1 / An Overview of Microelectronic Fabrication

1.1 A HISTORICAL PERSPECTIVE

In this volume we will develop an understanding of the basic processes used in monolithic integrated-circuit fabrication. Silicon is the dominant material used throughout the integrated-circuit industry today, and in order to conserve space only silicon processing will be discussed in this book. However, all of the basic processes discussed here are applicable to the fabrication of gallium arsenide integrated circuits (ICs) and thick- and thin-film hybrid ICs.

Germanium was one of the first materials to receive wide attention for use in semiconductor device fabrication, but it was rapidly replaced by silicon during the early 1960s. Silicon emerged as the dominant material because it was found to have major processing advantages. Silicon can easily be oxidized to form silicon dioxide. Silicon dioxide was found to be not only a high-quality insulator but also an excellent barrier layer for the selective diffusion steps needed in integrated-circuit fabrication.

Silicon was also shown to have a number of ancillary advantages. It is a very abundant element in nature, providing the possibility of a low-cost starting material. It has a wider bandgap than germanium and can therefore operate at higher temperatures than germanium. In retrospect it appears that the processing advantages were the dominant reasons for the emergence of silicon over other semiconductor materials.

The first successful fabrication techniques produced single transistors on a silicon die 1 to 2 mm on a side. Early integrated circuits, fabricated at Texas Instruments and Fairchild Semiconductor, included several transistors and resistors to make simple logic gates and amplifier circuits. From this modest beginning, we have reached integration levels of several million components on a 7 mm $\times$ 7 mm die. For example, a one-megabit dynamic random-access memory (DRAM) chip has more than 1,000,000 transistors and more than 1,000,000 capacitors in the memory array, as well as tens of thousands of transistors in the access and decoding circuits. The level of integration has been doubling every one to two years since the early 1960s.

One-megabit RAMs are currently being produced with photolithographic features measuring between 1 and 2 microns (μm). MOS transistors with dimensions approximately ten times smaller (0.1 to 0.2 μm) have already been fabricated in research laboratories. So we still have at least a factor of 100 to go in terms of integrated-circuit density, provided manufacturable fabrication processes can be developed for these submicron dimensions.

Early fabrication used silicon wafers which had 1- and then 2-in. diameters. The size of the wafers has steadily increased to the point where 4-, 5-, and 6-in. wafers are now in production. Wafers with 8-in. diameters have been successfully produced by silicon wafer manufacturers.

The larger the diameter of the wafer, the more integrated-circuit dice can be produced at one time. Many wafers are processed at the same time. The same silicon chip is replicated as many times as possible on a silicon wafer of a given size. Figure 1.1 shows the approximate number of 5 × 5 mm dice that fit on a wafer of a given diameter. Processing costs per wafer are relatively independent of wafer size, so the cost per die is lower for larger wafer sizes. Thus there are strong economic forces driving the integrated-circuit industry to continually move to larger and larger wafer sizes.

Here we see a problem with the units of measure used to describe integrated circuits. Horizontal dimensions were originally specified in mils (1 mil = 0.001 in.), whereas specification of the shallower vertical dimensions commonly made use of the metric system. Today, most of the dimensions are specified using the metric system, although English units are occasionally still used. Throughout the rest of this book, we will attempt to make consistent use of metric units.

1.2 AN OVERVIEW OF MONOLITHIC FABRICATION PROCESSES AND STRUCTURES

Monolithic integrated-circuit fabrication can be illustrated by studying the basic cross sections of MOS and bipolar transistors in Figs. 1.2 and 1.3. The *n*-channel MOS transistor is formed in a *p*-type substrate. Source/drain regions are formed by selectively converting shallow regions at the surface to *n*-type material. Thin and thick silicon dioxide regions on the surface form the gate insulator of the transistor and serve to isolate one device from another. A thin film of polysilicon is used to form the gate of the transistor, and aluminum is used to make contact to the source and drain. Interconnections between devices can be made using the diffusions and the layers of polysilicon and metal.

The bipolar transistor has alternating *n*- and *p*-type regions selectively fabricated on a *p*-type substrate. Silicon dioxide is again used as an insulator, and a metal such as aluminum is used to make electrical contact to the emitter, base, and collector of the transistor.

