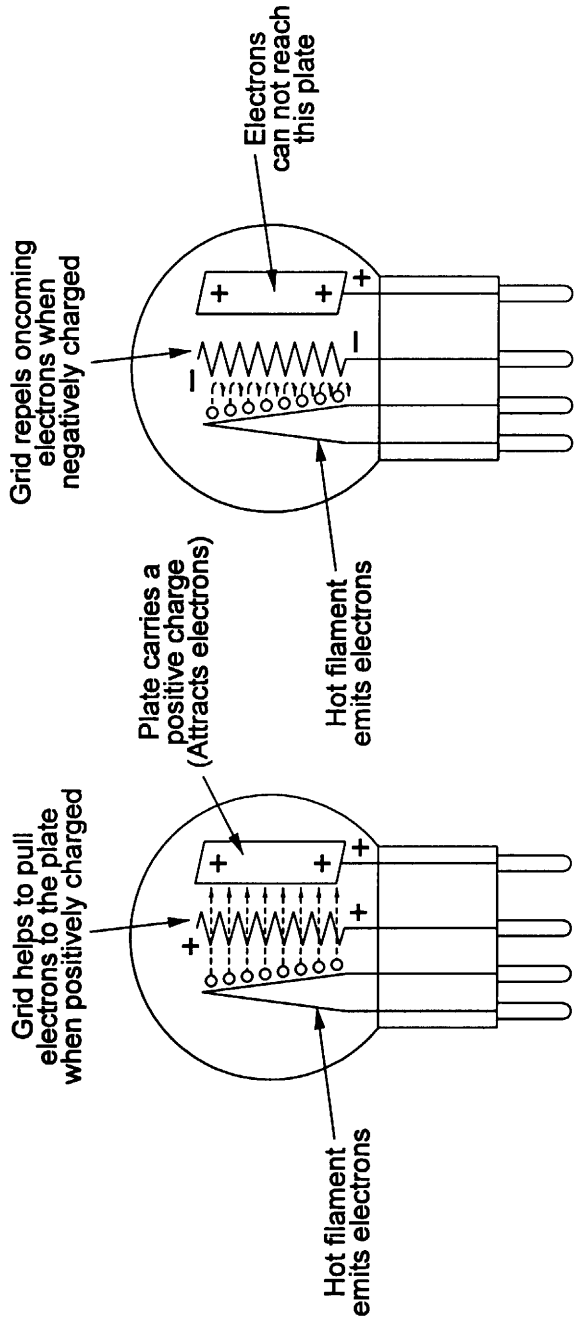

Figure 1.6 Grid controls the flow of electrons to the plate of a triode.

vantages of vacuum tubes, vacuum tubes started to fade from use in electronic products.

1.4 Semiconductor Theory

Semiconductor materials have physical characteristics that are totally different from those of metals. Whereas metals conduct electricity at all temperatures, semiconductors conduct well at some temperatures and poorly at others.

In the preceding section, it was shown that semiconductors are covalent solids. That is, the atoms form covalent bonds with themselves, the most important being silicon and germanium in Group IV of the periodic table (Fig. 1.1). Others may form semiconductor compounds where two or more elements form covalent bonds, such as gallium (Group III) and arsenic (Group V), which combine to form gallium arsenide.

Typical semiconductor materials used in the fabrication of IC chips are

- Elemental semiconductors
  - Silicon
  - Germanium
  - Selenium
- Semiconducting compounds
  - Gallium arsenide (GaAs)
  - Gallium arsenide–phosphide (FaAsP)
  - Indium phosphide (InP)

Germanium is an elemental semiconductor that was used to fabricate the first transistors and solid state devices. But, because it is difficult to process and inhibits device performance, it is rarely used now.

The other elemental semiconductor, silicon, is used in approximately 90 percent of the chips fabricated. Silicon's popularity can be attributed to its abundance in nature and retention of good electrical properties, even at high temperatures. In addition, its silicon dioxide (SiO_2) has many properties ideally suited to IC manufacturing.

Gallium arsenide is classified as a semiconducting compound. Some of its properties, such as faster operating frequencies (two to three times faster than silicon), low heat dissipation, resistance to radiation, and minimal leakage between adjacent components, makes GaAs an important semiconductor for use in high-performance applications. Its drawbacks are the difficulty of growing the ingots and fabricating the ICs.

An elemental or compound semiconductor that was not contaminated by the introduction of impurities is called an *intrinsic semiconductor*. At an absolute zero temperature, intrinsic semiconductors form stable covalent bonds that have valence shells completely filled with electrons. These covalent bonds are very strong, so that each electron is held very strongly to the atoms sharing it. Thus, there are no free electrons available, and no electrical conduction is pos-

sible. As the temperature is raised to relatively high temperatures, the valence bonds sometimes break, and electrons are released. The free electrons behave in the same way as free electrons in a metal; therefore, electrical conduction is now possible when an electric field is applied.

If an impurity, such as phosphorus or boron, is introduced into the crystal structure of an intrinsic semiconductor, its chemical state is altered to where the semiconductor will have an excess or deficiency of electrons, depending on the impurity type used. The process of adding a small quantity of impurities to an intrinsic semiconductor is called *doping*. As an example, consider an intrinsic silicon crystal structure with its covalent bonds, shown as a two-dimensional sketch in Fig. 1.7. Each atom is surrounded by four other atoms, with which it shares one pair of electrons, to form four covalent bonds. If the silicon crystal (Group IV) is doped with a controlled quantity of an impurity (dopant), such as phosphorus (Group V), the newly formed covalent bonds (Fig. 1.8) have an excess of electrons that are free to move from atom to atom when a voltage is applied across the semiconductor. The material thus altered is called an n-type (n for negative) semiconductor. Another semiconductor type, called p-type (p for positive), can be formed by doping the silicon crystal with a dopant from Group III, such as boron. The resultant combination (Fig. 1.9) has a deficiency of electrons and thus creates "holes," or electron vacancies, in the positively charged atoms. A single semiconductor crystal structure can be selectively doped with two different kinds of impurities that will form adjacent p-type and n-type semiconductors (Fig. 1.10). The transition between the two types of semiconductors is the p-n junction and is where electrons and holes recombine. As the electrons enter the p-type region, filling the holes, the atoms become negatively charged while the atoms left behind, with fewer elec-

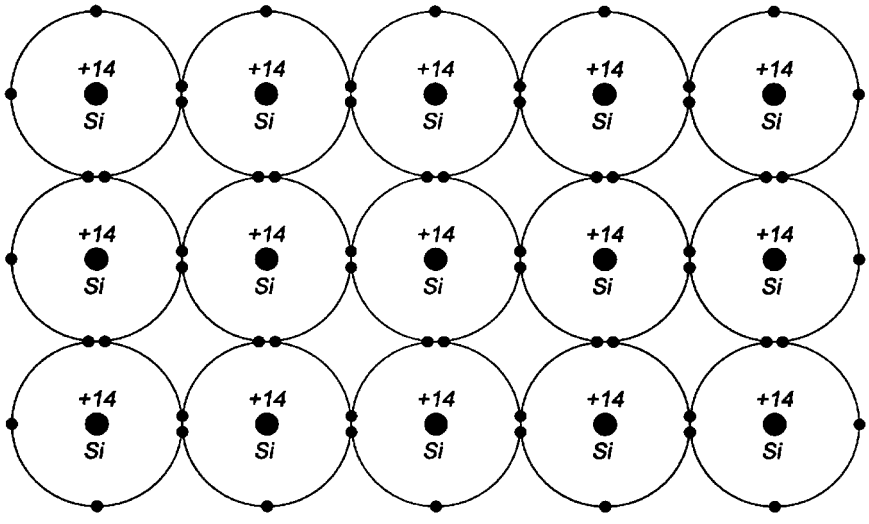

Figure 1.7 Two-dimensional representation of an intrinsic silicon crystal (only valence electrons are shown).

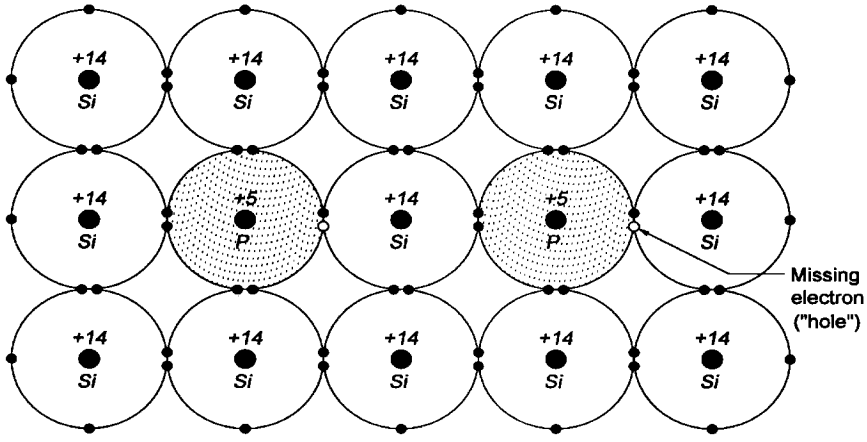

Figure 1.8 Two-dimensional representation of silicon crystal doped with phosphorus to create a p-type semiconductor (only valence electrons are shown).

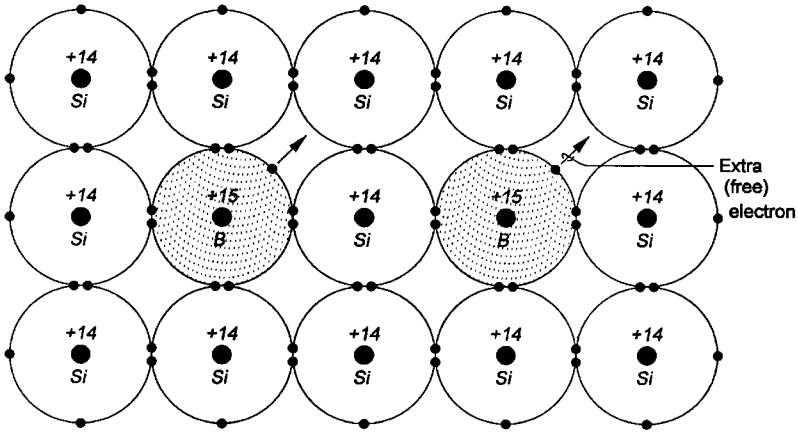

Figure 1.9 Two-dimensional representation of silicon crystal doped with boron to create an n-type semiconductor (only valence electrons are shown).

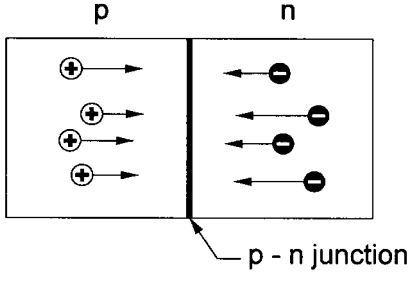

Figure 1.10 P-type/n-type semiconductor junction. *(After Tedeschi.[1])*

trons, and new holes, become positively charged (Fig. 1.11). The process can be considered as a flow of holes or a current flow of positively charged vacancies, which is opposite to the electron flow. Since there is a depletion of electrons and holes in the contact region, the p-n junction is referred to as the *depletion region*. The double layer of charged atoms sets up an electric field across the contact that prevents further intermixing of electrons and holes in the region, creating a barrier.[1]

1.4.1 The diode

When an external battery is placed across the p-n junction, with the positive (+) terminal of the battery connected to the n-type side of the semiconductor and the negative (–) terminal connected to the p-type side, a so-called reverse bias condition is created across the junction. As the electrons are attracted to the positive terminal of the battery, and the holes are attracted to the negative side, the electrons and holes move away from the junction, thus increasing the depletion region and preventing current flow (Fig. 1.12).

If the battery terminals are reversed (Fig. 1.13), the electrons in the n-material and the holes in the p-material are repelled by their respective negative and positive potentials of the battery and move toward the junction. This reduces the barrier junction, allowing electrons and holes to cross the junction and continue to recombine. As the electrons and holes recombine, new electrons from the (–) terminal of the battery enter the n-region to replace the electrons that crossed into the p-region. Similarly, the electrons in the p-region are attracted by the (+) terminal, leaving new holes behind, which are filled by electrons coming from the n-region. The continuous recombining process creates a forward current flow across the p-n region, which is referred to as *forward biased*. Thus, a p-n junction acts as a diode (rectifier); i.e., when the junction is forward biased, it conducts current, and when the bias is reversed, the current stops.

