The book is designed to be covered in one semester. On the quarter we cover Chapters 1–5 and most of the material from Chapters 9 and 10 on MOS and bipolar process integration have been included.

The text chip of Fig. 2.7 was fabricated during this laboratory, and a few of the problems are related to structures on this chip. We can provide pattern generator data for this flyway fabricated by previous classes on an "as-available" basis. A number of previous books which have obviously influenced the text. These include *The Theory and Practice of Microelectronics in Principles* by S. K. Ghandhi, *Basic Integrated Circuit Engineering* and W. G. Howard, *Integrated Circuit Engineering* by A. H. abak Sharpe, *Microelectronics Processing and Device Design* by H. L. Semiconductor Devices—Physics and Technology* by S. M. Sze. ly for putting up with the countless hours of work which have gone into this book. In particular, I want to thank my wife Joan for the many in the library tracking down and verifying errant references. During the writing, we found that the references in many popular, recently published texts are useless because they contained so many errors. We have done our best to keep this from happening.

To my colleagues who have helped with this book, especially for his suggestions, and to our laboratory managers, Walter Power who have been instrumental in developing the processes for the ss chips.

Richard C. Jaeger

## Contents

1. An Overview of Microelectronic Fabrication
   1.1 A Historical Perspective
   1.2 An Overview of Monolithic Fabrication Processes and Structures
   1.3 Metal-Oxide-Semiconductor (MOS) Processes
      1.3.1 Basic NMOS Process
      1.3.2 Basic Complementary MOS (CMOS) Process
   1.4 Basic Bipolar Processing
   1.5 Problems

2. Lithography
   2.1 The Photolithographic Process
      2.1.1 Wafer Cleaning
      2.1.2 Barrier Layer Formation
      2.1.3 Photoresist Application
      2.1.4 Soft Baking
      2.1.5 Mask Alignment
      2.1.6 Photoresist Exposure and Development
      2.1.7 Hard Baking
      2.2 Etching Techniques
      2.2.1 Wet Chemical Etching
      2.2.2 Dry Etching
      2.2.3 Photoresist Removal
   2.3 Photomask Fabrication
   2.4 Summary
   References
   Further Reading
   Problems
9 MOS Process Integration

9.1 Basic MOS Device Considerations
9.1.1 Gate-Oxide Thickness
9.1.2 Substrate Doping and Threshold Voltage
9.1.3 Threshold Adjustment
9.1.4 Field-Region Considerations

9.2 MOS Transistor Layout and Design Rules
9.2.1 Metal-Gate Transistor Layout
9.2.2 Polysilicon-Gate Transistor Layout
9.2.3 More-Aggressive Design Rules
9.2.4 Channel Length and Width Biases

9.3 Complementary MOS (CMOS) Technology

9.4 Other MOS Structures
9.5 Summary
9.6 References
9.7 Problems

10 Bipolar Process Integration

10.1 The Junction-Isolated Structure
10.2 Current Gain
10.3 Transit Time
10.4 Basewidth
10.5 Breakdown Voltages
10.5.1 Emitter-Base Breakdown Voltage
10.5.2 Collector-Base Breakdown Voltage
10.6 Layout Considerations
10.6.1 Buried-Layer and Isolation Diffusions
10.6.2 Base Diffusion to Isolation Diffusion Spacing
10.6.3 Emitter-Diffusion Design Rules
10.6.4 A Layout Example

10.7 Advanced Bipolar Structures
10.8 Other Bipolar Isolation Techniques
10.8.1 Collector-Diffusion Isolation (CDI)
10.8.2 V-Groove Isolation
10.8.3 Dielectric Isolation

10.9 Summary
10.10 References
10.11 Problems

Answers to Selected Problems

Index
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 × 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.
negabit RAMs are currently being produced with photolithographic features between 1 and 2 microns (μm). MOS transistors with dimensions approximately smaller than 0.1 to 0.2 μm have already been fabricated in researches. So we still have at least a factor of 100 to go in terms of integrated-circuit manufacturable fabrication processes can be developed for these dimensions.

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

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

we see a problem with the units of measure used to describe integrated circuits. Dimensions were originally specified in mils (1 mil = 0.001 in.), whereas today, most of the dimensions are specified using the metric system, although mils are occasionally still used. Throughout the rest of this book, we will make consistent use of metric units.

OVERVIEW OF MONOLITHIC FABRICATION PROCESSES AND URES

Integrated-circuit fabrication can be illustrated by studying the basic cross section of MOS and bipolar transistors in Figs. 1.2 and 1.3. The n-channel MOS is formed in a p-type substrate. Source/drain regions are formed by selectively etching shallow regions at the surface to n-type material. Thin and thick silicon oxide films on the surface form the gate insulator of the transistor and serve to isolate it from another. A thin film of polysilicon is used to form the gate of the p-type and aluminum is used to make contact to the source and drain. Interiors between devices can be made using the thin films and the layers of polysilicon.

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

![Integrated-circuit die diagram](image)

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 x 5 mm dice which can be fabricated on wafers of different diameters.
The basic structure of an n-channel metal-oxide-semiconductor (NMOS) transistor. (a) The vertical cross section through the transistor; (b) a composite top view of the mask to fabricate the transistor in (a).

These structures are fabricated through the repeated application of a number of basic 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
is bombarded with high-energy donor or acceptor atoms generated in a high-energy particle accelerator.

In order to build devices and circuits, the n- and p-type regions must be formed separately in the surface of the wafer. Silicon dioxide, silicon nitride, polysilicon, and materials can all be used to mask areas of the wafer surface to prevent penetration purities 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 surface from a mask through the use of optical techniques. The masks are also made using photographic reduction techniques.

photolithography includes the overall process of mask fabrication as well as the steps 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 steps is often used as a measure of complexity when comparing fabrication processes.

**METAL-OXIDE-SEMICONDUCTOR (MOS) PROCESSES**

**Basic NMOS Process**

The possible process flow for a basic n-channel MOS process (NMOS) is shown in Figures 1.4 and 1.5. The starting wafer is first oxidized to form a thin-pad oxide layer of silicon dioxide (SiO₂) which protects the silicon surface. Silicon nitride is then deposited low-pressure chemical vapor deposition (LPCVD) process. Mask #1 defines the 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 paddle layers are removed, a new thin layer of silicon is grown to serve as the gate oxide for the MOS transistors. Following gate-oxide deposition, a boron implantation is commonly used to adjust the threshold voltage to the desired value.

Silicon is deposited over the complete wafer using a CVD process. The second step is the polysilicon gate region of the transistors. Polysilicon is etched away where except over the gate regions and the areas used for interconnection. Next, source/drain regions are implanted through the thin oxide regions. The implanted ions may be driven in deeper with a high-temperature diffusion step. More oxide is added over the surface, and contact openings are defined by the third mask step. Metal is evaporated 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 thin layer of phosphosilicate glass (not shown in Fig. 1.4) is deposited on the 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 sub-micron processes. The entire surface of the wafer is first coated with a desired material, 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.
2 Basic Complementary MOS (CMOS) Process

Fig. 1.5 shows the mask sequence for a basic complementary MOS (CMOS) process. A new mask, beyond that of the NMOS process, is used to define the "p-well" or "n-well," which serves as the substrate for the n-channel devices. A second new mask is used to define the source/drain regions of the p-channel transistors. Additional masks are 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-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.

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.
BASIC BIPOLAR PROCESSING

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

![Diagram of basic bipolar process](image)

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

![Flowchart of basic bipolar process](image)

**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.
AN OVERVIEW OF MICROELECTRONIC FABRICATION

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

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

BLEMMS

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

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

Perform the cost per good die for a 75-mm wafer.

A certain silicon-gate NMOS transistor occupies an area of 25 \( \lambda^2 \) square microns. Find \( \lambda \) for a given MOS transistor on a 5 x 5 mm die if \( \lambda = 10 \, \mu \text{m} \)?

How many MOS transistors can fit on a 5 x 5 mm die if \( \lambda = 10 \, \mu \text{m} \)?

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

2 / Lithography

In order to produce an integrated circuit, thin films of various materials are used as barriers to the diffusion or implantation of impurity atoms, or as insulators between conductive materials and the silicon substrate. Holes or windows are cut through this barrier material in order to retrieve impurity penetration or contact is desired.

Masks contain the patterns of windows which are transferred to the surface of the silicon wafer using a process called photolithography. Photolithography makes use of a highly refined version of the photogradning process. The patterns are first transferred from the mask to a light-sensitive material called photoresist. Chemical or plasma etching is then used to transfer the pattern from the photoresist to the barrier material on the surface of the wafer. Each mask step requires successful completion of numerous processing steps, and the complexity of an integrated-circuit process is often measured by the number of photomask used during fabrication. This chapter will explore the lithographic process, including mask fabrication, photoresist processes, and etching.