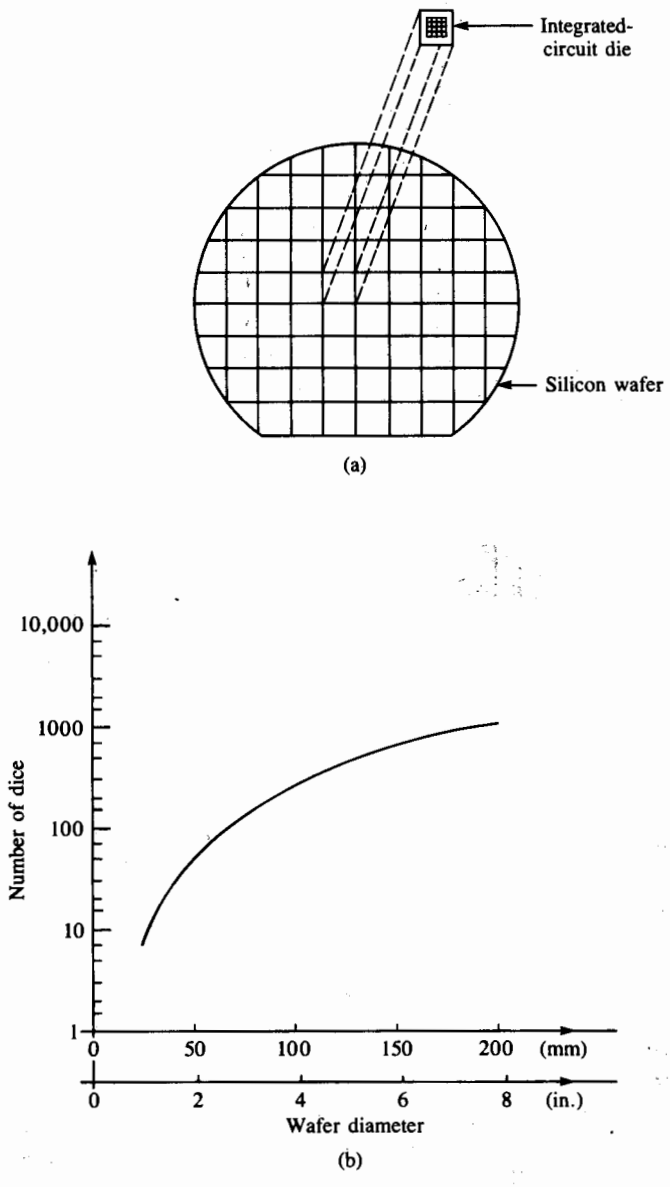

Fig. 1.1 (a) The same integrated-circuit die is replicated hundreds of times on a typical silicon wafer; (b) the graph gives the approximate number of 5 × 5 mm dice which can be fabricated on wafers of different diameters.