1.4.2 The junction-type bipolar transistor

Combining two or more p-n junction arrangements (p-n-p, n-p-n, etc.) into one device resulted in the development of the transistor. The transistor is a device

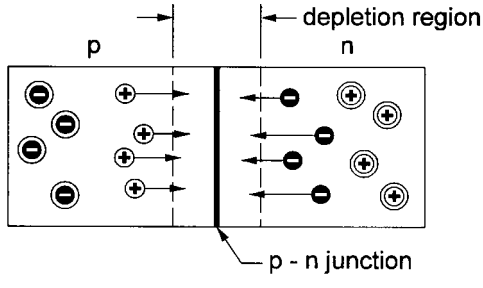

Figure 1.11 P-type/n-type semiconductor junction with depletion region. *(After Tedeschi.[1])*

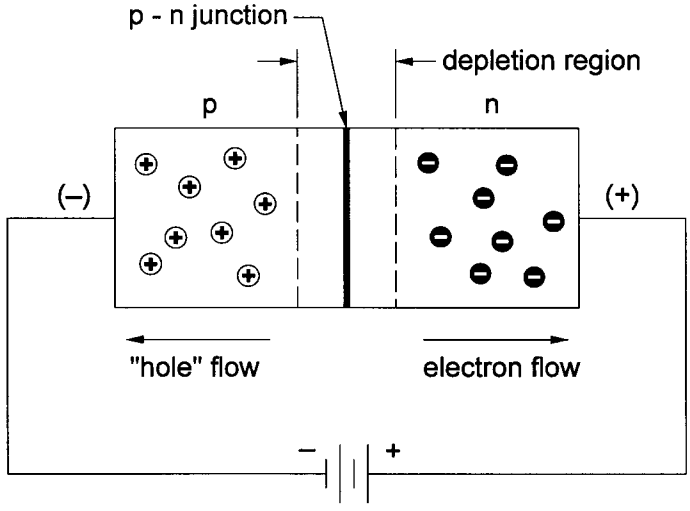

Figure 1.12 Reverse-biased p-n junction. *(After Tedeschi.[1])*

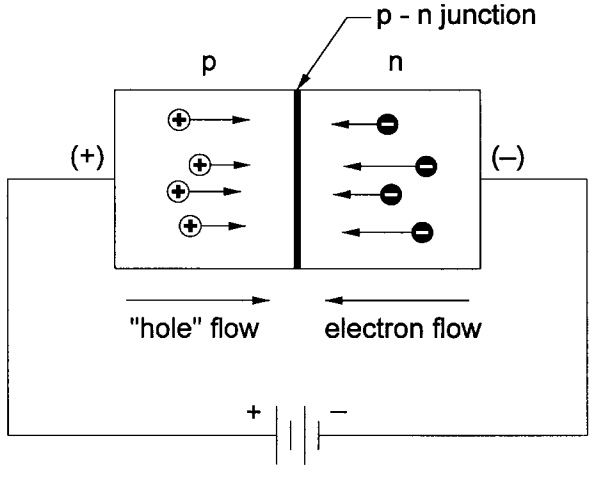

Figure 1.13 Forward-biased p-n junction. *(After Tedeschi.[1])*

capable of amplifying a signal or switching a current *on* and *off* billions of times per second. Its development dawned a new age in electronics.

Since its inception in 1948 by W. Shockley, J. Bardeen, and W. Brattain of Bell Laboratories, the transistor has evolved into many forms. The original device (Fig. 1.14) used point contacts to penetrate the body of a germanium semiconductor. Subsequent transistors were of the junction (bipolar) type with germanium as the semiconductor. The semiconductor material was later replaced with silicon.

To illustrate how a bipolar transistor works, an n-p-n semiconductor configuration (Fig. 1.15) is used as an example. In this structure, a very thin, lightly

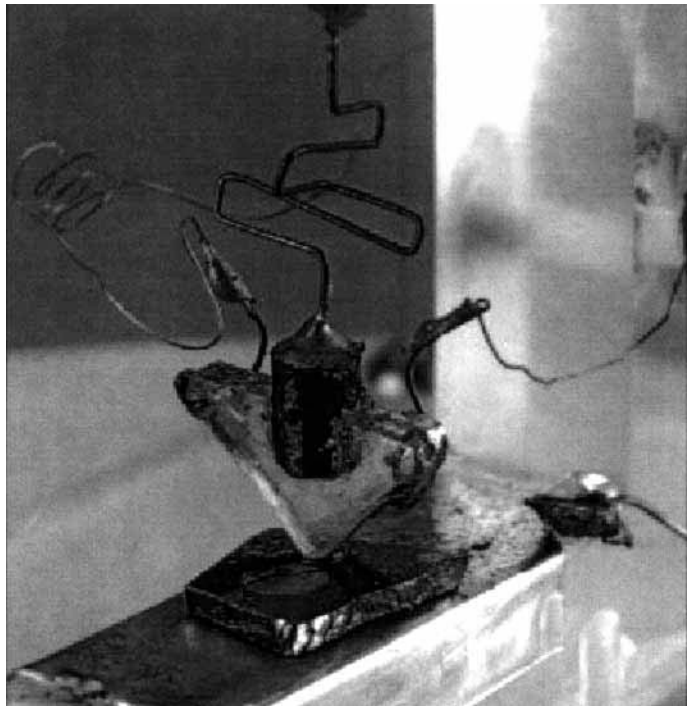

Figure 1.14 The original point-contact transistor. *(Courtesy of Bell Laboratories.)*

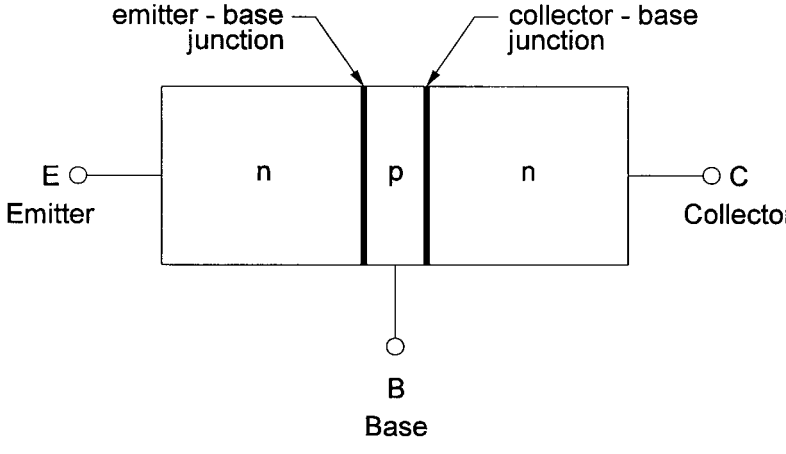

Figure 1.15 Typical n-p-n transistor.

doped p-region, called the base (B), is sandwiched between two thicker outer n-regions, called the emitter (E) and collector (C). The emitter generates electrons, the collector absorbs the electrons, and an input signal applied at the base controls the electron flow from emitter to collector.

Figure 1.16 shows a typical circuit of a bipolar transistor functioning as a digital switch. A supply voltage V_{CE} is applied across the emitter and collector terminals, with the (+) positive terminal of the voltage source connected through a load resistor R_L to the collector terminal. Applying a positive voltage between the base and emitter terminals, $V_{BE} > 0.5$ V, turns the transistor *on*. Since the emitter-base junction is forward biased, the electrons in the emitter region will cross the junction and enter the base region where a few of the electrons will recombine with holes in the lightly doped base. Because the base region is very thin, and the free electrons are close to the collector, the electrons are pulled across the collector-base junction by the positive potential of the collector and continue to flow through the external circuit. Decreasing the input voltage to zero no longer sustains a flow of electrons across the emitter-base junction and the transistor is turned *off*.

When the bipolar transistor is used as an amplifier, the strength of the emitter-to-collector current flow follows the variations in strength of the input voltage, but at a magnified level. That is, increasing the strength of the input voltage at the base causes proportionally more electrons to cross the emitter-base junction, thus increasing the current flow between the emitter and collector. Decreasing the input voltage causes the electrons to reduce their speed of

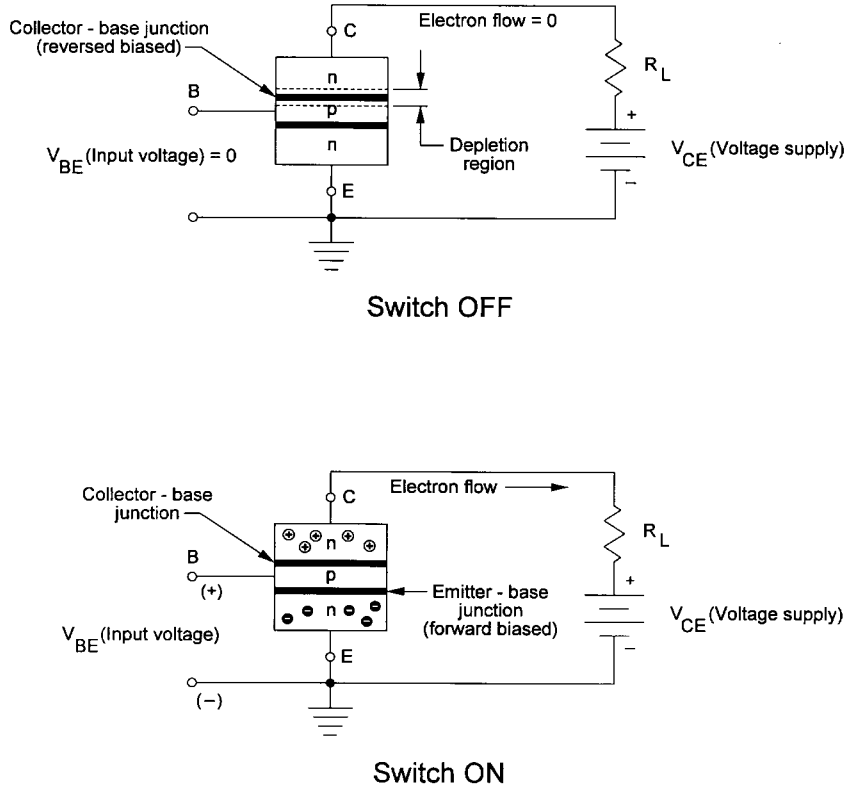

Figure 1.16 Bipolar transistor functioning as a digital switch. *(After Levine.[2])*

crossing the emitter-base junction, and the current flow decreases. Since the bipolar transistor can equally amplify both current and voltage, the transistor can also be considered a power amplifier.

The characteristic of the bipolar transistor is its high-frequency response capability, which equates with high switching speed. But to achieve high switching speeds, the transistor must operate at high emitter-to-collector current flow, causing increased power losses.[2]

1.4.3 The field-effect transistor (FET)

The FET transistor operates on a different principle from that of the bipolar transistor. The input voltage creates an electric field that changes the resistance of the output region, thus controlling the current flow. Its unique characteristic of having a very high input resistance will prevent a preceding device in the circuit from being loaded down, which could degrade its performance. The working principle of the FET transistor was known long before the bipolar transistor was developed, but, because of production difficulties, it was abandoned in favor of the bipolar transistor. The 1960s saw a revival of interest in FET transistors after the earlier production issues were resolved. The FET transistor has three semiconductor regions, similar to the bipolar transistor, but, because its principle of operation is different, the FET regions are called the *source,* the *drain,* and the *gate*. These regions are equivalent to the emitter, collector, and base of the bipolar transistor. If we again consider an n-p-n structure, the source and the drain regions are n-type semiconductors, and the gate region is a p-type material.

There are two types of FET transistors: the junction field-effect transistor (JFET) and the metal oxide semiconductor field-effect transistor (MOSFET).

1.4.4 The junction field-effect transistor (JFET)

In a junction field-effect transistor (JFET), the electrons do not cross the p-n junction but, rather, flow from the source to the drain along a so-called n-channel, which is formed between two p-type materials (Fig. 1.17). The n-channel is considered the output section of the transistor, and the gate-to-source p-n junction is the input section. In a typical JFET circuit (Fig. 1.18), where the transistor functions as a digital switch, the voltage supply V_{SD} is applied

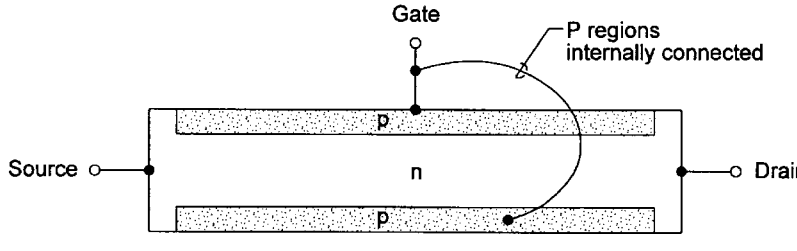

Figure 1.17 Junction field-effect transistor (JFET) construction. *(After Levine.[2])*

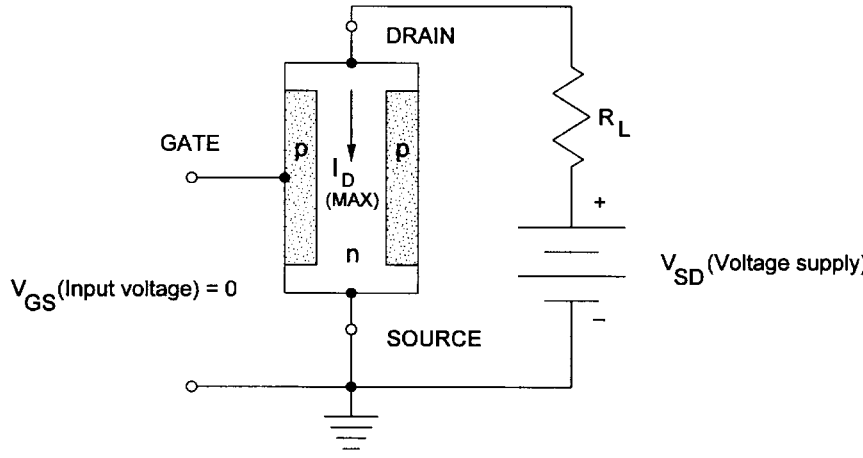

Figure 1.18 JFET functioning as an "on" switch, p-n junction forward biased. *(After Levine.[2])*

across the (–) source and the (+) drain terminals, through a load resistor R_L. The input voltage V_{GS} is connected between the gate and source terminals with the negative polarity on the gate. With a reversed bias input voltage, the effect of the electric field creates depletion areas around the two p-n junctions, which are characteristically devoid of electrons. As the input voltage increases, the depletion areas penetrate deeper toward the center of the channel, restricting the electron flow between the source and the drain (Fig. 1.19). If the input voltage is large enough, the depletion areas will totally fill the n-channel, choking off the flow of electrons. Reducing the input voltage V_{GS} to zero, the depletion areas disappear, and the n-p channel is wide open, with

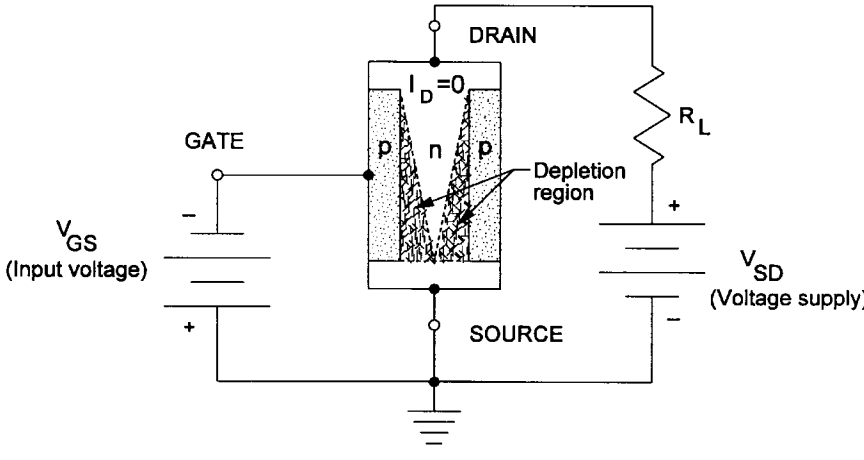

Figure 1.19 JFET functioning as an "off" switch, p-n junction reverse biased. *(After Levine.[2])*

very low resistance; thus, the electron flow rate will be at its maximum. When the JFET transistor is used as a linear amplifier, the input voltage variation will have an equivalent effect on the current flow in the n-channel and cause an output voltage gain across the source and drain terminals.[2]

1.4.5 The metal-oxide semiconductor field-effect transistor (MOSFET)

Another type of FET transistor is the metal-oxide semiconductor field-effect transistor (MOSFET). It operates on the same principle as the JFET transistor but uses the input voltage, applied across a built-in capacitor, to control the source-to-drain electron flow.