2.1 THE PHOTOLITHOGRAPHIC PROCESS

Photolithography encompasses all the steps involved in transferring a pattern from a mask to the surface of the silicon wafer. The various steps of the basic photolithographic process given in Figs. 2.1 and 2.2 will each be discussed in detail below.

Ultraclean conditions must be maintained during the lithography process. Any dust particles on the original substrate or that fall on the substrate during processing can result in defects in the final resist coating. Even if defects occur in only 10% of the chip sites at each mask step, less than 50% of the chips will be functional after a seven-mask process is completed. Vertical laminar-flow hoods in clean rooms are used to prevent particulate contamination throughout the fabrication process. Clean rooms use filtration to remove particles from the air and are rated by the maximum number of particles per cubic foot or cubic meter of air, as shown in Table 2.1. Clean rooms have evolved from the Class 10,000 to the Class 10 and even Class 1 facilities now being used for VLSI processing. For comparison, each cubic foot of ordinary room air has several million dust particles exceeding a size of 0.5 \( \mu \text{m} \).
1 Wafer Cleaning

To use, wafers are chemically cleaned to remove particulate matter on the surface as well as any traces of organic, ionic, and metallic impurities. A cleaning step in a solution of hydrofluoric acid is used to remove any oxide which may have formed on the surface. A typical cleaning process is presented in Table 2.2.

One very important chemical used in wafer cleaning and throughout microelectronic fabrication processes is deionized (DI) water. DI water is highly purified and filtered to Fig. 2.2 Drawings of wafer through various steps of the photolithographic process. (a) Substrate covered with silicon dioxide barrier layer; (b) positive photosensit applied to the surface of the wafer; (c) mask in close proximity to the surface of the resist-covered wafer; (d) substrate following resist exposure and development; (e) substrate following etching of the silicon dioxide layer; (f) oxide barrier on wafer surface after resist removal; (g) view of substrate with silicon dioxide pattern on the surface.

Fig. 2.1 Steps of the photolithographic process.
2.1 Ratings by Class of Effectiveness of Filtration in Clean Rooms.

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of 0.5-μm particles per ft² (m²)</th>
<th>Number of 5-μm particles per ft² (m²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>10,000 (350,800)</td>
<td>65 (23,000)</td>
</tr>
<tr>
<td>000</td>
<td>1,000 (35,000)</td>
<td>6.5 (230)</td>
</tr>
<tr>
<td>100</td>
<td>100 (3,100)</td>
<td>0.65 (23)</td>
</tr>
<tr>
<td>10</td>
<td>10 (350)</td>
<td>0.0065 (2.3)</td>
</tr>
<tr>
<td>1</td>
<td>1 (35)</td>
<td>0.00065 (2.3)</td>
</tr>
</tbody>
</table>

is very difficult to measure particulate counts below 10 per ft².

2.2 Silicon Wafer Cleaning Procedure

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.</td>
<td>Solvent Removal</td>
</tr>
<tr>
<td></td>
<td>1. Immerse in boiling trichloroethylene (TCE) for 3 min.</td>
</tr>
<tr>
<td></td>
<td>2. Immerse in boiling acetone for 3 min.</td>
</tr>
<tr>
<td></td>
<td>3. Immerse in boiling methyl alcohol for 3 min.</td>
</tr>
<tr>
<td></td>
<td>4. Wash in DI water for 3 min.</td>
</tr>
<tr>
<td>B.</td>
<td>Removal of Residual Organic/Ionic Contamination</td>
</tr>
<tr>
<td></td>
<td>1. Immerse in a (5:1:1) solution of H₂O–NH₃OH–H₂O₃, heat solution to 75–80 °C and hold for 10 min.</td>
</tr>
<tr>
<td></td>
<td>2. Quench the solution under running DI water for 1 min.</td>
</tr>
<tr>
<td></td>
<td>3. Wash in DI water for 5 min.</td>
</tr>
<tr>
<td>C.</td>
<td>Hydrous Oxide Removal</td>
</tr>
<tr>
<td></td>
<td>1. Immerse in a (1:50) solution of HF–H₂O for 15 sec.</td>
</tr>
<tr>
<td></td>
<td>2. Wash in running DI water with agitation for 30 sec.</td>
</tr>
<tr>
<td>D.</td>
<td>Heavy Metal Clean</td>
</tr>
<tr>
<td></td>
<td>1. Immerse in a (6:1:1) solution of H₂O–HCl–H₂O₃ for 10 min at a temperature of 75–80 °C.</td>
</tr>
<tr>
<td></td>
<td>2. Quench the solution under running DI water for 1 min.</td>
</tr>
<tr>
<td></td>
<td>3. Wash in running DI water for 20 min.</td>
</tr>
</tbody>
</table>

2.1.3 Photoresist Application

After formation of the SiO₂ layer, the surface of the wafer is coated with a light-sensitive material called photoresist. The surface must be clean and dry to ensure good photoresist adhesion. Freshly oxidized wafers may be directly coated, but if the wafers have been stored, they should be carefully cleaned and dried prior to application of the resist. A liquid adhesion promoter is often applied just prior to resist application.

Photoresist is typically applied in liquid form. The wafer is held on a vacuum chuck and then spun at high speed for 30 to 60 sec to produce a thin uniform layer. Speeds of 1000 to 5000 rpm result in layers ranging from 2.5 to 0.5 μm, respectively. The actual thickness of the resist depends on its viscosity and is inversely proportional to the square root of the spinning speed.

2.1.4 Soft Baking

A drying step called soft baking or prebaking is used to improve adhesion and remove solvent from the photoresist. Times range from 10 to 30 min in an oven at 80 to 90 °C in an air or nitrogen atmosphere. The soft-baking process is specified on the resist manufacturer’s data sheet and should be followed closely. After soft baking, the photoresist is ready for mask alignment and exposure.

2.1.5 Mask Alignment

A photomask, a square glass plate with a patterned emulsion or metal film on one side, is placed over the wafer. Each mask following the first must be carefully aligned to the previous pattern on the wafer. Much of the alignment has traditionally involved manual operation of alignment equipment. VLSI designs with minimum-size geometrical features measuring 1.25 μm (minimum linewidth or space) require an alignment tolerance of better than ±0.25 μm. Computer-controlled alignment equipment has been developed to achieve this level of alignment precision.

With manual alignment equipment, the wafer is held on a vacuum chuck and carefully moved into position below the mask using an adjustable x-y stage. The mask is
spaced 25 to 125 \( \mu \text{m} \) above the surface of the wafer during alignment. If contact printing is being used, the mask is brought into contact with the wafer after alignment.

Alignment marks are introduced on each mask and transferred to the wafer as part of the integrated-circuit pattern. The marks are used to align each new mask level to one of the previous levels. A sample set of alignment marks is shown in Fig. 2.3. For certain mask levels, the cross on the mask is placed in a box on the wafer. For other mask levels, the box on the mask is placed over a cross on the wafer. The choice depends on the type of resist used during a given photolithographic step. Split-field optics are used to simultaneously align two well-separated areas of the wafer.

2.1.6 Photoresist Exposure and Development

Following alignment, the photoresist is exposed through the mask with high-intensity ultraviolet light. Resist is exposed wherever silicon dioxide is to be removed. The photoresist is developed with a process very similar to that used for developing ordinary photographic film, using a developer supplied by the photoresist manufacturer. Any resist which has been exposed to ultraviolet light is washed away, leaving bare silicon dioxide in the exposed areas of Fig. 2.2d. A photoresist acting in the manner just described is called a positive resist, and the mask contains a copy of the pattern which will remain on the surface of the wafer. Windows are opened wherever the exposing light passes through the mask.

Negative photoresists can also be used. A negative resist remains on the surface wherever it is exposed. Figure 2.4 shows simple examples of the patterns transferred to a silicon dioxide barrier layer using positive and negative photoresists with the same mask. Negative resists were widely used in early integrated-circuit processing. However, positive resist yields better process control in small-geometry structures and is now the main type of resist used in VLSI processes.

Fig. 2.3 A simple set of alignment marks. At some steps a cross may be aligned within a box. At others, a box may be placed around the cross. The choice depends on the type of resist being used at a given mask step.

Fig. 2.4 Resist and silicon dioxide patterns following photolithography with positive and negative resists.
.7 Hard Baking
Low exposure and development, a baking step is used to harden the photoresist and promote adhesion to the substrate. A typical process involves baking in an oven for 20 to 30 min at 120 to 180 °C. Details of this step are again specified on the manufacturer's photoresist data sheets.

ETCHING TECHNIQUES

Chemical etching in liquid or gaseous form is used to remove any barrier material not masked by hardened photoresist. The choice of chemicals depends on the material to be etched. A high degree of selectivity is required so that the etchant will remove the protected barrier layer much more rapidly than it attacks the photoresist layer.