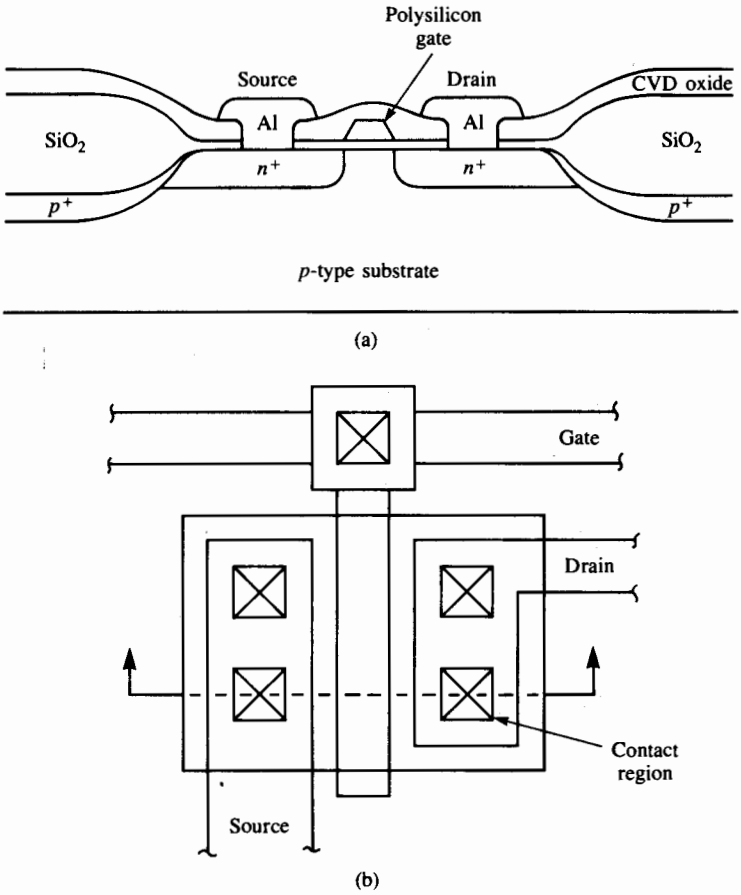

Fig. 1.2 The basic structure of an *n*-channel metal-oxide-semiconductor (NMOS) transistor structure. (a) The vertical cross section through the transistor; (b) a composite top view of the masks used to fabricate the transistor in (a).

These structures are fabricated through the repeated application of a number of basic processing steps:

- Oxidation
- Photolithography
- Etching
- Diffusion
- Evaporation or sputtering
- Chemical vapor deposition (CVD)
- Ion implantation
- Epitaxy

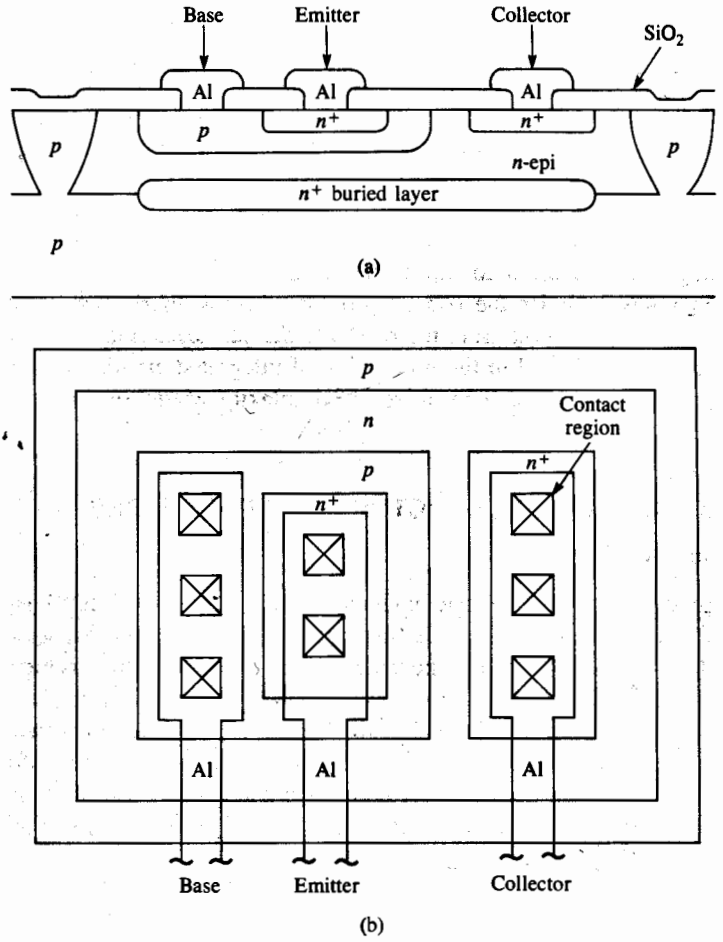

Fig. 1.3 The basic structure of a junction-isolated bipolar transistor. (a) The vertical cross section through the transistor; (b) a composite top view of the masks used to fabricate the transistor in (a).

Silicon dioxide can be formed by heating a silicon wafer to a high temperature (1000 to 1200 °C) in the presence of oxygen. This process is called *oxidation*. Metal films can be deposited through evaporation by heating the metal to its melting point in a vacuum. Thin films of silicon nitride, silicon dioxide, and polysilicon can all be formed through a process known as *chemical vapor deposition* (CVD), in which the material is deposited out of a gaseous mixture onto the surface of the wafer. Metals and insulators may also be deposited by a process called *sputtering*.