A MOSFET typically consists of a source and drain (n-type regions) embedded in a p-type material (Fig. 1.20). The gate terminal is connected to a metal (aluminum) layer that is separated from the p-type material by a silicon dioxide (SiO_2) insulator. This combination of metal, silicon dioxide (insulation), and p-type semiconductor layers forms a decoupling capacitor. The gate region is located between the source and drain regions, with a fourth region located under the gate, called the *substrate*. The substrate is either internally connected to the source or is used as an external terminal.

The flow of electrons from the source to the drain is controlled by whether the gate has a positive or negative voltage. If the input voltage applied to the gate is positive, free electrons will be attracted from the n-regions and the p-region to the underside of the silicon dioxide layer, at the gate region. The abundance of electrons under the gate forms an n-channel between the two n-regions, thus providing a conductive path for the current to flow from the source to the drain (Fig. 1.21). In this case, the MOSFET is said to be *on*. If the input voltage at the gate is negative, the electrons in the p-region under the gate are repelled, and no n-channel is formed. Since the resistance in the p-region between the two n-regions is infinite, no current will flow, thus turning the MOSFET *off*. Although the MOSFET used in the above description was of an n-p-n type, a p-n-p type MOSFET can also be constructed, but its voltage polarities are reversed.[2]

1.4.6 The CMOSFET transistor

When two MOSFET transistors, one an n-p-n type and the other a p-n-p type, are connected, the combination (Fig. 1.22) is called a complementary MOS-

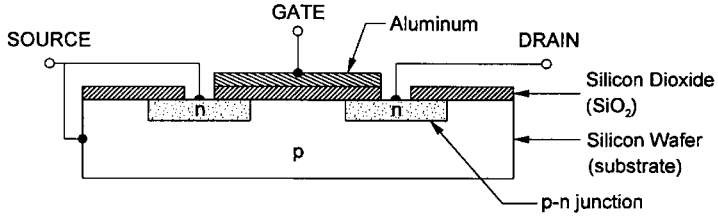

Figure 1.20 Typical construction of a MOSFET (metal-oxide semiconductor field-effect transistor). *(After Levine.[2])*

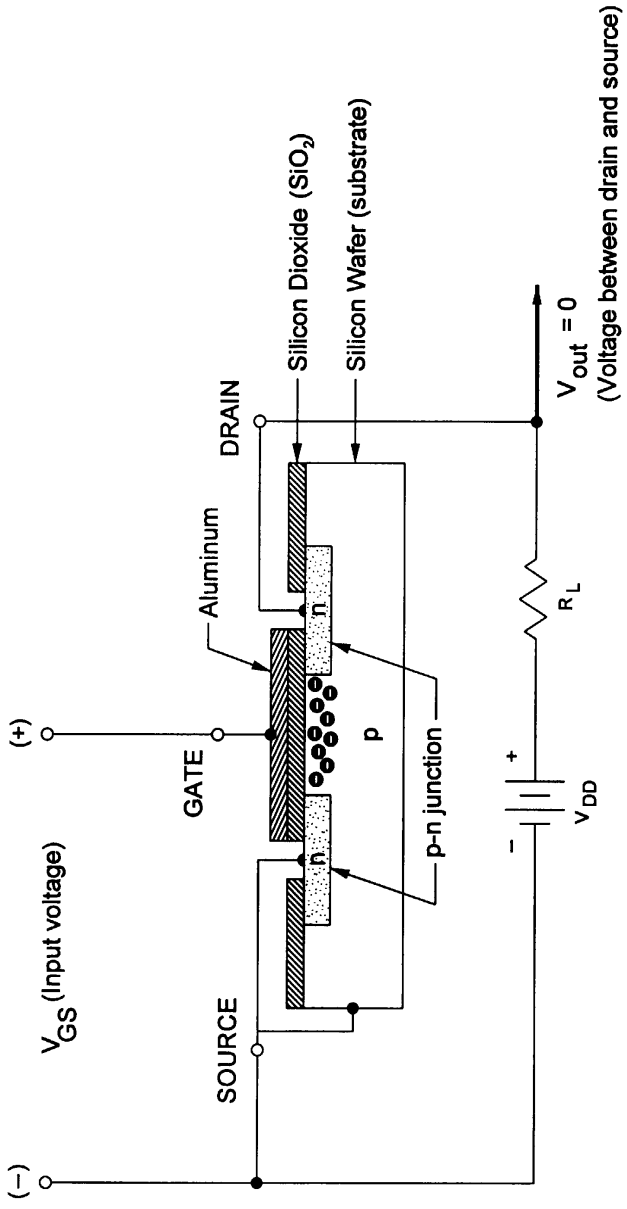

Figure 1.21 MOSFET functioning as an "on" switch. *(After Levine.[2])*

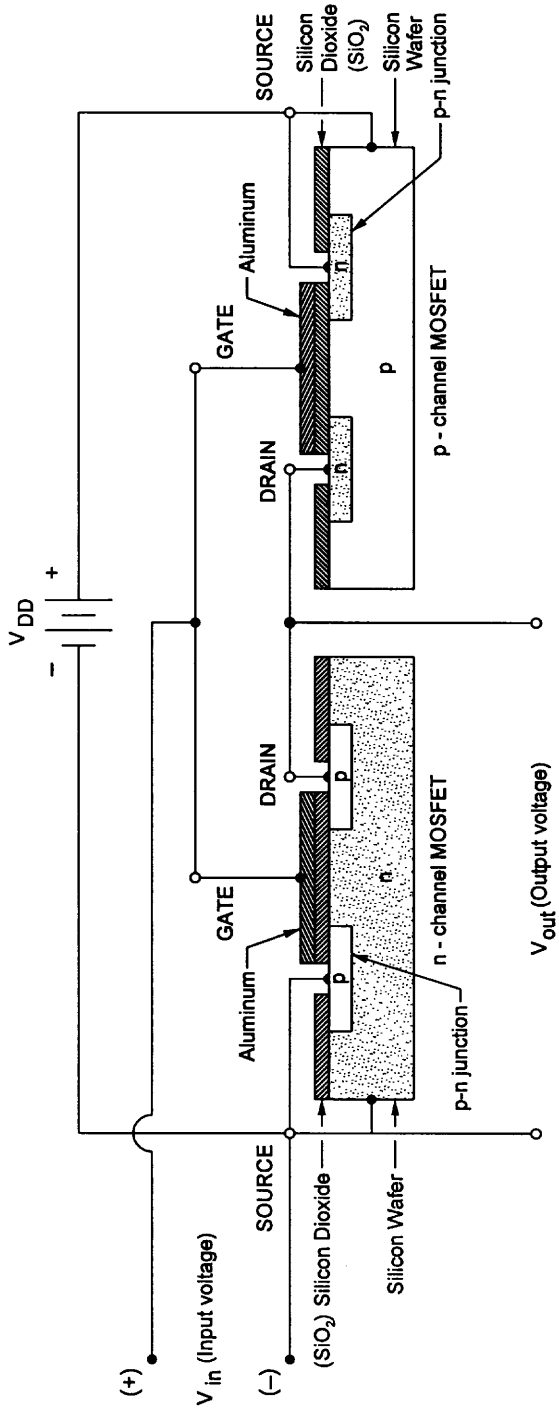

Figure 1.22 CMOSFET (n-p-n MOSFET connected to a p-n-p MOSFET to form a switch).

FET or CMOSFET. The advantages of a CMOSFET transistor are simplified circuitry (no load resistors required), very low power dissipation, and the capability to generate an output signal, which is the reverse of the input signal. For example, a positive input will have a zero output, or a zero input will create a positive output.

1.5 Fundamentals of Integrated Circuits

An integrated circuit (IC) chip is a collection of components connected to form a complete electronic circuit that is manufactured on a single piece of semiconductor material (Fig. 1.23). As described, the function of most solid state components is dependent on the properties of one or more p-n junctions

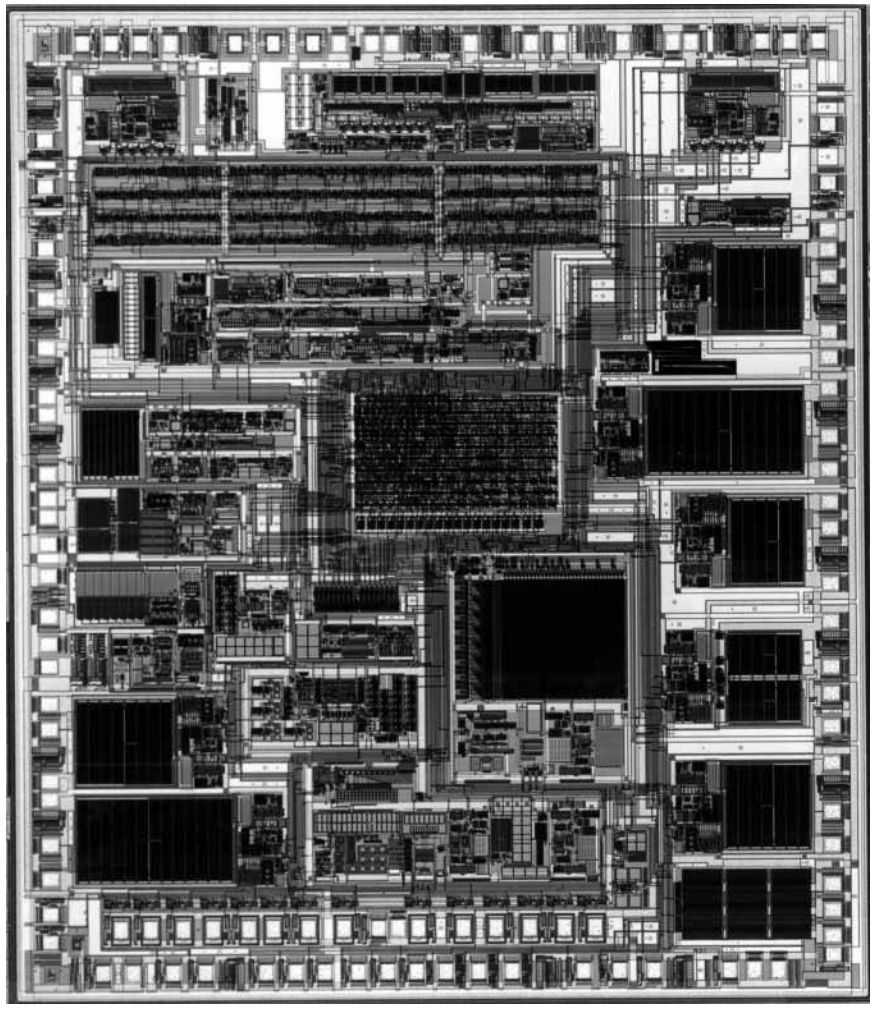

Figure 1.23 Typical IC chip. *(Courtesy of Agere Systems.)*

incorporated into their structures. Figure 1.24 illustrates the combination of various electrical components on an IC, showing their p-n junction structures.

Although the development of ICs was the result of contributions made by many people, Jack Kilby of Texas Instruments is credited with conceiving and constructing the first IC in 1958. In the Kilby IC, the various semiconductor components (transistors, diodes, resistors, capacitors, etc.) were interconnected with so-called "flying wires" (Fig. 1.25). In 1959, Robert Noyce of Fairchild was first to apply the idea of an IC in which the semiconductor

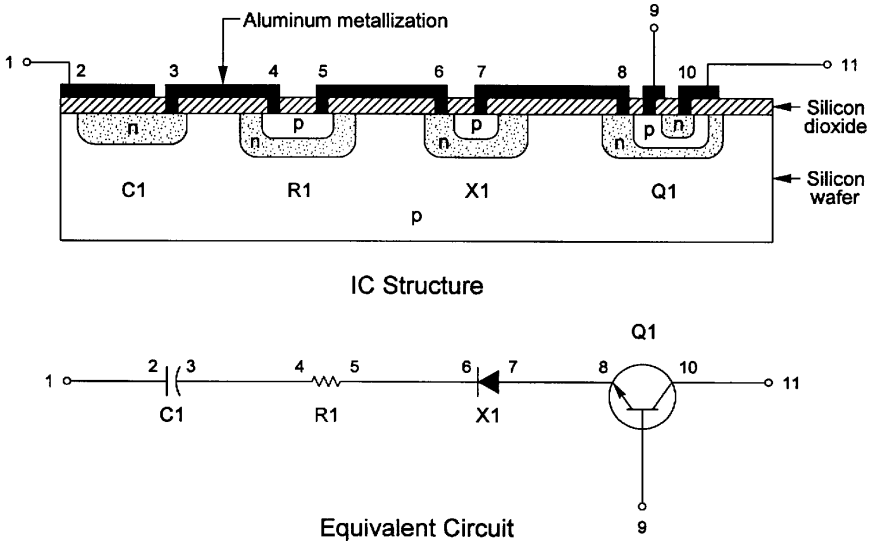

Figure 1.24 Typical silicon structure of electrical components.