.1 Wet Chemical Etching

Suffred oxide etch (BOE or BHF) is commonly used to etch windows in silicon dioxide layers. BOE is a solution containing hydrofluoric acid (HF), and etching is often achieved by immersing the wafers in the solution. At room temperature, HF etches silicon dioxide much more rapidly than it etches photoresist or silicon. The etch rate in Si ranges from 10 to 100 nm/min at 25 °C, depending on the density of the silicon film. Etch rate is temperature-dependent, and temperature is carefully monitored during the etching process. In addition, etch rates depend on the type of oxide present. Films grown in dry oxygen etch more slowly than those grown in the presence of water vapor. A high concentration of phosphorus in the oxide enhances the etch rate, whereas doped etch rates occur when a high concentration of boron is present. High concentrations of these elements convert the SiO2 layer to a phosphosilicate or borosilicate glass.

HF and water both wet silicon dioxide but do not wet silicon. The length of the etch may be controlled by visually monitoring test wafers which are etched along with actual integrated-circuit wafers. Occurrence of a hydrophobic condition on the wafer signals completion of the etch step.

Wet chemical etching tends to be an isotropic process, etching equally in all directions. Figure 2.5a shows the result of isotropic etching of a narrow line in silicon dioxide. The etching process has etched under the resist by a distance equal to the thickness of the film. This "etch bias" becomes a serious problem in processes requiring linewidths with dimensions similar to the thickness of the film.

2.3.2 Dry Etching

Dry etching processes are widely used in VLSI fabrication. Highly anisotropic etching profiles can be obtained as shown in Fig. 2.5b, avoiding the undercutting problem of Fig. 2.5a characteristic of wet processes. Dry processes require only small amounts of reactant gases, whereas wet etching requires disposal of relatively large amounts of liquid chemical wastes.

Plasma etching immerses the wafers in a gaseous plasma created by RF excitation in a vacuum system. The plasma contains fluorine or chlorine ions which etch silicon dioxide. The RF power source typically operates at a frequency of 13.56 MHz, which is set aside by the Federal Communications Commission for industrial and scientific purposes.

Sputter etching uses energetic noble gas ions such as Ar+ to bombard the wafer surface. Etching occurs by physically knocking atoms off the surface of the wafer. Highly anisotropic etching can be obtained, but selectivity is often poor. Metals can be used as barrier materials to protect the wafer from etching.

Reactive-ion etching combines the plasma and sputter etching processes. Plasma systems are used to ionize reactive gases, and the ions are accelerated to bombard the surface. Etching occurs through a combination of the chemical reaction and momentum transfer from the etching species.

2.2.3 Photoresist Removal

After windows are etched through the SiO2 layer, the photoresist is stripped from the surface, leaving a window in the silicon dioxide. Photoresist removal typically uses proprietary-liquid resist strippers, which cause the resist to swell and lose adhesion to the substrate. Dry processing may also be used to remove resist by oxidizing (burning) it in an oxygen plasma system, a process often called resist ashing.

2.3 PHOTOMASK FABRICATION

Photomask fabrication involves a series of photographic processes outlined in Fig. 2.6. An integrated-circuit mask begins with a large-scale drawing of each mask. Early photomasks were cut by hand in a material called ruby lith, a sandwich of a clear backing layer and a thin red layer of Mylar. The red layer was cut with a stylus and peeled off, leaving the desired pattern in red. The original ruby lith copy of the mask was 100 to 1000 times larger than the final integrated circuit and was photographically reduced to form a reticle for use in a step-and-repeat camera, as described later.

Today, computer graphics systems and optical pattern generators have largely supplanted the use of ruby lith. An image of the desired mask is created on a computer...
Reticule images range from one to ten times final size. A step-and-repeat camera is used to reduce the reticle image to its final size and to expose a two-dimensional array of images on a master copy of the final mask. On a 125-mm wafer, it is possible to get approximately 1900 copies of a 2.5 mm × 2.5 mm integrated-circuit chip! Figure 2.7 shows examples of a computer graphics plot, a reticle, and a final mask for a simple integrated circuit.

A final master copy of the mask is usually made in a thin film of metal, such as chrome, on a glass plate. The mask image is transferred to photoresist, which is used as an etch mask for the chrome. Working emulsion masks are then produced from the chrome master.

Each time a mask is brought into contact with the surface of the silicon wafer, the pattern can be damaged. Therefore, emulsion masks are used for only a few exposures before they are thrown away. Contact printing has been largely replaced by proximity and projection printing systems, illustrated in Fig. 2.8. In proximity printing, the mask is brought in very close proximity to the wafer but does not come in contact with the wafer during exposure, thus preventing damage to the mask. Projection printing uses a dual-lens system to project a portion of the mask image onto the wafer surface. The wafer and masks may be scanned or the system may operate in a step-and-repeat mode. The actual mask and lenses are mounted many centimeters from the wafer surface.

Fig. 2.6 Outline of steps in the mask fabrication process.

graphics system. Once the image is complete, files containing the commands needed to drive a pattern generator are created on magnetic tape or disks. The pattern generator uses a flash lamp to expose the series of rectangles composing the mask image directly onto a photographic plate called the reticle.

Fig. 2.7 Mask fabrication. (a) Composite computer graphics plot of all masks for a simple integrated circuit; (b) 10× reticle of metal-level mask; (c) final-size emulsion mask with 400 copies of the metal level of the integrated circuit in (a). (Figure continued on p. 24.)
In large-diameter wafers, it is difficult to maintain alignment between mask levels across the complete wafer, particularly with features whose size approaches 1 \( \mu m \). High-resolution systems now use direct step-and-repeat techniques. A projection system is used with a 1X or 10X reticle to expose the integrated-circuit die pattern directly on the wafer. No step-and-repeat masks of the circuit are produced. The pattern is aligned and exposed separately at each die site.

Masks requiring geometrical features smaller than 1.25 \( \mu m \) can be produced by writing the pattern on the wafer in a special electron-sensitive resist using electron beams. Electron-beam systems are also commonly used to make 1X reticles for direct step-on-wafer systems.

### 2.4 SUMMARY

Photolithography is used to transfer patterns from masks to photoresist on the surface of silicon wafers. The resist protects portions of the surface while windows are etched in

---

**Fig. 2.8** Artist's conception of various printing techniques. (a) Contact printing, in which wafer is in intimate contact with mask; (b) proximity printing, in which wafer and mask are in close proximity; (c) projection printing, in which light source is scanned across the mask and focused on the wafer. Copyright, 1983, Bell Telephone Laboratories, Incorporated. Reprinted by permission from ref. [2].

---

**Fig. 2.7** (continued)
urer layers such as silicon dioxide, silicon nitride, or metal. The windows may be
drilled using either wet or dry processing techniques. Wet chemical etching tends to etch
der the edge of the mask, causing a loss of linewidth control at small dimensions. Dry
etching can yield highly anisotropic etching profiles and is required in most VLSI
processing.

After etching, impurities can be introduced into the wafer through the windows using
implantation and/or high-temperature diffusion, or metal can be deposited on the
surface making contact with the silicon through the etched windows. Masking operations
are performed over and over during integrated-circuit processing, and the number of
mask steps required is used as a basic measure of process complexity.

Mask fabrication uses computer graphics systems to draw the chip image at 100 to
100 times final size. Reticles on to ten times final size are made from this computer
image, using optical pattern generators or electron-beam systems. Step-and-repeat can-
as are used to fabricate final masks from the reticles, or direct step-on-wafer systems
may be used to transfer the patterns directly to the wafer.

Today we are reaching the limits of optical lithography. Present equipment can define
windows that are approximately 1.25 μm wide. (Just a few years ago, experts thought
at 2 μm would be the limit! Today it appears that it may be possible to extend optical
lithography to submicron dimensions.) The wavelength of light is too long to produce
such smaller geometrical features because of fringing and interference effects. Electron-
beam and X-ray lithography are now being used to fabricate devices with geometrical
features smaller than 0.25 μm, and lithography test structures have reproduced shapes
that minimum feature sizes of 0.1 μm.

REFERENCES
1. L. F. Thompson, C. G. Wilson, and M. I. Bowden, Eds., Introduction to Microlithography,
4. W. Kern and D. A. Poitinen, "Cleaning Solutions Based upon Hydrogen Peroxide for Use in
5. W. Kern, "Purifying Si and SiO₂ Surfaces with Hydrogen Peroxide," Semiconductor Interna-

ADDITIONAL READING

PROBLEMS
1. A complex CMOS fabrication process requires 15 masks. What fraction of the dice must be
good (i.e., what yield must be obtained) during each mask step if we require 90% of the final dice
to be good?

2. The mask set for a simple rectangular pn junction diode is shown in Fig. P2.2. The diode is
formed in a p-type substrate. Draw a picture of the horizontal layout for the diode which results
when a worst-case misalignment of 3 μm occurs in both the x and y directions on each mask level.