Shallow *n*- and *p*-type layers are formed by high-temperature (1000 to 1200 °C) *diffusion* of donor or acceptor impurities into silicon or by *ion implantation*, in which the

wafer is bombarded with high-energy donor or acceptor atoms generated in a high-voltage particle accelerator.

In order to build devices and circuits, the *n*- and *p*-type regions must be formed selectively in the surface of the wafer. Silicon dioxide, silicon nitride, polysilicon, and other materials can all be used to mask areas of the wafer surface to prevent penetration of impurities during ion implantation or diffusion. Windows are cut in the masking material by etching with acids or in a plasma. Window patterns are transferred to the wafer surface from a mask through the use of optical techniques. The masks are also produced using photographic reduction techniques.

Photolithography includes the overall process of mask fabrication as well as the process of transferring patterns from the masks to the surface of the wafer. The photolithographic process is critical to the production of integrated circuits, and the number of mask steps is often used as a measure of complexity when comparing fabrication processes.

1.3 METAL-OXIDE-SEMICONDUCTOR (MOS) PROCESSES

1.3.1 Basic NMOS Process

A possible process flow for a basic *n*-channel MOS process (NMOS) is shown in Figs. 1.4 and 1.5. The starting wafer is first oxidized to form a thin-pad oxide layer of silicon dioxide (SiO_2) which protects the silicon surface. Silicon nitride is then deposited by a low-pressure chemical vapor deposition (LPCVD) process. Mask #1 defines the active transistor areas. The nitride/oxide sandwich is etched away everywhere except where transistors are to be formed. A boron implantation is performed and followed by an oxidation step. The nitride serves as both an implantation mask and an oxidation mask. After the nitride and thin oxide padding layers are removed, a new thin layer of oxide is grown to serve as the gate oxide for the MOS transistors. Following gate-oxide growth, a boron implantation is commonly used to adjust the threshold voltage to the desired value.

Polysilicon is deposited over the complete wafer using a CVD process. The second mask defines the polysilicon gate region of the transistors. Polysilicon is etched away everywhere except over the gate regions and the areas used for interconnection. Next, the source/drain regions are implanted through the thin oxide regions. The implanted impurity may be driven in deeper with a high-temperature diffusion step. More oxide is deposited on the surface, and contact openings are defined by the third mask step. Metal is deposited over the wafer surface by evaporation or sputtering. The fourth mask step is used to define the interconnection pattern which will be etched in the metal. A passivation layer of phosphosilicate glass (not shown in Fig. 1.4) is deposited on the wafer surface, and the final mask (#5) is used to define windows so that bonding wires can be attached to pads on the periphery of the IC die.

This simple process requires five mask steps. Note that these mask steps use subtractive processes. The entire surface of the wafer is first coated with a desired material, and then most of the material is removed by wet chemical or plasma etching.