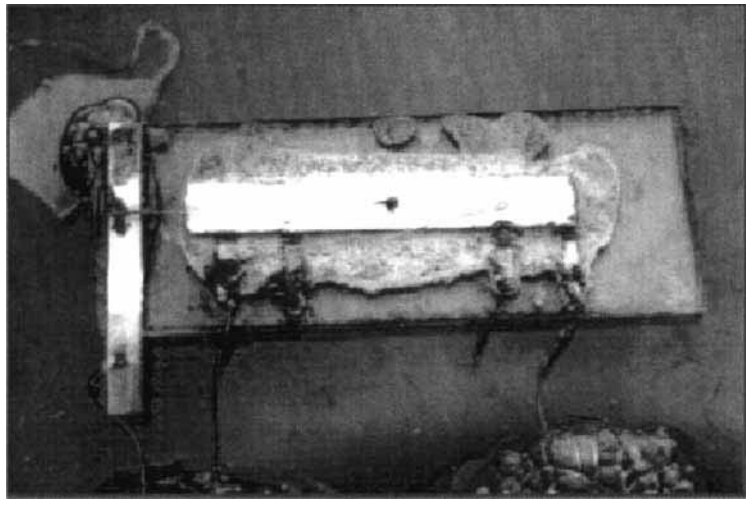

Figure 1.25 Jack Kilby's first integrated circuit.

components are interconnected within the chip using a planar fabrication process, thus eliminating the flying wires[4] (Fig. 1.26).

Over the last four decades, the electronics industry has grown very rapidly, with increases of over an order of magnitude in sales of ICs. In the 1960s, bipolar transistors dominated the IC market but, by 1975, digital metal-oxide semiconductor (MOS) devices emerged as the predominant IC group. Because of MOS's advantage in device miniaturization, low power dissipation, and high yields, its dominance in market share has continued to this day.

IC complexity has also advanced from small-scale integration (SSI) in the 1960s, to medium-scale (MSI), to large-scale integration (LSI), and finally to very large-scale integration (VLSI), which characterizes devices containing 10^5 or more components per chip. This rate of growth[3] is exponential in nature (Fig. 1.27) and, at the current rate of growth, the complexity is expected to reach about 5×10^9 devices per chip by the year 2005.

Continued reduction of the minimum IC feature dimensions[3] (Fig. 1.28) is a major factor in achieving the complexity levels mentioned. The feature size has recently been shrinking at an approximate annual rate of 11 percent. Thus, by the year 2006, it is expected to reach a minimum feature size of 10^2 nm (0.10 μm).

Device miniaturization has further improved the circuit-level performance, one improvement being the reduction of power consumption at the per-gate level. Figure 1.29 illustrates the exponentially decreasing trend in the power

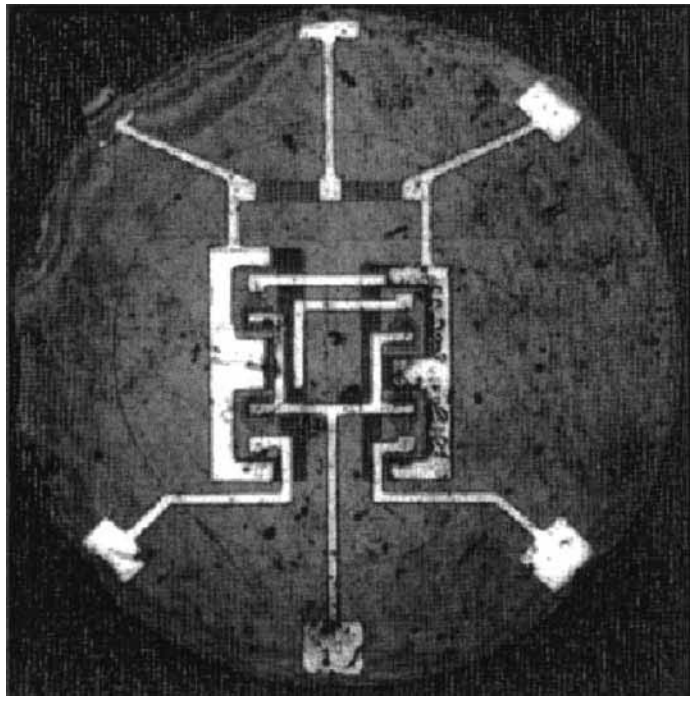

Figure 1.26 Early Fairchild IC using planar fabrication process.

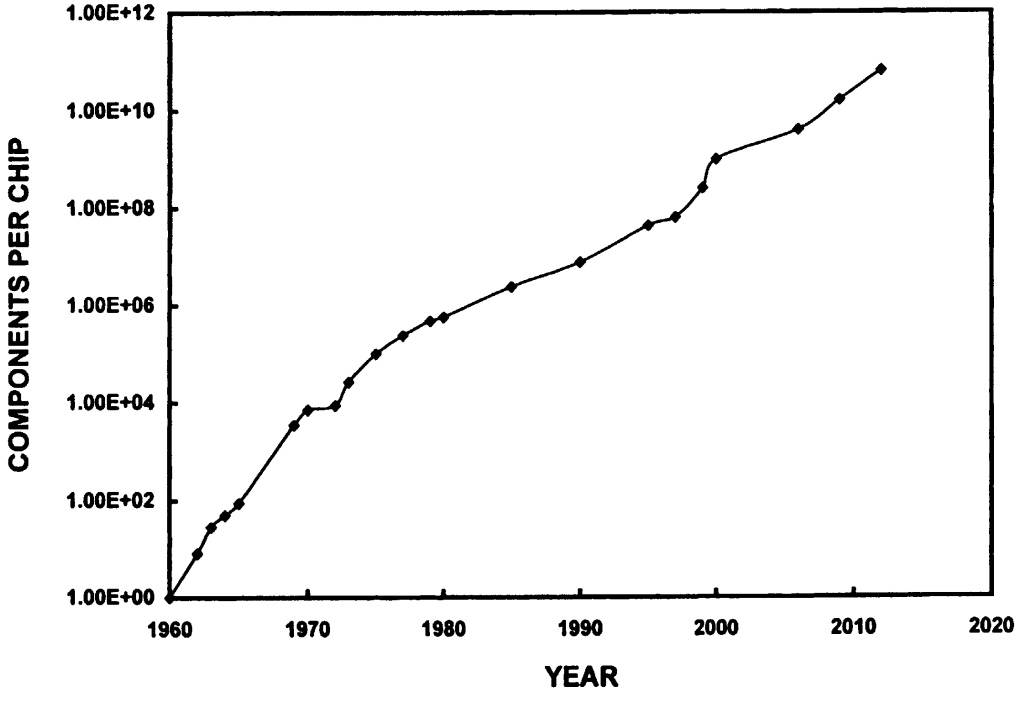

Figure 1.27 Exponential growth of components per IC ship for MOS memory. *(After Harper.[3])*

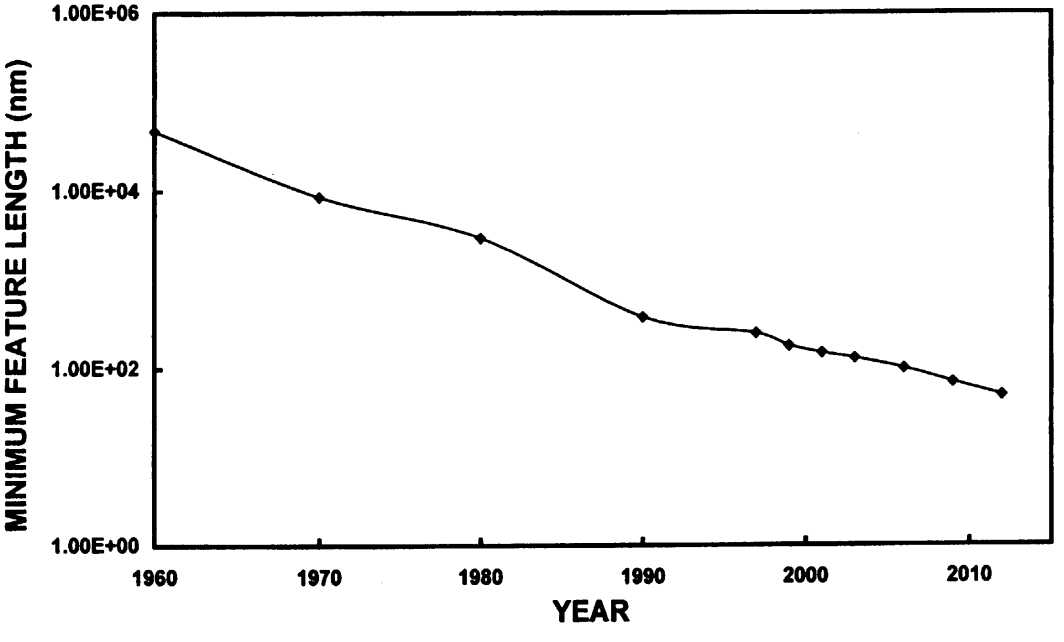

Figure 1.28 Exponential decrease of minimum device dimensions. *(After Harper.[3])*

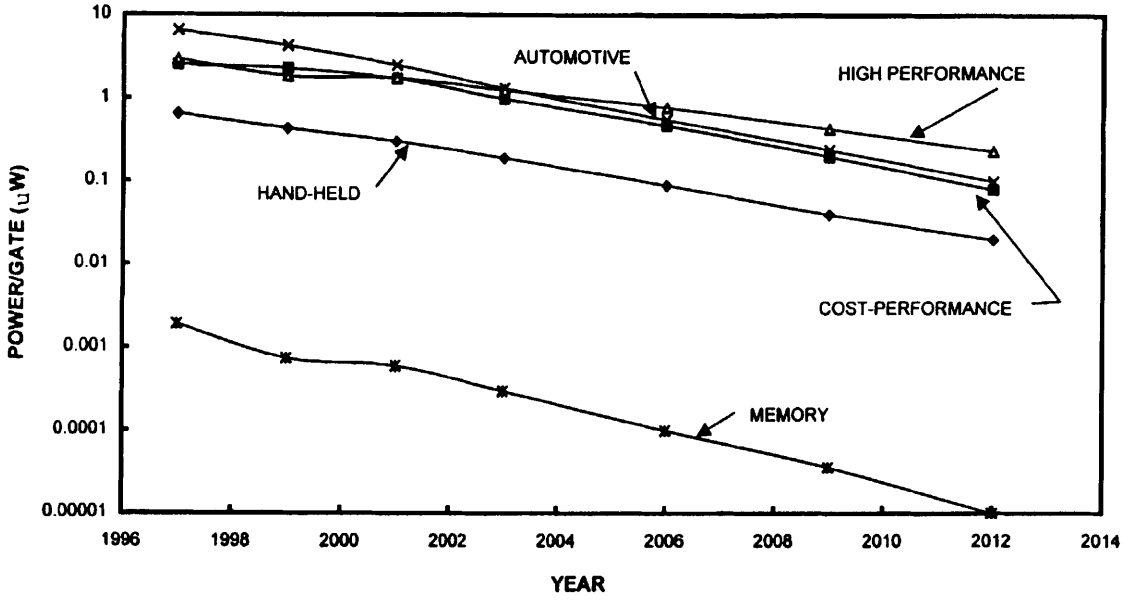

Figure 1.29 Trends in circuit power dissipation per gate. *(After Harper.[3])*

per gate for five major IC application groups: automotive, high-performance, cost-performance, hand-held, and memory. Figure 1.30, on the other hand, shows that the power dissipation per chip actually increased over the same period of time for the high-performance and cost-performance groups, whereas, for the automotive, hand-held, and memory groups, the power dissipation remained relatively constant. This is explained by the fact that, while the power per gate scales linearly with feature size, the power dissipation per chip, P, is largely influenced by the inverse square of the feature-size, as shown below.

$$P = f(Freq, C, V^2, Gate\ Count)$$

where $Freq$ = clock frequency
C = capacitance
V = voltage
$Gate\ Count$ = chip area / (feature size)^2

While the clock frequency and gate count have been increasing exponentially over the years (Figs. 1.27 and 1.31), the capacitance and voltage have been decreasing. Therefore, the increase in chip power dissipation is primarily due to the greater number of gates on a chip made possible by the decrease in the feature size.