![Fig. P2.2]
3 / Thermal Oxidation of Silicon

Upon exposure to oxygen, the surface of a silicon wafer oxidizes to form silicon dioxide. This native silicon dioxide film is a high-quality electrical insulator and can be used as a barrier material during impurity diffusion. These two properties of silicon dioxide were the primary process factors leading to silicon becoming the dominant material in use today for the fabrication of integrated circuits. This chapter discusses the theory of oxide growth, oxide growth processes, factors affecting oxide growth rate, impurity redistribution during oxidation, and techniques for selective oxidation of silicon. Methods for determining the thickness of the oxide film are also presented.

3.1 THE OXIDATION PROCESS

Thermal oxidation of silicon is easily achieved by heating the wafer to a high temperature, typically 900 to 1200 °C, in an atmosphere containing either pure oxygen or water vapor. Both water vapor and oxygen move (diffuse) easily through silicon dioxide at these high temperatures (see Fig. 3.1). Oxygen arriving at the silicon surface can then combine with silicon to form silicon dioxide. The chemical reaction occurring at the silicon surface is:

$$ \text{Si} + \text{O}_2 \rightarrow \text{SiO}_2 $$  \hspace{1cm} (3.1)

for dry oxygen and

$$ \text{Si} + 2\text{H}_2\text{O} \rightarrow \text{SiO}_2 + 2\text{H}_2 $$  \hspace{1cm} (3.2)

for water vapor. Silicon is consumed as the oxide grows, and the resulting oxide expands during growth, as shown in Fig. 3.2. The final oxide layer is approximately 54% above the original surface of the silicon and 46% below the original surface.
3.1 Diffusivities of hydrogen, oxygen, sodium, and water vapor in silicon glass. Copyright \( \text{\textcopyright} \) Wiley & Sons, Inc. Reprinted with permission from ref. [2].

MODELING OXIDATION

Order for oxidation to occur, oxygen must reach the silicon interface. As the oxide grows, oxygen must pass through more and more oxide, and the growth rate decreases as time goes on. A simple model for oxidation can be developed by assuming that oxygen uses through the existing oxide layer. Fick's first law of diffusion states that the flux flow per unit area, \( J \) (called particle flux), is directly proportional to the concentration gradient of the particle:

\[
J = -D \frac{\partial N(x, t)}{\partial x} \tag{3.3}
\]

where \( D \) is the diffusion coefficient and \( N \) is the particle concentration. The negative sign indicates that particles move from a region of high concentration to a region of low concentration.

For our case of silicon oxidation, we will make the approximation that the oxygen flux passing through the oxide in Fig. 3.3 is constant everywhere in the oxide (oxygen does not accumulate in the oxide). The oxygen flux \( J \) is then given by

\[
J = -D(N_t - N_0)/X_0 \quad \text{(number of particles/cm}^2\text{-sec)} \tag{3.4}
\]

where \( N_t \) is the thickness of the silicon dioxide layer at any time \( t \), \( J \) is the constant flux of oxygen diffusing through the layer, and \( N_0 \) and \( N_t \) represent the oxygen concentration at the oxide surface and silicon dioxide-silicon interface, respectively.

Fig. 3.2 Formation of a silicon dioxide layer on the surface of a silicon wafer consumes silicon during growth of the layer. The oxide expands to fill a region approximately 54% above and 46% below the original surface of the wafer. The exact percentages depend on the density of the oxide.

Fig. 3.3 Model for thermal oxidation of silicon. \( X_0 \) is the thickness of the silicon dioxide layer at any time \( t \), \( J \) is the constant flux of oxygen diffusing through the layer, and \( N_0 \) and \( N_t \) represent the oxygen concentration at the oxide surface and silicon dioxide-silicon interface, respectively.
THERMAL OXIDATION OF SILICON

\[ X_0 \] is the thickness of the oxide at a given time, and \( N_o \) and \( N_i \) are the concentrations of oxidizing species in the oxide at the oxide surface and silicon dioxide–silicon face, respectively. At the silicon dioxide–silicon interface, we assume that the reaction rate is proportional to the concentration of the oxidizing species so that the flux \( J \) at the interface is

\[ J = k_s N_i \]  
(3.5)

\( k_s \) is called the rate constant for the reaction at the Si-SiO\(_2\) interface. Eliminating eqs. (3.4) and (3.5), the flux \( J \) becomes

\[ J = DN_0/(X_0 + D/k_s) \]  
(3.6)

The rate of change of thickness of the oxide layer with time is then given by the oxidizing divided by the number of molecules \( M \) of the oxidizing species that are incorporated in unit volume of the resulting oxide:

\[ \frac{dX_o}{dt} = J/M = (DN_0/M)/(X_0 + D/k_s) \]  
(3.7)

The differential equation is easily solved using the boundary condition \( X_o(t = 0) = X_{i_0} \), which yields

\[ X_i^2 + AX_i = Bt \]  
(3.8)

\( A = 2D/k_s, B = 2DN_0/M, \) and \( \tau = X_{i_0}/B + X_i/(B/A). \) \( X_i \) is the initial thickness of the wafer. A thin native oxide layer (10 to 20 Å) is always present on silicon, and atmospheric oxidation, or \( X_i \), may represent a thicker oxide grown during previous ion steps. Solving eq. (3.8) for \( X_o(t) \) yields

\[ X_o(t) = 0.5A \left[ \left( 1 + \frac{4B}{A^2}(t + \tau) \right)^{1/2} - 1 \right] \]  
(3.9)

For short times with \( t + \tau \gg A^2/4B \),

\[ X_o(t) = (B/A)(t + \tau) \]  
(3.10)

growth is proportional to time, and the ratio \( B/A \) is called the linear (growth) rate constant. In this region, growth rate is limited by the reaction at the silicon interface.

3.3 FACTORS INFLUENCING OXIDATION RATE

For long times with \( t + \tau \gg A^2/4B, t \gg \tau \),

\[ X_0 = \sqrt{Bt} \]  
(3.11)

Oxide growth is proportional to the square root of time, and \( B \) is called the parabolic rate constant. The oxidation rate is diffusion-limited in this region.

3.3 FACTORS INFLUENCING OXIDATION RATE

There is good experimental agreement with this simple theory. Figures 3.4 and 3.5 show experimental data for the parabolic and linear rate constants. The rate-constant data follow straight lines when plotted on a semilogarithmic scale versus reciprocal temperature. This type of behavior occurs in many natural systems and is referred to as an Arrhenius relationship. A mathematical model for this behavior is as follows:

\[ D = D_0 \exp(-E_A/kT) \]  
(3.12)

![Fig. 3.4 Dependence of the parabolic rate constant \( B \) on temperature for the thermal oxidation of silicon in pyrogenic H\(_2\)O (640 torr) or dry O\(_2\). Reprinted by permission of the publisher, The Electrochemical Society, Inc., from ref. 7.](image-url)
Thermal Oxidation of Silicon

Temperature (°C)

<table>
<thead>
<tr>
<th>1200</th>
<th>1100</th>
<th>1000</th>
<th>900</th>
<th>800</th>
<th>700</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.0</td>
<td>9.0</td>
<td>8.0</td>
<td>7.0</td>
<td>6.0</td>
<td>5.0</td>
</tr>
</tbody>
</table>

Linear rate constant, B/A (μm/hr)

- H₂O (640 torr)
  \( \frac{B/A}{(111) Si} = 2.05 \text{ eV} \)
- Dry O₂
  \( \frac{B/A}{(111) Si} = 2.05 \text{ eV} \)

dependence of the linear rate constant \( B/A \) on temperature for the thermal oxidation on in pyrogenic H₂O (640 torr) or dry O₂. Reprinted by permission of the publisher, The Chemical Society, Inc., from ref. [7].

Values for Coefficient \( D₀ \) and Activation Energy \( Eₐ \) for Wet and Dry Oxygen.*

<table>
<thead>
<tr>
<th>Wet O₂ (( Xₐ = 0 ) nm)</th>
<th>Dry O₂ (( Xₐ = 25 ) nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D₀ )</td>
<td>( Eₐ )</td>
</tr>
<tr>
<td>(100) Silicon</td>
<td></td>
</tr>
<tr>
<td>Linear (B/A)</td>
<td>9.70 \times 10⁷ μm/hr</td>
</tr>
<tr>
<td>Parabolic (B)</td>
<td>386 μm²/hr</td>
</tr>
<tr>
<td>(100) Silicon</td>
<td></td>
</tr>
<tr>
<td>Linear (B/A)</td>
<td>1.63 \times 10⁷ μm/hr</td>
</tr>
<tr>
<td>Parabolic (B)</td>
<td>386 μm²/hr</td>
</tr>
</tbody>
</table>

*Data from ref. [7].

Fig. 3.6 Wet and dry silicon dioxide growth for (100) silicon calculated using the data from Table 3.1.
3.7 Wet and dry silicon dioxide growth for (111) silicon calculated using the data from Fig. 3.1.