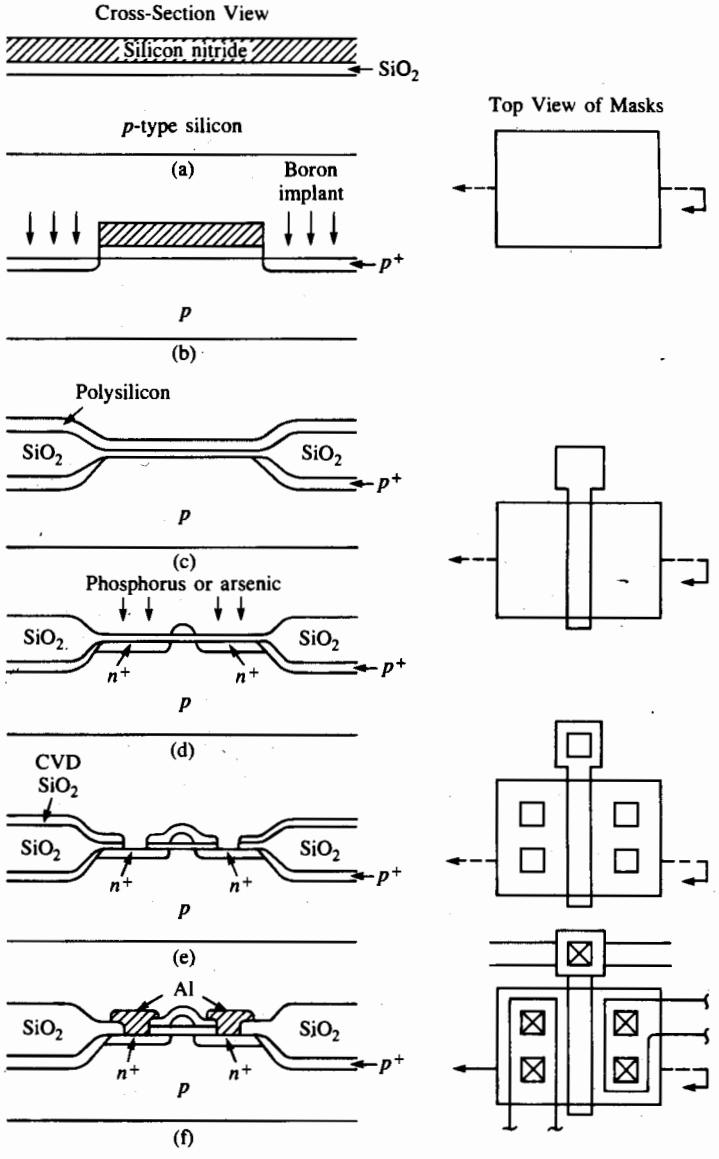

Fig. 1.4 Process sequence for a semirecessed oxide NMOS process. (a) Silicon wafer covered with silicon nitride over a thin padding layer of silicon dioxide; (b) etched wafer after first mask step. A boron implant is used to help control field oxide threshold; (c) structure following nitride removal and polysilicon deposition; (d) wafer after second mask step and etching of polysilicon; (e) the third mask has been used to open contact windows following silicon dioxide deposition; (f) final structure following metal deposition and patterning with fourth mask.

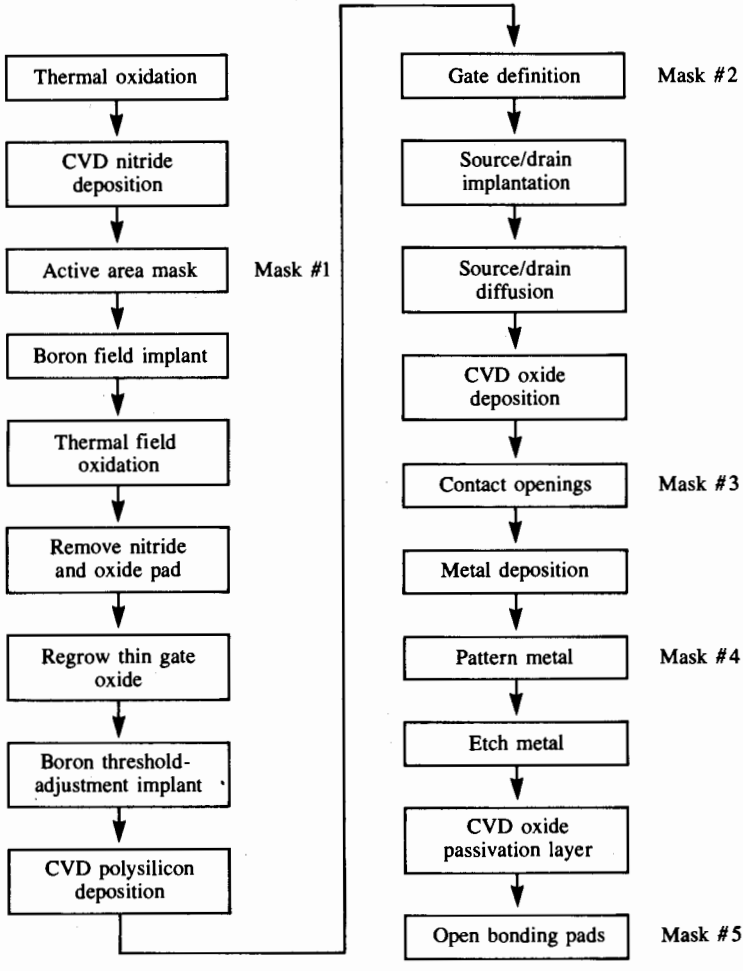

Fig. 1.5 Basic NMOS process flowchart.