Device miniaturization has resulted in significant improvements in on-chip switching speeds. Off-chip driver rise-time trends for ECL, CMOS, and GaAs are shown in Fig. 1.32. MOS circuits are known to be more sensitive to loading

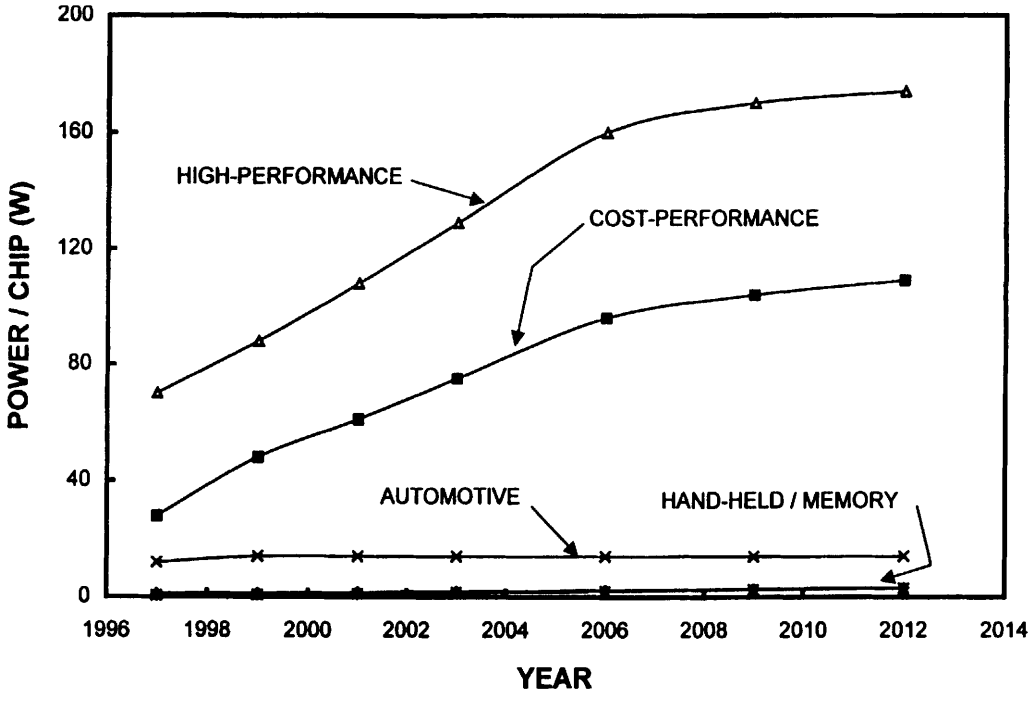

Figure 1.30 Trends in circuit power dissipation per chip. *(After Harper.[3])*

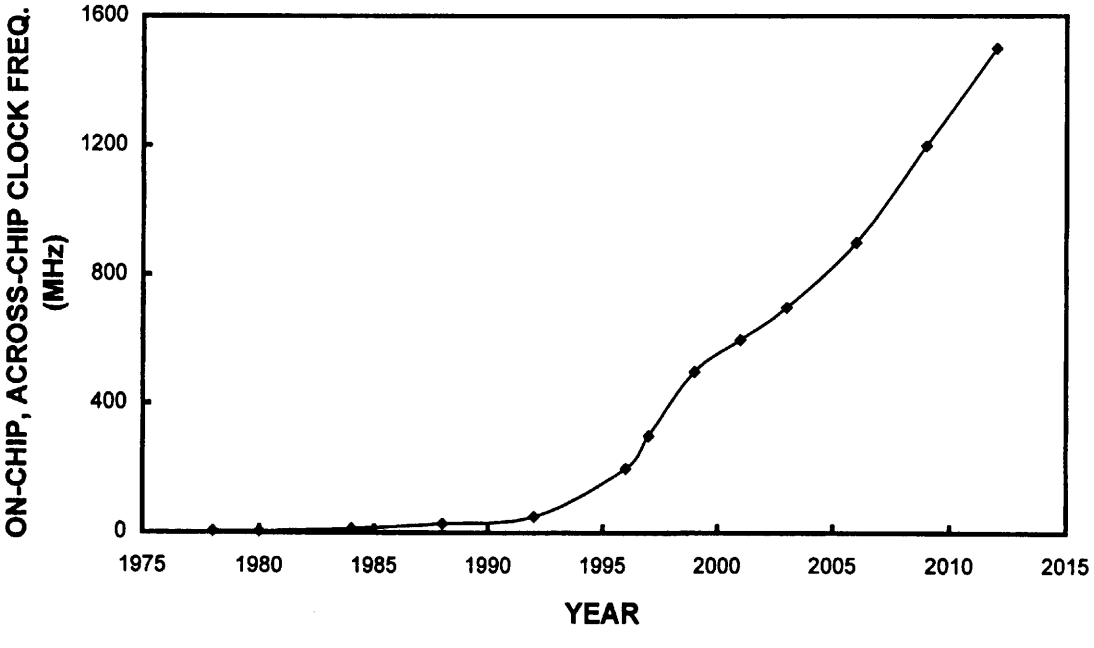

Figure 1.31 Frequency trends of high-performance ASIC chips. *(After Harper.[3])*

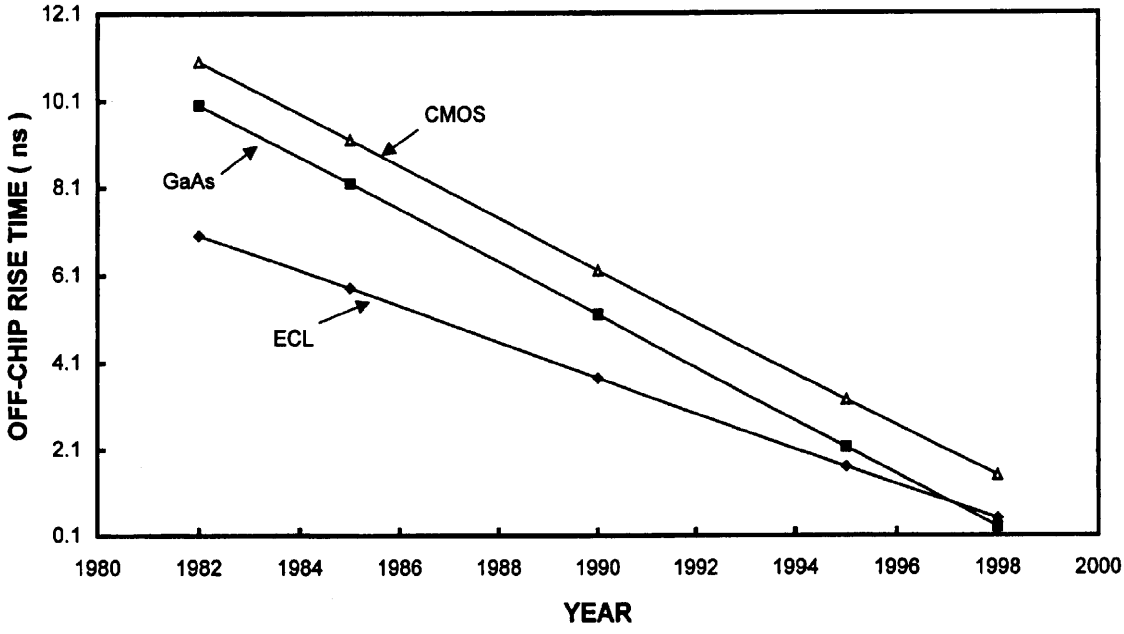

Figure 1.32 Off-chip rise times (typical loading). *(After Harper.[3])*

conditions due to their relatively high output impedance. Hence, interconnect density is more important in MOS systems than for bipolar designs. As the applications for these devices tend toward the nanosecond and subnanosecond signal rise times, more attention will be directed to the electrical design consideration of packages and interconnections.

Reduced unit cost per function is a direct result of miniaturization. The cost per bit of memory chips was cut in half every two years for successive generations of DRAMs. By the year 2005, the cost per bit is projected to be between 0.1 and 0.2 microcents for a 1-Gb memory chip. Similar cost reductions are projected for logic ICs.

1.6 IC Chip Fabrication

This section describes wafer preparation and the processes involved in fabricating the solid state components (ICs). The IC chips, which are configured on the wafer in a step-and-repeat pattern, are formed in a batch process. The pitch of the chip array pattern is dependent on the IC chip size and the width of the "saw street" separating the chips from each other. The width of separation is equal to the thickness of the saw used in singulation. The economics of chip fabrication dictate that as many chips as possible be processed at the same time on a given wafer. Thus, reducing the size of the chips by decreasing their feature dimensions and using larger-diameter wafers are the most cost-effective ways of fabricating ICs. Figure 1.33 shows a typical wafer with chips covering the entire wafer surface.

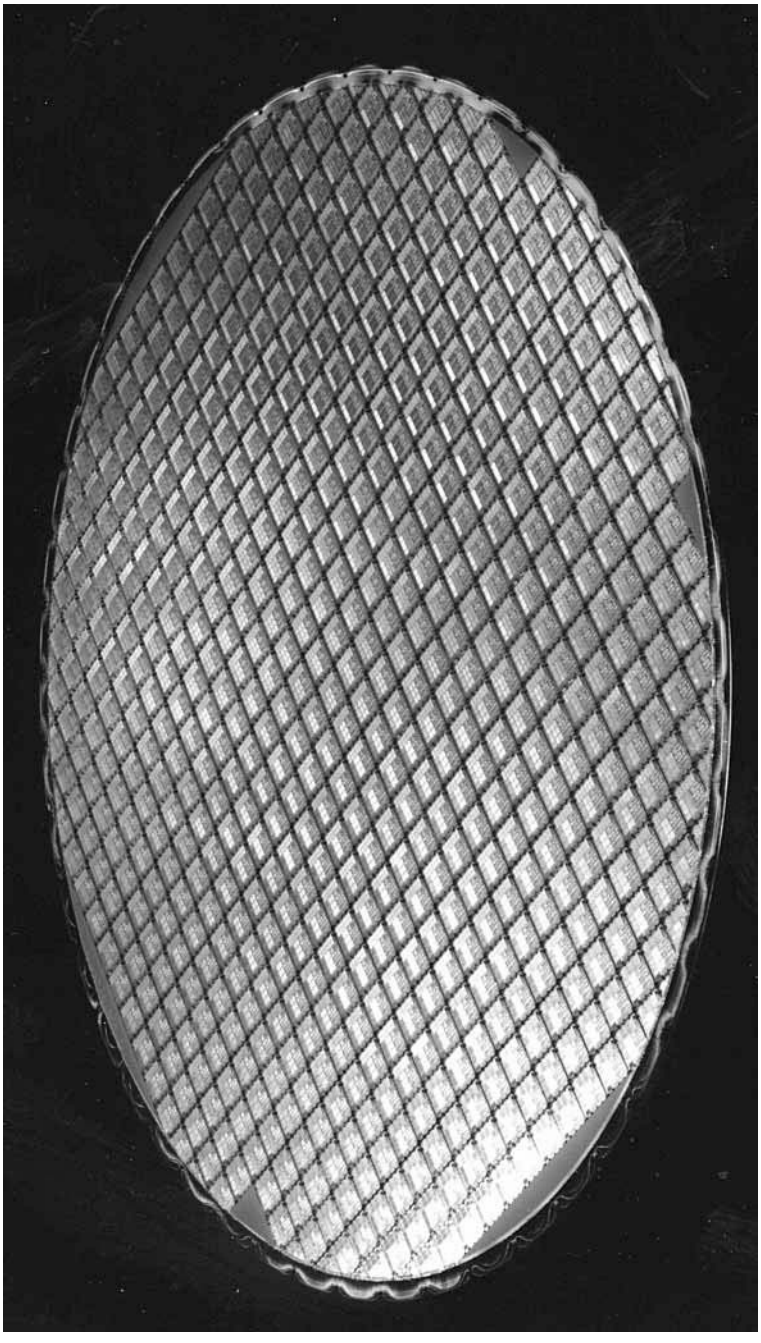

Figure 1.33 Typical wafer with an array of chips. *(Courtesy of Agere Systems.)*

Of all the semiconductor materials described in Sec. 1.4, silicon is used the most, because it is found in abundance in nature and its silicon dioxide (SiO_2) has many properties ideal for IC fabrication. As a result, this section will use silicon as the exemplary material to describe IC fabrication.

IC fabrication comprises many physical and chemical process steps (Fig. 1.34) that involve state-of-the-art equipment in ultra-clean environments. The following are the step-by-step processes used to fabricate ICs.

1.6.1 Ingot growth and wafer preparation

Before starting on the fabrication of ICs, the silicon wafer, defined as the semiconductor substrate upon which ICs are formed, must be fabricated.

The first step in producing a silicon wafer is to refine raw silicon, which is obtained from either beach sand or quartz mined from agatized rock formations. The sand or quartz is heated along with reacting gases at approximately 1700°C to separate and remove the impurities. The remaining material is chemically purified silicon (nuggets), which has a polycrystalline structure that lacks uniformity in the orientation of its cells. The polycrystalline silicon cannot be used to fabricate wafers but has to be further processed to convert it into a monocrystalline structure containing a single-crystal silicon with uniform cell structures. The silicon nuggets are placed in a quartz crucible (Fig. 1.35) and heated to 1415°C (the melting point of silicon). From the molten silicon, a single-crystal ingot is grown and then sliced into wafers upon which ICs are fabricated.

There are several methods used to grow silicon ingot, but the Czochralski (CZ) method is the most popular. A single silicon crystal seed is placed at the end of a rotating shaft and lowered into the heated crucible until the seed touches the surface of the molten silicon (Fig. 1.35). By continually rotating the shaft and crucible in opposite directions and simultaneously pulling the seed away from the molten silicon, a silicon crystal is formed at the seed/melt interface with an identical crystal structure as the seed. The monocrystalline silicon ingot continues to be formed as the seed is slowly withdrawn from the crucible and the supply of molten silicon is replenished. To grow an n- or p-type crystal structure, small amounts of impurities (dopants) are introduced to the melt. For example, a phosphorus dopant, when mixed with the pure silicon melt, will produce an n-type crystal, whereas a boron dopant will produce a p-type.

The shape of the ingot consists of a thin circular neck formed at the seed end [approx. 0.12 in (3.0 mm) dia.], followed by the main cylindrical body, and ending with a blunt tail. The length and diameter of the ingot is dependent on the shaft rotation, withdrawal rate of the seed, and the purity and temperature of the silicon melt. Ingot sizes vary from 3 in (75 mm) to 12 in (300 mm) dia. and have a maximum length of approx. 79 in (2 m) (Fig. 1.36). The ingots are grown at a rate of about 2.5 to 3.0 in/hr (63.5 to 76.2 mm/hr).