The rate in a wet atmosphere. Slower growth results in a denser, higher-quality oxide, which is usually used for MOS gate oxides. More rapid growth in wet oxygen is used for interconnecting layers.

Eq. (3.8) shows that both the linear and parabolic rate constants are proportional to $N_0$, which is proportional to the partial pressure of the oxidizing species, so pressure can be used to control oxide growth rate. There is great interest in developing low-temperature processes for VLSI fabrication. High pressure is being used to increase oxidation rates at temperatures (see Fig. 3.8). In addition, very thin oxides (50 to 200 Å) are used for VLSI, and low-pressure oxidation is being investigated as a means of controlling growth of thin oxides.

Figures 3.4 through 3.7 also show the dependence of oxidation rate on substrate orientation for the (111) and (100) materials most commonly used in bipolar and S processes, respectively. The crystal orientation changes the number of silicon bonds available at the silicon surface, which influences the oxide growth rate and quality of the silicon-silicon dioxide interface.

Example 3.1: Using Fig. 3.6, a 1 hr oxidation of (100) silicon at 1000 °C in dry oxygen will produce a silicon dioxide film approximately 580 Å (0.0058 μm) thick. The same oxidation in wet oxygen will yield a film 3900 Å (0.39 μm) thick.

Example 3.2: A (100) wafer has a 2000-Å oxide on its surface. (a) How long did it take to grow this oxide at 1100 °C in dry oxygen? (b) The wafer is put back in the furnace in wet oxygen at 1000 °C. How long will it take to grow an additional 3000 Å of oxide? Solve this problem graphically using Fig. 3.6 and 3.7 as appropriate. (c) Repeat part (b) using the oxidation theory presented in eqs. (3.3) through (3.12).

Solution: (a) Using Fig. 3.6, it would take 2.75 hr to grow a 0.2-μm oxide in dry oxygen at 1100 °C.
Heavy doping of silicon also changes its oxidation characteristics. Phosphorus doping increases the linear rate constant without altering the parabolic rate constant. Boron doping, on the other hand, increases the parabolic rate constant but has little effect on the linear rate constant. These effects are related to impurity redistribution during oxidation, which is discussed in the next section.

**DOPANT REDISTRIBUTION DURING OXIDATION**

Oxide oxidation, the impurity concentration changes in the silicon near the silicon dioxide interface. Boron and gallium tend to be depleted from the surface, whereas phosphorus, arsenic, and antimony pile up near the surface. Impurity depletion and pileup depend on both the diffusion coefficient and the segregation coefficient of the impurity in the oxide. The segregation coefficient \( m \) is
defined as the ratio of the equilibrium concentration of the impurity in silicon to that of the oxide in the oxide. Various possibilities are depicted in Fig. 3.9. The value of \( m \) for \( \text{Si} \) is temperature-dependent and is less than 0.3 at normal diffusion temperatures. It also diffuses slowly through \( \text{SiO}_2 \). Thus, boron is depleted from the silicon surface remains in the oxide (Fig. 3.9a). The presence of hydrogen during oxide growth or nitride diffusion greatly enhances the diffusion of boron through oxide, resulting in increased depletion of boron at the silicon surface (Fig. 3.9b).

The value of \( m \) is approximately 1.0 for phosphorus, antimony, and arsenic. These elements are rejected by the oxide, and they diffuse slowly in the oxide, resulting in pileup at the silicon surface (Fig. 3.9c).

Gallium has a segregation coefficient of 20. However, it diffuses very rapidly through silicon dioxide. This combination causes depletion of gallium at the surface, as shown in Fig. 3.9d.

The effects of boron depletion and phosphorus pileup are particularly important in bipolar and MOS processing. Process design must take both problems into account, and it may be necessary to add or change processing steps to overcome the effects of each phenomenon.

---

**3.5 MASKING PROPERTIES OF SILICON DIOXIDE**

One of the most important properties of silicon dioxide is its ability to mask impurities during high-temperature diffusion. The diffusivities of antimony, arsenic, boron, and phosphorus in silicon dioxide are all orders of magnitude smaller than the corresponding values in silicon. Thus \( \text{SiO}_2 \) films can be used effectively to mask these elements. Relatively deep diffusion can take place in unprotected regions of silicon, whereas no significant impurity penetration will occur in regions covered by silicon dioxide.

Figure 3.10 gives the \( \text{SiO}_2 \) thickness required to mask boron and phosphorus diffusions as a function of diffusion time and temperature. Note that silicon dioxide is four to five times more effective in masking boron than in masking phosphorus. Arsenic and antimony diffuse more slowly than phosphorus, so an oxide thick enough to mask
phosphorus is also sufficient to mask arsenic and antimony. Masking oxide thicknesses of 0.5 to 1.0 μm are typical in integrated-circuit processes. The masking oxide would be considered to have failed if the impurity level under the mask were to reach a significant fraction (10%) of the background concentration in the silicon.

The graph for boron is valid for an environment which contains no hydrogen! As mentioned earlier, the presence of hydrogen greatly enhances the boron diffusivity. Wet oxidation releases hydrogen, and care must be taken to avoid boron diffusion in the presence of water vapor.

As mentioned in Section 3.4, gallium diffuses rapidly through SiO₂, as does aluminum, and silicon dioxide cannot be used as a mask for these elements. However, silicon nitride can be used effectively as a mask for these impurities.

### 3.6 TECHNOLOGY OF OXIDATION

Thermal oxidation of silicon is typically carried out in a high-temperature furnace tube. The tubes may be made of quartz, polycrystalline silicon, or silicon carbide and are specially fabricated to prevent sodium contamination during oxidation. The wafers are placed upright on edge in a quartz boat and pushed slowly into the furnace. The furnace is maintained at a temperature between 800 and 1200 °C. Three-zone resistance-heated furnaces maintain the temperature within a fraction of a degree over a distance of 0.5 m in the center zone. A photograph of a typical six-tube furnace used for oxidation and diffusion appears in Fig. 3.11.

The furnace is continually purged with an inert gas such as nitrogen prior to oxidation. Oxidation begins by introducing the oxidizing species into the furnace in gaseous form. Extremely high-purity oxygen is available and is used for dry oxidation. Water vapor may be introduced by passing oxygen through a bubbler containing deionized water heated to 95 °C. The oxygen serves as a transport gas to carry the water vapor into the furnace. High-purity water vapor can also be obtained by burning hydrogen and oxygen in the furnace tube. Steam is not often used because it tends to pit the silicon surface.

### 3.7 OXIDE QUALITY

Wet oxidation is used to grow relatively thick oxides used for masking. An oxidation growth cycle usually consists of a sequence of dry/wet/dry oxidations. Most of the oxide
THERMAL OXIDATION OF SILICON

rown during the wet oxidation phase since the growth rate is much higher in the ence of water. Dry oxidation results in a higher-density oxide than that achieved with oxidation. Higher density in turn results in a higher breakdown voltage (5 to 15V/cm). In order to maintain good process control, the thin gate oxides (<1000 Å) MOS devices are usually formed using dry oxidation.

MOS devices are usually fabricated on wafers having a (100) surface orientation. The (100) orientation results in the smallest number of unsatisfied silicon bonds at the Si-SiO₂ interface, and the choice of the (100) orientation yields the lowest number of interface states.

Sodium ions are highly mobile in SiO₂ films (see Fig. 3.1), and contamination of Si gate oxides was a difficult problem to overcome in the early days of the integrated-circuit industry. Bipolar devices are much more tolerant of oxide contamination than MOS devices, and this was a primary factor in the early dominance of bipolar on-chip circuits.

Sodium-ion contamination results in mobile positive charge in the oxide. In addition, substantial level of positive fixed oxide exists at the Si-SiO₂ interface. These charge centers attract electrons to the surface of MOS transistors, resulting in a negative threshold voltage of the MOS devices. NMOS devices become depletion-mode devices. PMOS devices remain enhancement-mode devices but have more negative threshold voltages. The first successful MOS processes were therefore PMOS processes. The industry was able to improve overall oxide quality, NMOS processes became rarer because of the mobility advantage of electrons over holes.

It was discovered that the effects of sodium contamination can be greatly reduced by using chlorine during oxidation. Chlorine is incorporated into the oxide and immobile the sodium ions. A small amount (6% or less) of anhydrous HCl can be added to oxidizing gas. Gaseous chlorine, oxygen, or nitrogen can also be bubbled through chloroethylene (C₂H₄Cl). It should also be noted that the presence of chlorine during oxidation results in an increase in both the linear and parabolic rate constants.