1.3.2 Basic Complementary MOS (CMOS) Process

Figure 1.6 shows the mask sequence for a basic complementary MOS (CMOS) process. One new mask, beyond that of the NMOS process, is used to define the "*p*-well" or "*p*-tub," which serves as the substrate for the *n*-channel devices. A second new mask step is used to define the source/drain regions of the *p*-channel transistors. Additional masks may be used to adjust the threshold voltage of the MOS transistors and are very common in state-of-the-art NMOS and CMOS processes.

Some recent CMOS processes use an *n*-well instead of a *p*-well. The *n*-well can be added to an existing NMOS process with a minimum of change, and it permits high-

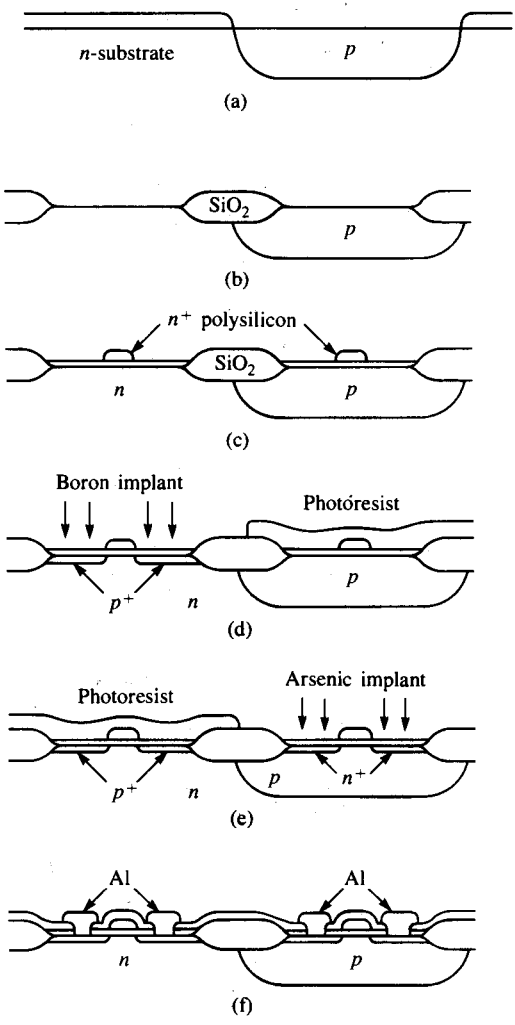

Fig. 1.6 Cross-sectional views at major steps in a basic CMOS process. (a) Following *p*-well diffusion, (b) following selective oxidation, and (c) following gate oxidation and polysilicon gate definition; (d) PMOS source/drain implantation; (e) NMOS source/drain implantation; (f) structure following contact and metal mask steps.

performance NMOS and CMOS on the same chip. Twin-well processes have also been developed recently. Both a *p*-well and an *n*-well are formed in a lightly doped substrate, and the *n*- and *p*-channel devices can each be optimized for highest performance. Twin-well very large-scale integration (VLSI) processes use lightly doped layers grown on heavily doped substrates to suppress a CMOS failure mode called *latchup*.

1.4 BASIC BIPOLAR PROCESSING

Basic bipolar fabrication is somewhat more complex than single-channel MOS processing, as indicated in Figs. 1.7 and 1.8. A *p*-type silicon wafer is oxidized, and the first