The following are typical processing steps to prepare a silicon wafer for IC fabrication (Figs. 1.37 and 1.38):

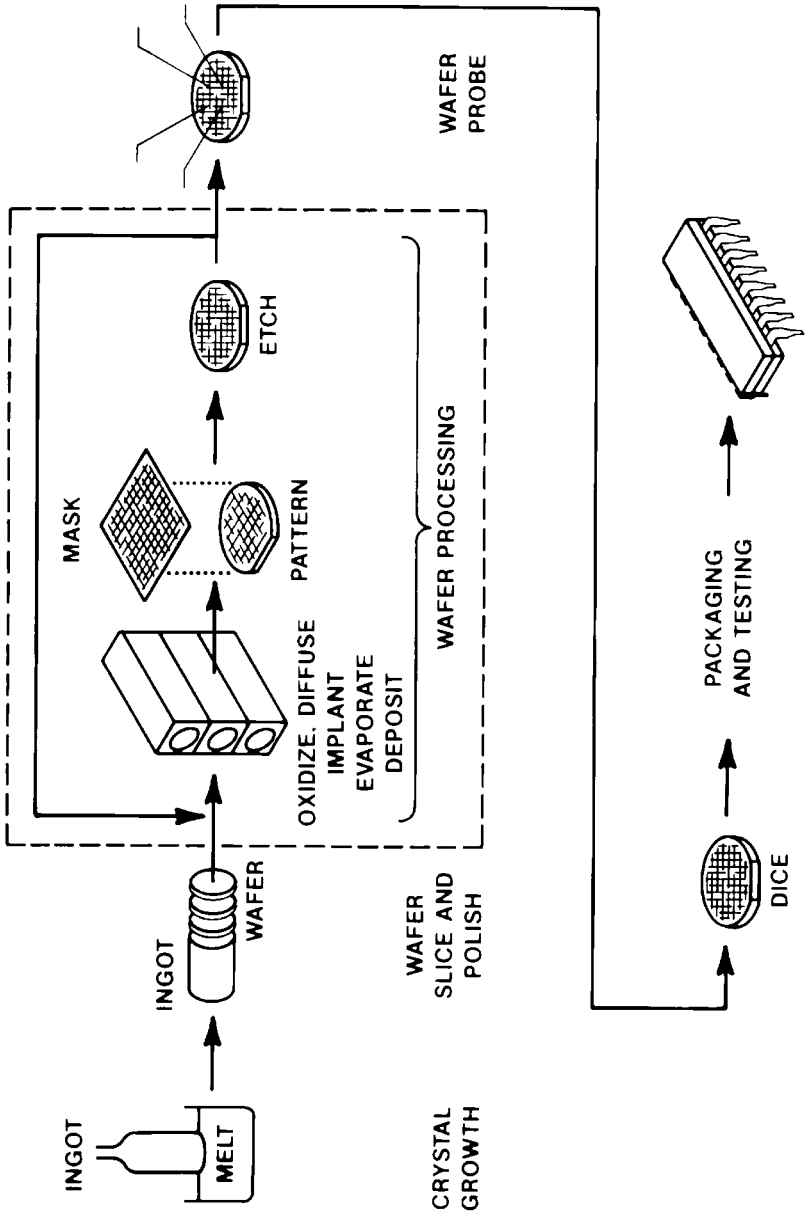

Figure 1.34 Typical IC fabrication processes.

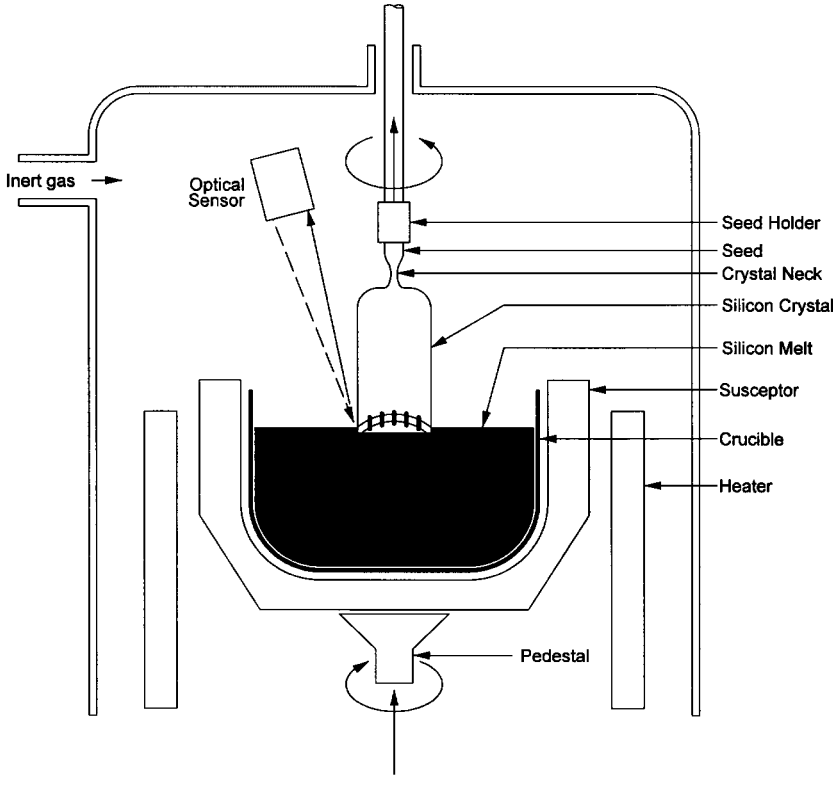

Figure 1.35 The Czochralski (CZ) method of growing a silicon ingot.

1. The ingot is cut to a uniform diameter and then checked for crystal orientation, conductivity type (n- or p-type), and resistivity (amount of dopant used).
2. A flat is ground along the axis to be used as reference for crystal orientation, wafer imaging alignment, and electrical probing of the wafer. Sometimes, a secondary, smaller flat is also ground, whose position with respect to the major flat signifies the orientation and type of conductivity (p- or n-type) the crystal has. Larger-diameter ingots may use a notch for this purpose.
3. The ingot is now ready to be sliced into thin disks, called wafers, which may vary in thickness from 0.020 in (0.50 mm) to 0.030 in (0.75 mm), depending on the wafer diameter. Wafers are sliced with either an inner diameter saw blade or a wire saw. The saw blade slicing technique consists of a 0.006-in (0.152-mm) thick stainless steel blade with an inside diameter cutting edge that is coated with diamonds. The cutting edge, being on the inner diameter of a large hole cut out of a thin circular blade, is fairly rigid. The slicing process is sequential; that is, one wafer is cut at a time, which takes approximately nine minutes.

Figure 1.36 Typical silicon ingots. *(Courtesy of Agere Systems.)*

The wire saw, on the other hand, slices the wafers in a batch process, cutting all the wafers at once in a 16-in (410-mm) length of ingot. The process consists of a wire-winding mechanism, which positions the wires parallel to each other at a pitch equal to the wafer thickness to be cut. The wires are 0.007 in (0.170 mm) dia. and are made of stainless steel coated with brass. The slicing equipment includes a wire guiding unit and a tensioning and wire feed-rate mechanism. The wires continually travel in a closed loop by winding up on one spool and unwinding from another. A silicon carbide slurry, which acts as an abrasive, coats the wires prior to cutting through the silicon ingot. The wires travel about 10 m/s, and it takes approximately 5.5 hr to cut through all the wafers at once.

4. The wafers are laser marked for identification.
5. The sliced wafers are lapped, to remove any imperfections caused by sawing, and then deburred and polished on the top side, to a mirror-like finish. This provides a flat surface for subsequent IC fabrication processes.

1.6.2 Cleanliness

The processes explained so far involved preparation of the wafers for the next phase of IC fabrication, i.e., forming the circuitry. Before proceeding to describe new processes, we must first examine a critical aspect of IC fabrication that affects the yield at every step, namely the cleanliness of the environment where ICs are being produced. Contamination control in the fabrication area is of great concern, because lower yields, caused by unwanted particles, chemicals, or metallic ions in the atmosphere, increase IC costs.

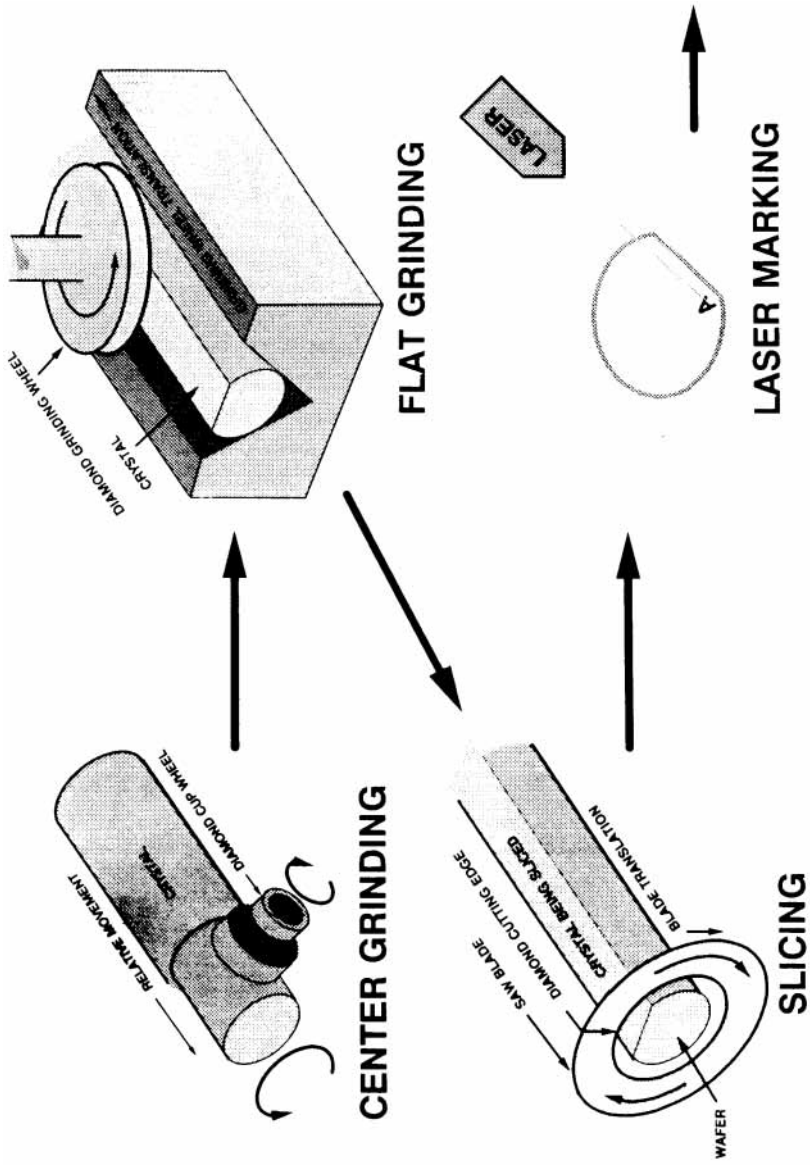

Figure 1.37 Cutting silicon ingot into wafers.

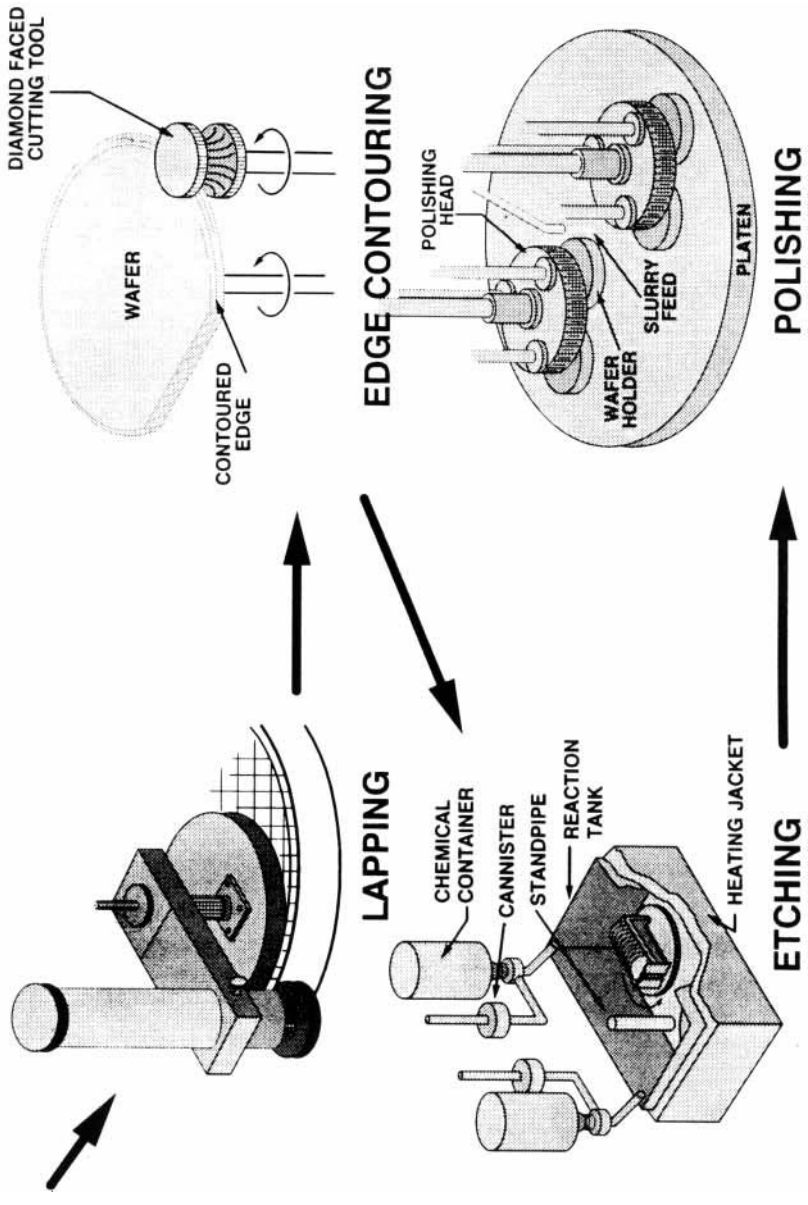

Figure 1.38 Wafer processing.

To control the environment, all IC fabrication processes are housed in clean rooms that are classified by how many particles, 0.5 μm in diameter, are allowed in one cubic foot of air. In general, clean rooms range in classification from Class 1 to Class 100,000, with particle size distributions as shown in Figure 1.39. For example, a Class 1000 clean room can have 1000, 0.5-μm size particles in one cubic foot. For IC fabrication, clean rooms range from Class 1 to Class 1,000, depending on the needs of the process.