SELECTIVE OXIDATION

oxidation processes described above generally form an oxide film over the complete surface of the silicon wafer. The ability to selectively oxidize the silicon surface has been very important in high-density bipolar and MOS processes. Selective oxidation processes result in improved device packing density and more planar final structures. Oxygen and water vapor do not diffuse well through silicon nitride. Figure 3.12 shows an MOS process with selective oxidation in which silicon nitride is used as the etch mask. A thin layer (10 to 20 nm) of silicon dioxide is first grown on the wafer to etch the silicon wafer. Next, a layer of silicon nitride is deposited over the surface patterned using photolithography. The wafer then goes through a thermal oxidation step. Oxide grows wherever the wafer is not protected by silicon nitride. This process results in the so-called semirecessed oxide structure.

Some oxide growth occurs under the edges of the nitride and causes the nitride to bend up at the edges of the masked area. The penetration of the oxide underneath the nitride results in a "bird's beak" structure. Formation of the bird's beak in Fig. 3.12 leads to loss of geometry control in VLSI structures. Hence, minimization of the bird's beak phenomenon is an important goal in VLSI process design.

A fully recessed oxide can be formed by etching the silicon prior to oxidation. This process can yield a very planar surface after the removal of the nitride mask. Subsequent processing reduces the advantage of this process over the semirecessed version, and most processes today use some form of semirecessed oxidation.

3.9 OXIDE THICKNESS CHARACTERIZATION

One of the simplest methods for determining the thickness of an oxide is to compare the color of the wafer with the reference color chart in Table 3.2. When a wafer is illuminated with white light perpendicular to the surface, the light penetrates the oxide film and is reflected by the underlying silicon wafer. Constructive interference causes enhancement of a certain wavelength in the reflected light, and the color of the wafer corresponds to the enhanced wavelength. Constructive interference occurs when the...
Table 3.2 (continued)

<table>
<thead>
<tr>
<th>Film Thickness (µm)</th>
<th>Color and Comments</th>
<th>Film Thickness (µm)</th>
<th>Color and Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05 Tan</td>
<td></td>
<td>0.54 Yellow green</td>
<td></td>
</tr>
<tr>
<td>0.07 Brown</td>
<td></td>
<td>0.56 Green yellow</td>
<td></td>
</tr>
<tr>
<td>0.10 Dark violet to red violet</td>
<td></td>
<td>0.57 Yellow to “yellowish” (not yellow but in the position where yellow is to be expected; at times appears to be light creamy gray or metallic)</td>
<td></td>
</tr>
<tr>
<td>0.12 Royal blue</td>
<td></td>
<td>0.60 Carnation pink</td>
<td></td>
</tr>
<tr>
<td>0.15 Light blue to metallic blue</td>
<td></td>
<td>0.63 Violet red</td>
<td></td>
</tr>
<tr>
<td>0.17 Metallic to very light yellow green</td>
<td></td>
<td>0.68 “Bluish” (not blue but borderline between violet and blue green; appears more like a mixture between violet red and blue green and looks grayish)</td>
<td></td>
</tr>
<tr>
<td>0.20 Light gold or yellow; slightly metallic</td>
<td></td>
<td>0.72 Blue green to green (quite broad)</td>
<td></td>
</tr>
<tr>
<td>0.22 Gold with slight yellow orange</td>
<td></td>
<td>0.77 “Yellowish”</td>
<td></td>
</tr>
<tr>
<td>0.25 Orange to melon</td>
<td></td>
<td>0.80 Orange (rather broad for orange)</td>
<td></td>
</tr>
<tr>
<td>0.30 Blue to violet blue</td>
<td></td>
<td>0.82 Salmon</td>
<td></td>
</tr>
<tr>
<td>0.31 Blue</td>
<td></td>
<td>0.85 Doll, light red violet</td>
<td></td>
</tr>
<tr>
<td>0.32 Blue to blue green</td>
<td></td>
<td>0.86 Violet</td>
<td></td>
</tr>
<tr>
<td>0.34 Light green</td>
<td></td>
<td>0.87 Blue violet</td>
<td></td>
</tr>
<tr>
<td>0.35 Green to yellow green</td>
<td></td>
<td>0.89 Blue</td>
<td></td>
</tr>
<tr>
<td>0.36 Yellow green</td>
<td></td>
<td>0.92 Blue green</td>
<td></td>
</tr>
<tr>
<td>0.37 Green yellow</td>
<td></td>
<td>0.95 Doll yellow green</td>
<td></td>
</tr>
<tr>
<td>0.39 Yellow</td>
<td></td>
<td>0.97 Yellow to “yellowish”</td>
<td></td>
</tr>
<tr>
<td>0.41 Light orange</td>
<td></td>
<td>0.99 Orange</td>
<td></td>
</tr>
<tr>
<td>0.42 Carnation pink</td>
<td></td>
<td>1.00 Carnation pink</td>
<td></td>
</tr>
<tr>
<td>0.44 Violet red</td>
<td></td>
<td>1.24 Carnation pink to salmon</td>
<td></td>
</tr>
<tr>
<td>0.46 Red violet</td>
<td></td>
<td>1.25 Orange</td>
<td></td>
</tr>
<tr>
<td>0.47 Violet</td>
<td></td>
<td>1.28 “Yellowish”</td>
<td></td>
</tr>
<tr>
<td>0.48 Blue violet</td>
<td></td>
<td>1.32 Sky blue to green blue</td>
<td></td>
</tr>
<tr>
<td>0.49 Blue</td>
<td></td>
<td>1.40 Orange</td>
<td></td>
</tr>
<tr>
<td>0.50 Blue green</td>
<td></td>
<td>1.45 Violet</td>
<td></td>
</tr>
<tr>
<td>0.52 Green (broad)</td>
<td></td>
<td>1.46 Blue violet</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.50 Blue</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.54 Dull yellow green</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.56 Violet</td>
<td></td>
</tr>
</tbody>
</table>

The path length in the oxide (2X_0) is equal to an even multiple of one wavelength of light in the oxide:

\[2X_0 = k \lambda / n\]  \hspace{1cm} (3.13)

where the number \(k\) is any integer greater than zero, and \(n\) is the refractive index of the oxide (\(n = 1.46\) for SiO_2).

As an example, a wafer with a 5000-Å silicon dioxide layer will appear blue green. Color-chart comparisons are quite subjective, and the colors vary periodically with thickness. In addition, care must be exercised to determine the color from a position perpendicular to the wafer. The color chart (Table 3.2) is only valid for vertical illumination with fluorescent light.

Accurate thickness measurement can be achieved with an instrument called an ellipsometer, and this instrument is often used to make an accurate reference color chart. Polarized monochromatic light is used to illuminate the wafer at an angle to the surface. Light is reflected from both the oxide and silicon surfaces. The differences in polarization are measured, and the oxide thickness can then be calculated.169

A mechanical surface profiler can also be used to measure film thickness. The oxide is partially etched from the surface of a test wafer to expose a step in the wafer and oxide surfaces. A stylus is mechanically scanned over the surface of the wafer, and thickness variations are recorded on a strip-chart recorder. Films ranging from less than 0.01 µm to more than 5 µm can be measured with this instrument.

Accurate film thickness measurements can also be achieved using light-interference effects in microscopy, and automated interference-based equipment is commercially available for thin-film characterization.
3.10 SUMMARY

Silicon dioxide provides a high-quality insulating barrier on the surface of the silicon wafer. In addition, this layer can serve as a barrier layer during subsequent impurity-diffusion process steps. These two factors have allowed silicon to become the dominant semiconductor material in use today.

A native oxide layer several tens of angstroms thick forms on the surface of silicon immediately upon exposure to oxygen even at room temperature. The thickness of this oxide layer may be readily measured from the accumulation-region capacitance of a MOS test capacitor. Thicker layers of silicon dioxide are conveniently grown in high-temperature oxidation furnaces using both wet and dry oxygen. Oxidation occurs much more rapidly in wet oxygen than in dry oxygen. However, the dry-oxygen environment produces a higher-quality oxide and is usually used for the growth of MOS gate oxides. Oxide cleanness is extremely important for MOS processes, and great care is exercised to prevent sodium contamination of the oxide. The addition of chlorine during oxidation improves oxide quality. Finally, oxidation alters the impurity distribution at the surface of the silicon wafer. Boron tends to be depleted from the silicon surface, whereas phosphorus tends to pile up at the silicon surface.

Oxidation thickness can be accurately measured using ellipsometers, interferometry, microscopes, and mechanical surface profilers or can be estimated from the apparent color of the oxide under vertical illumination with white light.

REFERENCES


PROBLEMS

3.1 How long does it take to grow 100 nm of oxide in wet oxygen at 1000 °C (assume (100) silicon)? Is dry oxygen? Which process would be preferred?

3.2 A 1.2-μm silicon dioxide film is grown on a (100) silicon wafer in wet oxygen at 1000 °C. How long does it take to grow the first 0.4 μm? The second 0.4 μm? The final 0.4 μm?