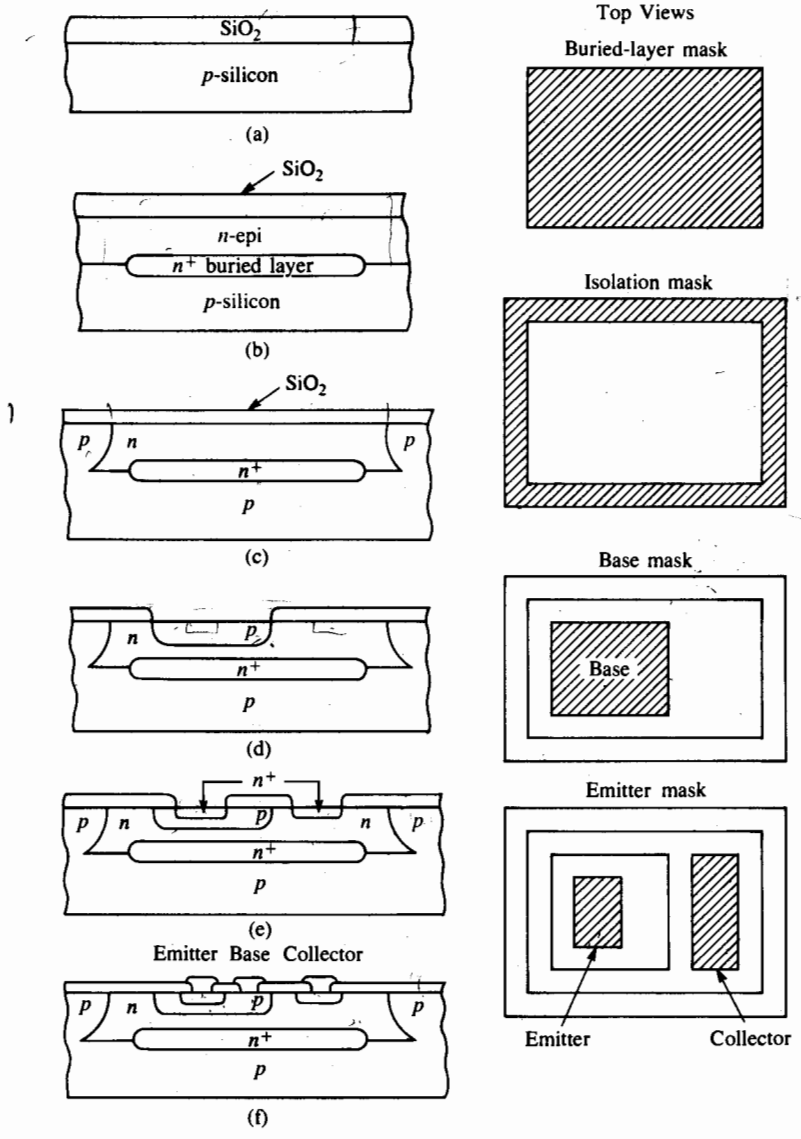

Fig. 1.7 Cross-sectional view of the major steps in a basic bipolar process. (a) Wafer with silicon dioxide layer; (b) following buried-layer diffusion using first mask, and subsequent epitaxial layer growth and oxidation; (c) following deep-isolation diffusion using second mask; (d) following boron-base diffusion using third mask; (e) fourth mask defines emitter and collector contact regions; (f) final structure following contact and metal mask steps.