1.6.3 IC fabrication

Having explained the importance of cleanliness on IC fabrication, let us resume with the processes involved in forming the circuitry in and on the wafers. The following ten basic IC fabrication processes are described:

- Oxidation
- Photolithography
- Diffusion
- Epitaxial deposition
- Metallization
- Passivation
- Backside grinding
- Backside metallization
- Electrical probing
- Die separation

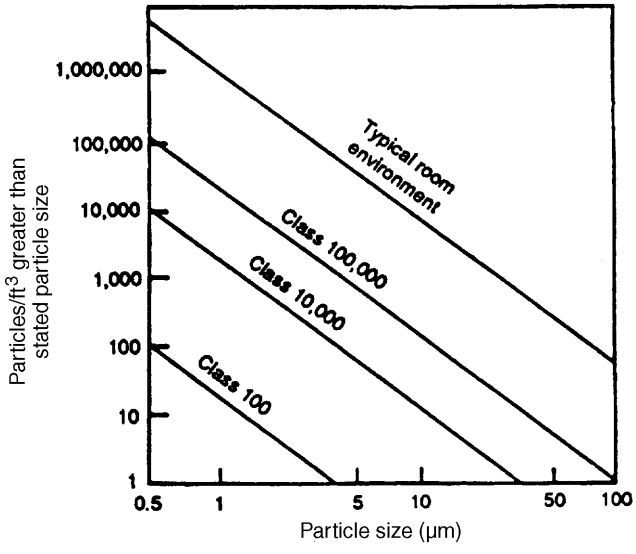

Figure 1.39 Particle size distribution in typical clean room atmosphere and in three classes of clean environments. *(After Harper.[3])*

1.6.3.1 Oxidation. Oxidation is the process of forming a silicon dioxide (SiO_2) layer on the surface of the silicon wafer. The silicon dioxide is an effective dielectric that is used to construct IC components, such as capacitors and MOS transistors. Because it acts as a barrier to doping and can easily be removed with a chemical solvent, the silicon dioxide is also an ideal template when used in the doping process. Silicon dioxide is formed by heating the wafer in an atmosphere of pure oxygen at a temperature between 900 and 1200°C, depending on the oxidation rate required. The oxidation can be speeded up if water vapor is introduced into the oxygen. The silicon dioxide growth on the silicon wafer, as the oxygen in contact with the wafer surface diffuses through the oxide layer to combine with the silicon atoms. As the oxide layer grows, it takes longer for the oxygen to reach the silicon, and the rate of growth slows. The growth of a 0.20-μm thick layer of silicon dioxide, at 1200°C and in dry oxygen, takes approximately 6 min, whereas, to double the oxide layer thickness to 0.40 μm takes 220 min or 36 times as long.

The parameters that affect the silicon dioxide growth rate are

- Use of dry oxygen or in combination with a water vapor
- Ambient pressure within the furnace
- Temperature in the furnace
- Crystal orientation
- Time

The silicon dioxide layers vary in thickness from 0.015 to 0.05 μm for MOS gate dielectrics or 0.2 to 0.5 μm thick when used for masking oxides or surface passivation.

1.6.3.2 Photolithography. Photolithography is a patterning process whereby the elements representing the IC circuit are transferred onto the wafer by photomasking and etching. Photolithography has similarities to photographic processes. The images of the various semiconductor element layers are formed on reticles or photomasks made of glass, which are then transferred to a photoresist material on the surface of the silicon wafer. The resist may be of a type that changes its structure and properties to either UV light or laser. If it is a negative-acting photoresist, the areas that are exposed to UV light polymerize (harden) and thus are insoluble during development, whereas the unexposed areas are washed away. This results in a negative image of the photomask being formed in the photoresist. An alternative to the negative image forming photoresist is a positive-reacting photoresist, where the material behaves in the opposite way. Areas exposed to UV light become unpolymerized, or soluble when immersed in chemical solvents.

Until the advent of VLSI circuits in the mid 1980s, the negative-reacting photoresist, because of its superior developing characteristics, was the resist most commonly used in the industry. However, due to its poor resolution capability, the negative photoresist could no longer provide the requirements de-

manded by the high-density features of VLSI circuits. As a result, the semiconductor industry has transitioned to the positive-reacting photoresist because of its superior resolution capability. The transition was difficult, because not only was the photomask or reticle changed to a positive image, but the industry had to overcome the resist's lower adhesion capability and reduced solubility differences between polymerized and unpolymerized areas. Photomasking is used for patterning both the silicon dioxide and the metallization layers.

The increasing need for ICs to be smaller and operate at higher speeds has forced the industry to develop ICs with ever smaller features (see Sec. 1.5 for feature size trends). As feature sizes decrease, the patterning technology has to advance to where the requirements of high resolution, tight pattern registration (alignment), and highly accurate dimensional control are met. Photomasking is the most critical element of the IC fabrication process in that alignment of the different photomask overlays and mask contamination have an overwhelming effect on fabrication yield. The following photomasking methods are used for patterning:

- Optical exposure
  - Contact printing
  - Proximity printing
  - Scanning projection printing
  - Direct wafer stepping
- Non-optical exposure
  - Electron beam
  - X-ray lithography

The characteristics of each patterning method are described in Table 1.1.

A typical photolithography process for patterning the silicon dioxide layer consists of the following steps (Fig. 1.40):

1. The silicon wafer undergoes an oxidation process (Sec. 1.6.3.1) where a silicon dioxide (SiO_2) layer is grown over its entire surface.
2. A drop of positive photoresist is applied to the SiO_2, and the wafer is spin coated uniformly across the surface.
3. If contact printing is used, a photomask, containing transparent and opaque areas that define the pattern to be created, is placed directly over the photoresist. In areas where the photoresist is exposed to UV light, projected through the mask, it becomes unpolymerized (does not harden), and where the UV light is blocked, the material polymerizes (hardens).
4. The photomask is removed, and the resist is developed to dissolve the unpolymerized areas, exposing the silicon oxide below.
5. The wafer is then wet or dry etched to remove the exposed oxide, resulting in a pattern identical to the photomask. Wet etching typically consists of

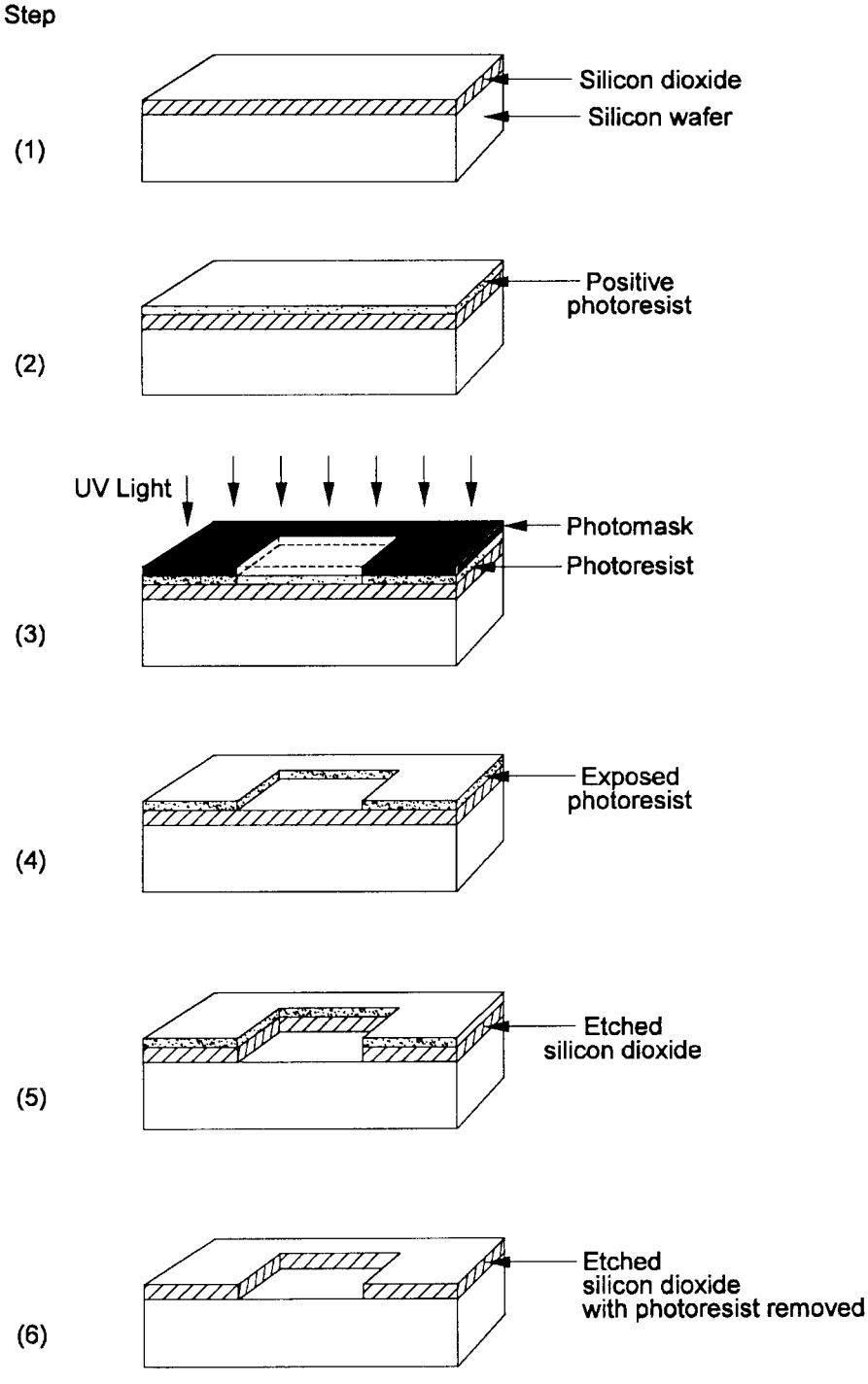

Figure 1.40 Typical photolithographic process for selective removal of silicon dioxide.

TABLE 1.1 Characteristics of Patterning Methods

Patterning methods	Description	Advantages	Disadvantages
Contact printing	Mask is placed directly on resist with UV light shining through the mask onto the wafer.	Good resolution High throughput	Causes defects such as scratching of mask and resist Adherence of dirt to the mask may block the light
Proximity printing	Small separation between mask and resist with UV light shining through the mask onto the wafer.	Less damage to mask and resist High throughput	Poor resolution due to some light scattering Not used for VLSI photomasking
Scanning projection printing	The UV light source is a slit projected through the mask. By use of optics, the pattern image of one slit width at a time is projected onto the wafer, exposing the resist	Good resolution High throughput	Alignment problems Possible image distortion from dust and glass damage
Direct wafer stepping	Based on refractive optics, the image of one or several chip sites is projected onto the wafer exposing the resist. The process is step repeated until the whole wafer is patterned.	Good resolution with fewer defects Better alignment Less vulnerable to dust and dirt Most used for VLSI fabrication Medium throughput	Tight maintenance requirement of humidity and temperature control
Electron beam	An electron beam produces a small diameter spot that is directed in an x-y direction, onto the wafer. The electron beam is capable of being turned ON and OFF to expose the resist as needed to form the pattern.	No mask is used Excellent resolution	High cost Low throughput
X-ray lithography	The process resembles the UV light system of the proximity printing method, but high energy X-rays are used instead.	Produces smaller pattern then light sources Excellent resolution	Requires masks made from gold or other refractory materials capable of blocking X-rays Low throughput

immersing the wafer in a diluted solution of hydrofluoric acid for a specified time that will result in complete etching. The wafers are then rinsed and dried. Wet etching is primarily used for wafers with IC feature sizes greater than 3 μm. For high-density etching, the dry etching technique is used, because it's more precise. Dry etching can be accomplished by three different etching techniques: plasma, ion beam milling, and a reactive ion etch. All three techniques use gases as the etching medium.

6. The remaining photoresist is removed with a chemical solvent.

This process is repeated a number of times to create the desired semiconductor elements on the wafer surface.

1.6.3.3 Diffusion. As was discussed in Sec. 1.4, when forming solid state components, silicon is not used in its natural or intrinsic state but is converted to either an n-type or p-type semiconductor. The n- or p-type materials, by themselves, are of little value unless they are joined to form a p-n junction. Diffusion or doping is the process of implanting impure atoms in a single crystal of pure silicon so as to convert it into n-type or p-type material. Depending on the dopant element used, antimony, arsenic, and phosphorus will produce an n-type material, whereas boron will produce a p-type structure. The basic dopant elements are available either in solid, liquid, or gaseous states as shown in Table 1.2. Type of dopant, dopant concentration, time of exposure, and temperature affect the diffusion process.

1.6.3.4 Epitaxial Deposition. This is a process whereby a thin layer of silicon (approximately 25 μm thick) is deposited upon the surface of an existing silicon wafer and doped using the same dopant types and delivery systems used in the diffusion process. Thus, this is another technique for fabricating p-n junctions. Although there are several deposition methods available, chemical

TABLE 1.2 Common Dopant Sources (after Zant[6])

Type	Element	Compound name	State
n-type	Antimony	Antimony trioxide	Solid
	Arsenic	Arsenic trioxide	Solid
		Arsine	Gas
	Phosphorus	Phosphorus oxychloride	Liquid
		Phosphorus pentoxide	Solid
		Phosphine	Gas
p-type	Boron	Boron tribromide	Liquid
		Boron trioxide	Solid
		Diborane	Gas
		Boron trichloride	Gas
		Boron nitride	Solid

vapor deposition (CVD) is the most commonly used technique. The basic CVD process consists of the following:

1. Silicon wafers are placed in a reaction chamber with an inert gas and heated to a temperature that depends on the reaction and parameters of the deposition method used and layer thickness required.
2. Reactant gases are introduced into the reaction chamber at a specified flow rate, where they come in contact with the wafer surface.
3. As the reactants are absorbed by the silicon wafer, the chemical reaction forms the deposition layer. The surface reaction rate is dependent on the temperature; increasing the temperature increases the reaction rate.
4. To dope the deposition layer, dopant gases are introduced into the reaction chamber where they combine with the deposited layer to form an n- or p-type material.
5. The gaseous by-products are flushed from the reaction chamber.
6. The wafers are removed from the chamber, and the deposited layer is checked for thickness, coverage, purity, cleanliness, and n- or p-type composition.