3.3 Derive eq. (3.8) by solving differential eq. (3.7).

3.4 How much oxide is needed to mask 4 hr boros diffusion at 1150 °C? A 1 hr phosphorus diffusion at 1050 °C?

3.5 A square window is etched through 200 nm of oxide prior to a second oxidation as in Example 3.2. The second oxidation grows 300 nm of oxide in the thick oxide region. Make a scale drawing of the cross section of the wafer after the second oxidation. What are the colors of the various regions under vertical illumination by white light?

3.6 Write a computer program to calculate the linear and parabolic rate constants for wet and dry oxidation for temperatures of 950, 1000, 1050, 1100, 1150, and 1200 °C. Assume (100) silicon.

3.7 A (100) silicon wafer has 400 nm of oxide on its surface. How long will it take to grow an additional 1 μm of oxide in wet oxygen at 1000 °C? Compare graphical and mathematical results. What is the color of the final oxide under vertical illumination by white light?

3.8 Yellow light has a wavelength of approximately 570 nm. Calculate the thicknesses of silicon dioxide which will appear yellow under vertical illumination by white light. Consider oxide thicknesses less than 1.5 μm. Compare with the color chart (Table 3.2).

3.9 Write a computer program to calculate the time required to grow a given thickness of oxide based on the theory of Section 3.2. The user should be able to specify desired oxide thickness, wet or dry oxidation conditions, temperature, and orientation of the silicon wafer.
4 / Diffusion

High-temperature diffusion has historically been one of the most important processing steps used in the fabrication of monolithic integrated circuits. Diffusion has been the primary method of introducing impurities such as boron, phosphorus, and antimony into silicon to control the majority-carrier type and resistivity of layers formed in the wafer. In this chapter, we explore the theoretical and practical aspects of the diffusion process, the characterization of diffused layer sheet resistance, and the determination of junction depth. Physical diffusion systems and solid, liquid, and gaseous impurity sources are all discussed.

4.1 THE DIFFUSION PROCESS

The diffusion process begins with the deposition of a high concentration of the desired impurity on the silicon surface through windows etched in the protective barrier layer. At high temperatures (900 to 1200 °C), the impurity atoms move from the surface into the silicon crystal via the substitutional or interstitial diffusion mechanisms illustrated in Fig. 4.1.

In the case of substitutional diffusion, the impurity atom hops from one crystal lattice site to another. The impurity atom thereby "substitutes" for a silicon atom in the lattice. Vacancies must be present in the silicon lattice in order for the substitutional process to occur. Statistically, a certain number of vacancies will always exist in the lattice. At high temperatures, vacancies may also be created by displacing silicon atoms from their normal lattice positions into the vacant interstitial space between lattice sites. The substitutional diffusion process in which silicon atoms are displaced into interstitial sites is called interstitially diffusion.

Considerable space exists between atoms in the silicon lattice, and certain impurity atoms diffuse through the crystal by jumping from one interstitial site to another. Since this mechanism does not require the presence of vacancies, interstitial diffusion proceeds much more rapidly than substitutional diffusion. The rapid diffusion rate makes interstitial diffusion difficult to control.
4.2 MATHEMATICAL MODEL FOR DIFFUSION

The basic one-dimensional diffusion process follows Fick's first law of diffusion, presented in Chapter 3:

\[ J = -D \frac{\partial N}{\partial x} \]  \hspace{1cm} (4.1)

where \( J \) is the particle flux of the donor or acceptor impurity species, \( N \) is the concentration of the impurity, and \( D \) is the diffusion coefficient.

Fick's second law of diffusion may be derived using the continuity equation for the particle flux:

\[ \frac{\partial N}{\partial t} = -\nabla \cdot J \]  \hspace{1cm} (4.2)

Eq. (4.2) states that the rate of increase of concentration with time is equal to the negative of the divergence of the particle flux. For the one-dimensional case, the divergence is equal to the gradient. Combining eqs. (4.1) and (4.2) yields Fick's second law of diffusion:

\[ \frac{\partial N}{\partial t} = D \frac{\partial^2 N}{\partial x^2} \]  \hspace{1cm} (4.3)

in which the diffusion coefficient \( D \) has been assumed to be independent of position. This assumption is violated at high impurity concentrations (see Section 4.8).

The partial differential equation in eq. (4.3) can be solved by variable separation or Laplace transform techniques. Two specific types of boundary conditions are important in modeling impurity diffusion in silicon. The first is the constant-source diffusion, in which the surface concentration is held constant throughout the diffusion. The second is called a limited-source diffusion, in which a fixed quantity of the impurity species is deposited in a thin layer on the surface of the silicon.

4.2.1 Constant-Source Diffusion

During a constant-source diffusion, the impurity concentration is held constant at the surface of the wafer. Under this boundary condition, the solution to eq. (4.3) is given by

\[ N(x, t) = N_i \text{erfc}(x/2\sqrt{Dt}) \]  \hspace{1cm} (4.4)

for a semi-infinite wafer in which \( N_i \) is the impurity concentration at the wafer surface \((x = 0)\). Such a diffusion is called a complementary error function (erfc) diffusion, shown graphically in Fig. 4.2. As time progresses, the diffusion front proceeds further and further into the wafer with the surface concentration remaining constant. The total number of impurity atoms per unit area in the silicon is called the dose, \( Q \), with units of atoms/cm\(^2\). \( Q \) increases with time, and an external impurity source must supply a
Fig. 4.2 A constant-source diffusion results in a complementary error function impurity distribution. The surface concentration \( N_0 \) remains constant and the diffusion moves deeper into the silicon wafer as the \( Dt \) product increases. \( Dt \) can change as a result of increasing diffusion time, increasing diffusion temperature, or a combination of both.

Continual flow of impurity atoms to the surface of the wafer. The dose is found by integrating the diffused impurity concentration throughout the silicon wafer.

\[
Q = \int_0^\infty N(x) \, dx = 2N_0 \sqrt{Dt} \frac{1}{\sqrt{\pi}}
\]  

4.2.2 Limited-Source Diffusion

A limited-source diffusion is modeled mathematically using an impulse function at the silicon surface as the initial boundary condition. The magnitude of the impulse is equal to the dose \( Q \). For this boundary condition in a semi-infinite wafer, the solution to eq. (4.3) is given by the Gaussian distribution,

\[
N(x, t) = \left( Q / \sqrt{\pi Dt} \right) \exp \left( -x^2 / 2Dt \right)
\]  

which is displayed graphically in Fig. 4.3. The dose remains constant throughout the limited-source diffusion process. As the diffusion front moves into the wafer, the surface concentration must decrease so that the area under the curve can remain constant with time.

On a normalized logarithmic plot, the shapes of the Gaussian and complementary error function curves appear similar, as illustrated in Fig. 4.4. The erfc curve, however, falls off more rapidly than the Gaussian curve.

4.3 THE DIFFUSION COEFFICIENT

Figure 4.5 shows the temperature dependence of the diffusion coefficient \( D \) for substitutional and interstitial diffusers in silicon. The large difference between these coefficients is readily apparent. In order to achieve reasonable diffusion times with substitutional diffusers, temperatures in the range of \( 900 \rightarrow 1200 \, ^\circ C \) are typically used. Interstitial
Fig. 4.4 A graph comparing the Gaussian and complementary error function (erfc) profiles. We will use this curve to evaluate the erfc and its inverse.

Diffusers are difficult to control because of their large diffusion coefficients (see problem 4.13).

Diffusion coefficients depend exponentially on temperature and follow the Arrhenius behavior discussed in Chapter 3:

\[ D = D_0 \exp\left(-\frac{E_a}{kT}\right) \]  \hspace{1cm} (4.7)

Values for \( D_0 \) and \( E_a \) can be determined from Fig. 4.5. Typical values for a number of impurities are given in Table 4.1.

Wide variability exists in diffusion coefficient data reported in the literature. We will use eq. (4.7) and Table 4.1 in the examples and problems throughout the rest of this book. In general, calculations based on eq. (4.7) and Table 4.1 can be used as guides. Most processes are then experimentally calibrated under the specific diffusion conditions in each laboratory.
Example 4.1: Calculate the diffusion coefficient for boron at 1100 °C.

Solution: From Table 4.1, \( D_B = 10.5 \text{ cm}^2/\text{sec} \) and \( E_A = 3.69 \text{ eV} \). \( T = 1373 \text{ K} \). \( D = 10.5 \exp\left(-\frac{3.69}{(8.614 \times 10^{-19})} \right) = 2.96 \times 10^{-3} \text{ cm}^2/\text{sec} \).