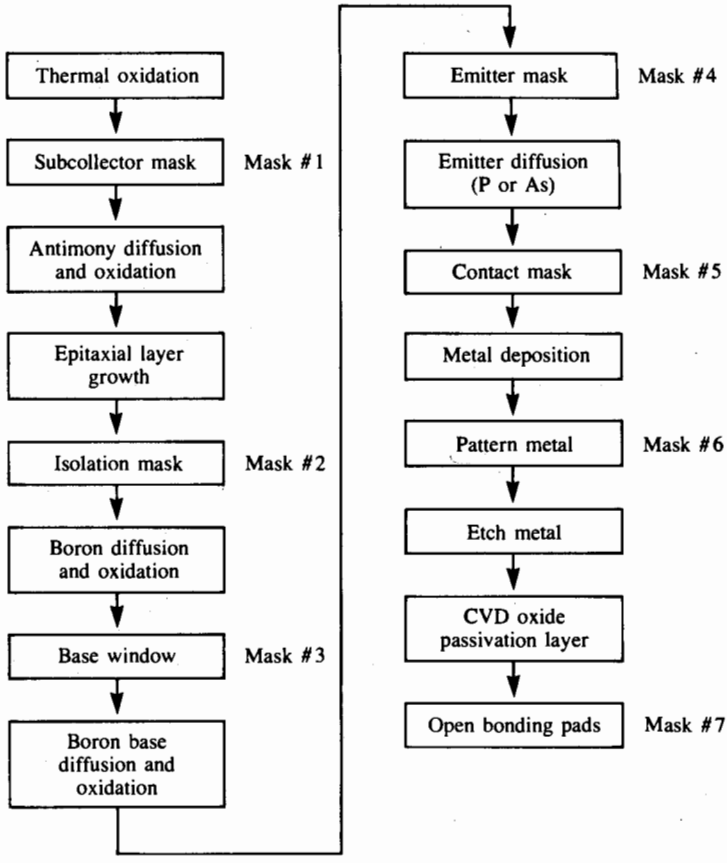

Fig. 1.8 Basic bipolar process flowchart.

mask is used to define a diffused region called the *buried layer* or *subcollector*. This diffusion is used to reduce the collector resistance of the bipolar transistor. Following the buried-layer diffusion, a process called *epitaxy* is used to grow single-crystal *n*-type silicon on top of the silicon wafer. The epitaxial growth process results in a high-quality silicon layer with the same crystal structure as the original silicon wafer. An oxide layer is then grown on the wafer. Mask two is used to open windows for a deep *p*-diffusion, which is used to isolate one bipolar transistor from another. Another oxidation follows the isolation diffusion. Mask three opens windows in the oxide for the *p*-type base diffusion. The wafer is usually oxidized during the base diffusion, and mask four is used to open windows for the emitter diffusion. The same diffusion step places an n^+ region under the collector contact to ensure that a good ohmic contact will be formed during subsequent metallization. Masks five, six, and seven are used to open contact windows, pattern the metallization layer, and open windows in the passivation layer just as in the NMOS process described in Section 1.3. Thus the basic bipolar process requires seven mask levels compared with five for the basic NMOS process.

After the MOS or bipolar process is completed, each die on the wafer is tested, and bad dice are marked with ink. The wafer is then sawed apart. Good dice are mounted in various packages for final testing and subsequent sale or use.

The rest of this book concentrates on the basic processes used in the fabrication of monolithic integrated circuits. Chapters 2 through 8 discuss mask making and pattern definition, oxidation, diffusion, ion implantation, film deposition, interconnections and contacts, and packaging and yield. The last two chapters introduce the integration of process, layout, and device design for MOS and bipolar technologies.

PROBLEMS

1.1 The curve in Fig. 1.1b represents the approximate number of chips on a wafer of a given diameter. Determine the exact number of 5 × 5 mm dice that will fit on a wafer with a diameter of 100 mm. (The number indicated on the curve is 254.)

1.2 The cost of processing a wafer in a particular process is $400. Supposing that 35% of the dice fabricated are good, find the number of dice, using Fig. 1.1b.

(a) Determine the cost per good die for a 75-mm wafer.

(b) Repeat for a 150-mm wafer.

1.3 A certain silicon-gate NMOS transistor occupies an area of 25 λ^2 where λ is the minimum lithographic feature size.

(a) How many MOS transistors can fit on a 5 × 5 mm die if $\lambda = 10$ μm?

(b) 2.5 μm?

(c) 1 μm?

1.4 A simple *pn* junction diode is shown in cross section in Fig. P1.4. Make a possible process flowchart for fabrication of this structure, including mask steps.

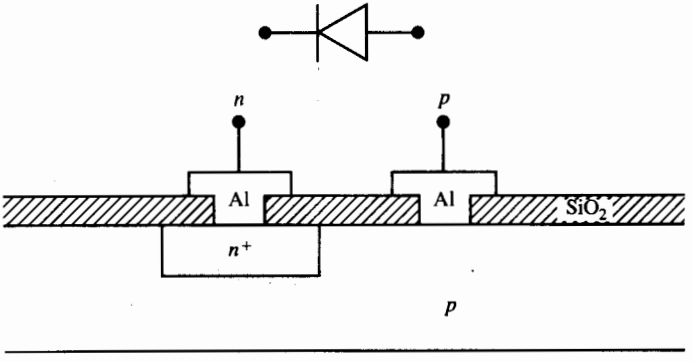

Fig. P1.4

1.5 Draw a set of contact and metal masks for the bipolar transistor of Fig. 1.7. Use square contact windows with one contact to the emitter and two contacts to the base and collector regions.