Variations in the CVD techniques, involving changes in vapor pressure and temperature in the chamber, have resulted in process enhancements. There are three different CVD techniques used in the industry:

- Atmospheric pressure CVD (APCVD)
- Low-pressure CVD (LPCVD)
- Plasma-enhanced CVD (PECVD)

The characteristics of the above techniques are shown in Table 1.3.

TABLE 1.3 CVD Techniques (after Wolfe[5])

Process	Advantages	Disadvantages	Application	Temp. range
APCVD	Low chemical reaction temperature Simple horizontal tube furnace Fast deposition	Poor coverage Particle contamination	Low temperature oxides, both doped and undoped	300–500°C
LPCVD	Good coverage and uniformity Vertical loading of wafers for increased productivity	High temperature Low deposition rate	High temperature oxides, both doped and undoped	580–900°C
PECVD	Lower chemical reaction temperature Good composition, coverage and throughput	High equipment cost Particulate contamination	Low temperature insulators over metals or passivation	200–500 °C

1.6.3.5 Metallization. The deposition of a conductive material, to form the interconnection leads between the circuit component parts and the bonding pads on the surface of the chip, is referred to as the *metallization* process. As chip density increases, interconnection can no longer be accomplished on a single level of metal but requires multilevel metallization with contact holes or vias interconnecting the various levels.

Materials such as aluminum, aluminum alloys, platinum, titanium, tungsten, molybdenum, and gold are used for the various metallization processes. Of these, aluminum is the most commonly used metallization material, because it adheres well to both silicon and silicon dioxide (low contact resistance), it can be easily vacuum deposited (it has a low boiling point), it has a relatively high conductivity, and it patterns easily with photoresist processes. In addition to pure aluminum, alloys of aluminum are also used for different performance related reasons; i.e., small amounts of Cu are added to the aluminum to reduce the potential for electromigration effects. Electromigration may occur during circuit operation, when high currents are carried by the long aluminum conductors, inducing mass transport of metal between the conductors. Sometimes small amounts of silicon or titanium are added to the aluminum to reduce the formation of metal "spikes," that occur over contact holes.

The aluminum metallization process consists of depositing aluminum on the wafer surface and again using the photoresist process to etch away the unwanted metallization. One of the techniques used to apply the aluminum is the vacuum deposition process wherein the aluminum is evaporated in a high-vacuum system and redeposited over the wafer surface. This process has the disadvantage of nonuniform metal coverage. Sputtering is another method for depositing aluminum metallization. Because it offers better control of the metallization quality than the vacuum deposition method, it's currently being used in the majority of IC metallization processes. Sputtering is a physical (nonchemical) method of deposition, which is performed by ionizing inert gas (Argon) particles in an electric field and then directing them toward an aluminum target. There, the energy of the incoming particles dislodge or "sputters off" atoms of the aluminum target, which are then deposited onto the wafer.

One of the problems encountered when pure aluminum is in contact with silicon, while being heated, is the formation of an eutectic aluminum-silicon alloy. The alloy formation penetrates into the wafer, where it can reach shallow junctions, causing leakages or shorting. To alleviate this problem, a metal barrier such as titanium tungsten (TiW) or titanium nitrate (TiN) is placed between the aluminum and the silicon. Adding silicon (1 to 1.5 percent by weight) to the aluminum is another way of preventing the formation of aluminum-silicon alloy, although this is less effective. Some alloying with the silicon wafer still occurs, but to a lesser extent.

The electrical performance of any given type of metallization is dependent on its resistivity, contact resistance, and the length and thickness of the conductor. To improve electrical performance in MOS circuits, the resistivity and contact resistance of the conductors are reduced through the use of barriers made of refractory metals such as titanium, tungsten, platinum, and molybde-

num, in combination with silicon, to form silicides of $TiSi_2$, WSi_2, $PtSi_2$, and $MoSi_2$, respectively. The silicides can also be used as conductors or via plugs.

As more and more chips are required to operate at higher frequencies, the current aluminum metallization can no longer meet the lower resistances needed to prevent data processing delays. As a result, copper has started to replace aluminum because of its lower resistance and reduced electromigration problems.

1.6.3.6 Passivation. The passivation layer is deposited at the end of the chip metallization process and is used to protect the aluminum interconnecting circuitry from moisture and contamination. An insulating or passivation layer of silicon dioxide or silicon nitride is vapor deposited over the chip circuitry (Fig. 1.41), with bond pads remaining exposed for wire bonding or flip-chip interconnection.

1.6.3.7 Backside grinding. At the end of the IC fabrication process, after the passivation layer is applied, wafers are sometimes thinned to fit the overall package height requirements. The thinning process consists of back grinding the wafer, similar to the procedure used in lapping the wafer, to remove any imperfections caused by sawing (Fig. 1.38)

1.6.3.8 Backside metallization. In cases in which the chip is to be eutectically bonded to a ceramic package, or where the back of the chip has to make electrical contact with the die attach area, it is necessary for the chip to have a gold film backing. The gold film is deposited by vacuum evaporation or sputtering and is done after backgrinding.

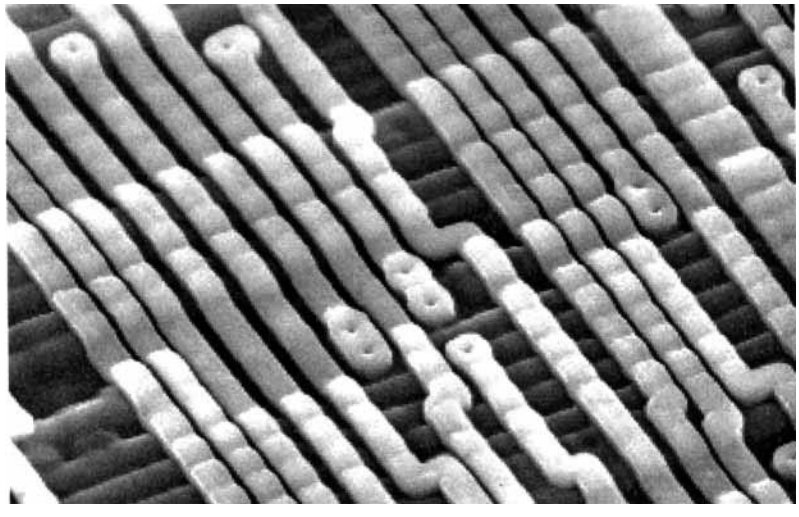

Figure 1.41 IC circuitry covered by a passivation layer. *(Courtesy of Agere Systems.)*

1.6.3.9 Electrical probing. The last step in wafer processing is to test the die. A test probe makes contact with the bonding pads on the surface of the wafer, and the chips are electrically tested against predetermined specifications. Chips thought to be faulty are inked, or an electronic map is developed indicating the bad chips.

1.6.3.10 Die separation. After the chips have been electrically tested, the chips are separated by two different methods:

1. *For chips thinner than 0.010 in (0.25 mm):* The chips are separated by first scribing shallow, fine, diamond-cut lines between the chips and then mounting the wafer onto a release tape affixed to a steel ring. Pressure from a roller is then exerted on the wafer, breaking it up into individual chips. The individual chips that tested good are removed by pushing the chip up (with a pin) from the underside of the tape and then picking them up with a vacuum tool called the *collet*. The chips are placed in a tray or are automatically transferred to the die attach process for IC packaging. This type of separation method may cause rough and cracked edges on the chip.
2. *For chips greater than 0.010 in (0.25 mm) thick:* As above, the wafer is mounted onto a release tape affixed to a steel ring and then cut between the chips, through the silicon thickness, using a diamond-impregnated round saw. The method for removing the good chips from the tape is similar to that for thinner chips. Unlike the break-up method, this separation process leaves smooth edges on the chip.

1.6.3.11 Typical construction of a p-n-p bipolar transistor (Fig. 1.42)

1. A silicon dioxide (SiO_2) layer is grown on a p-doped silicon wafer (Sec. 1.6.3.1).
2. A positive photoresist layer is applied to the SiO_2 (Sec. 1.6.3.2).
3. A photomask is created with opaque and clear areas, patterning the clear areas in locations where windows in the SiO_2 are to be formed. The photomask image is transferred onto the positive photoresist, which becomes polymerized in the areas where it is not exposed to the UV light (opaque areas in the photomask).
4. The resist is developed, and the unpolymerized areas dissolve, forming a window that exposes the SiO_2.
5. The silicon dioxide is etched away in the photoresist windows, exposing the silicon wafer.
6. The photoresist is removed.
7. Using phosphorus as the dopant, an n-type region in the p-type silicon base is created by diffusion (Sec. 1.6.3.3).

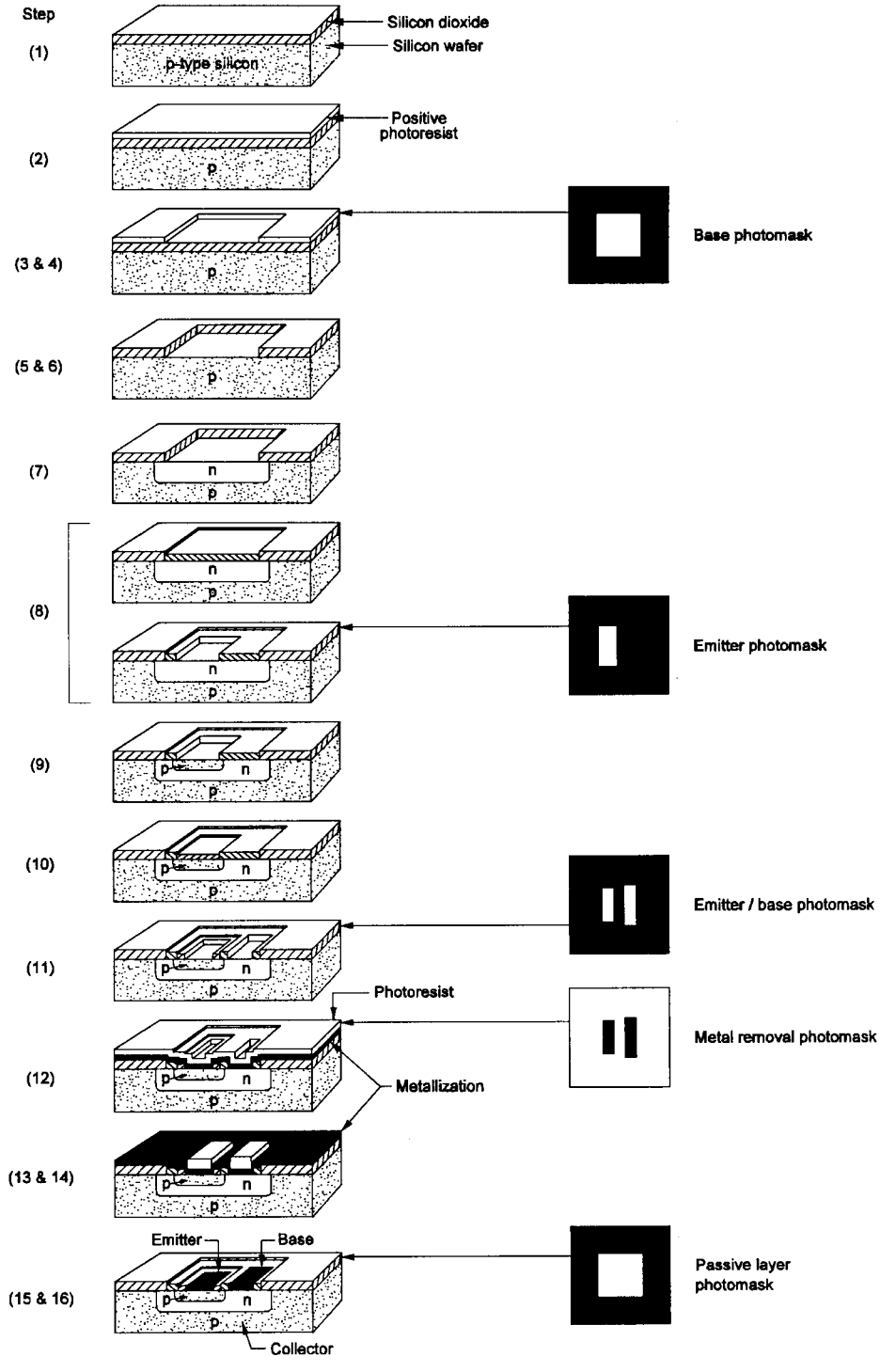

Figure 1.42 Typical process sequence in the fabrication of a silicon planar bipolar transistor.

8. A new layer of silicon dioxide is grown on the surface of the n-region, and steps (2) through (6) are repeated to create a new window in the SiO_2.
9. A second diffusion creates the p-type region in the n-type base by using boron as the dopant.
10. Silicon dioxide (SiO_2) is again grown over the exposed silicon wafer, and the photoresist is applied over the SiO_2.
11. The photomask, containing the two clearances for the emitter and base, is placed over the positive photoresist, and steps (2) through (6) are repeated.
12. The structure is now ready for metallization. An aluminum film is deposited over the entire surface, followed by a coating of positive photoresist.
13. The photomask, with the emitter and base areas opaque, is placed over the photoresist and exposed to UV light.
14. The photoresist is developed, leaving the resist over the emitter and base areas.
15. The exposed metallization is etched away, followed by the removal of the resist over the emitter and base areas.
16. A passivation layer of silicon nitride is applied to the circuitry, leaving the bond pads exposed.
17. The silicon planar bipolar transistor is now complete.

References

1. F. P. Tedeschi and M. R. Taber, *Solid-State Electronics,* Van Nostrand Reinhold, 1976.
2. S. Levine, *Discrete Semiconductors and Optoelectronics,* Vol. 2, Electro-Horizons Publications, 1987.
3. C. Harper, ed., *Electronic Packaging and Interconnection Handbook,* 3rd ed., McGraw-Hill, 2000, Chap. 7.
4. T. R. Reid, *The Chip,* Simon and Schuster, 1984.
5. S. Wolf and R. N. Tauber, *Silicon Processing for the VLSI Era*, Vol. 1, Lattice Press, 1986.
6. P. V. Zant, *Microchip Fabrication,* 4th ed., McGraw-Hill, 2000.