4.4 SUCCESSIVE DIFFUSIONS

We are ultimately interested in the final impurity distribution after all processing is complete. A wafer typically goes through many time-temperature cycles during predeposition, drive-in, oxide growth, CVD, etc. For example, the base diffusion in a bipolar transistor will be followed by several high-temperature oxidations as well as the emitter predeposition and drive-in cycles. These steps take place at different temperatures for different lengths of time. The effect of these steps is determined by calculating the total \( Dt \) product, \( (Dt)_\text{tot} \), for the diffusion. \( (Dt)_\text{tot} \) is equal to the sum of the \( Dt \) products for all high-temperature cycles affecting the diffusion:

\[
(Dt)_\text{tot} = \sum_i D_i t_i \tag{4.8}
\]

\( D_i \) and \( t_i \) are the diffusion coefficient and time associated with the \( i \)th processing step. \( (Dt)_\text{tot} \) is then used in eq. (4.4) or (4.6) to determine the final impurity distribution.

4.5 SOLID-SOLUBILITY LIMITS

At a given temperature, there is an upper limit to the amount of an impurity which can be absorbed by silicon. This quantity is called the solid-solubility limit for the impurity and is indicated by the solid lines in Fig. 4.6 for boron, phosphorus, antimony, and arsenic at normal diffusion temperatures. As can be seen in Fig. 4.6, surface concentrations achieved through solid-solubility-limited diffusions will be quite high. For example, the solid-solubility limit of boron is approximately \( 3.3 \times 10^{10}/\text{cm}^2 \) at 1100 °C, and \( 1.2 \times 10^{11}/\text{cm}^2 \) for phosphorus at the same temperature. High concentrations are desired for the emitter and subcollector diffusions in bipolar transistors and the source and drain diffusions in MOSFETs. However, solid-solubility-limited concentrations are too heavy for the base regions of bipolar transistors and for many resistors. The two-step diffusion process described in Section 4.2.3 was developed to overcome this problem.

At high concentrations, only a fraction of the impurities actually contribute to either silicon or electrons for conduction. The dotted lines in Fig. 4.6 show the "electrically active" portion of the impurity concentration. These curves will be referred to again in Section 4.7.2.

---

**Table 4.1** Typical Diffusion Coefficient Values for a Number of Impurities.

<table>
<thead>
<tr>
<th>Element</th>
<th>( D_f (\text{cm}^2/\text{sec}) )</th>
<th>( E_A (\text{eV}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>10.5</td>
<td>3.69</td>
</tr>
<tr>
<td>Al</td>
<td>8.00</td>
<td>3.47</td>
</tr>
<tr>
<td>Ga</td>
<td>3.60</td>
<td>3.51</td>
</tr>
<tr>
<td>In</td>
<td>16.5</td>
<td>3.90</td>
</tr>
<tr>
<td>P</td>
<td>10.5</td>
<td>3.69</td>
</tr>
<tr>
<td>As</td>
<td>0.32</td>
<td>3.56</td>
</tr>
<tr>
<td>Sb</td>
<td>5.60</td>
<td>3.95</td>
</tr>
</tbody>
</table>
4.6 JUNCTION FORMATION AND CHARACTERIZATION

4.6.1 Vertical Diffusion and Junction Formation

The goal of most diffusions is to form $p-n$ junctions by converting $p$-type material to $n$-type material or vice versa. In Fig. 4.7, for example, the wafer is uniformly doped $n$-type material with a concentration indicated by $N_n$, and the diffusing impurity is boron.

Fig. 4.7 Formation of a $p-n$ junction by diffusion. (a) An example of a $p$-type Gaussian diffusion into a uniformly doped $n$-type wafer; (b) net impurity concentration in the wafer. The metallurgical junction occurs at the point $x = x_j$, where the net concentration is zero. The material is converted to $p$-type to the left of $x_j$ and remains $n$-type to the right of $x_j$.

The point at which the diffused impurity profile intersects the background concentration is the metallurgical junction depth, $x_j$. The net impurity concentration at $x_j$ is zero. Setting $N(x)$ equal to the background concentration $N_b$ at $x = x_j$ yields

$$x_j = 2\sqrt{D \tau \ln(N_b/N_b)}$$  \hspace{1cm} (4.9a)
and

$$x_i = 2\sqrt{D_t} \text{ erfc}^{-1}(N_b/N_a) \quad (4.9b)$$

for the Gaussian and complementary error function distributions, respectively. In Fig. 4.7, the boron concentration $N$ exceeds $N_b$ to the left of the junction, and this region is $p$-type. To the right of $x_i$, $N$ is less than $N_a$, and this region remains $n$-type.

We can use our scientific calculators to evaluate eq. (4.9a), and we will learn to evaluate the complementary error function expression using Fig. 4.4. In order to calculate the junction depth, we must know the background concentration $N_b$ of the original wafer. Figure 4.8 gives the resistivity of $n$- and $p$-type silicon as a function of doping concentration. The background concentration can be determined using this figure when uniform concentrations of either donor or acceptor impurities are present in the silicon wafer.

**Example 4.2:** A boron diffusion is used to form the base region of an $n$-type transistor in a 0.18-ohm-cm $n$-type silicon wafer. A solid-solubility-limited boron predeposition is performed at 900 °C for 15 min followed by a 5 hr drive-in at 1100 °C. Find the surface concentration and junction depth (a) following the predeposition step and (b) following the drive-in step.

**Solution:** The predeposition step is a solid-solubility-limited constant-source diffusion. Using Fig. 4.6, the boron surface concentration is approximately $1.1 \times 10^{20}/\text{cm}^3$. The temperature of 900 °C equals 1173 K, which yields a diffusion coefficient $D_t = 1.45 \times 10^{-15} \text{ cm}^2/\text{sec}$, and $t_i = 900$ sec. The constant-source diffusion results in an erfc profile, and the impurity profile following predeposition is given by

$$N(x) = 1.1 \times 10^{20} \text{ erfc}(x/2\sqrt{D_t}t_i) \text{ boron atoms/cm}^3$$

To find the junction depth $x_j$, we must find the point at which the concentration $N(x)$ is equal to the background concentration $N_b$. Using Fig. 4.8, we find that a 0.18-ohm-cm $n$-type wafer corresponds to a doping concentration of $3 \times 10^{19} / \text{cm}^3$. Thus,

$$1.1 \times 10^{20} \text{ erfc}(x_i/2\sqrt{D_t}t_i) = 3 \times 10^{19}$$

Solving for $x_i$ yields

$$x_i = 2\sqrt{D_t}t_i \text{ erfc}^{-1}(0.000273) = 2(\sqrt{1.31 \times 10^{-15}})(2.57) \text{ cm} = 0.0587 \mu\text{m}$$

The dose in silicon is needed for the drive-in step and is equal to

$$Q = 2N_b\sqrt{D_t}t_i/\pi = 2(1.1 \times 10^{20})\sqrt{(1.45 \times 10^{-15}/\pi) \text{ boron atoms/cm}^2}$$

$$Q = 1.42 \times 10^{19} \text{ boron atoms/cm}^2$$

At the drive-in temperature of 1100 °C (1373 K), $D_t = 2.96 \times 10^{-13} \text{ cm}^2/\text{sec}$, and the drive-in time of 5 hr = 18,000 sec. Assuming that a Gaussian profile results from the drive-in step, the final profile is given by

$$N(x) = 1.1 \times 10^{20} \exp-(x/2\sqrt{D_t}t_i)^2 \text{ boron atoms/cm}^3 \quad (4.2.1)$$

Setting eq. (4.2.1) equal to the background concentration yields the final junction depth of 2.77 μm. Figure 4.9 shows the concentrations at various points in the diffusion process.

We must check our assumption that the drive-in step results in a Gaussian profile. The $D_t$ product for the predeposition step is $1.31 \times 10^{-15} \text{ cm}^2$, and the $D_t$ product for the drive-in step is $5.33 \times 10^{-15} \text{ cm}^2$. Thus $D_t \gg D_t$, and our assumption is justified.
62 DIFFUSION

![Graph showing diffusion profiles](image)

**Fig. 4.9** Calculated boron impurity profiles for Example 4.2. (a) Following the predeposition step at 900 °C for 15 min; (b) following a subsequent 5-hr drive-in step at 1100 °C. The final junction depth is 2.77 μm with a surface concentration of \(1.1 \times 10^{18}/\text{cm}^2\). The initial profile approximates an impulse.

### 4.6 Lateral Diffusion

During diffusion, impurities not only diffuse vertically but also move laterally under the edge of any diffusion barrier. Figure 4.10 presents the results of computer simulation of the two-dimensional diffusion process. The normalized impurity concentrations can be used to find the ratio of lateral to vertical diffusion. Lateral diffusion is an important effect-coupling device and process design and was an important factor driving the development of self-aligned polysilicon-gate MOS processes. The interaction of lateral diffusion and device layout will be discussed in greater detail in Chapters 9 and 10.

**Example 4.3:** An erf-diffusion results in a junction depth of 2 μm and a surface concentration of \(1 \times 10^{20}/\text{cm}^2\). The background concentration of the wafer is \(1 \times 10^{15}/\text{cm}^2\). What is the lateral diffusion underneath the edge of the mask?

![Normalized two-dimensional Gaussian and complementary error function diffusions](image)
**Solution:** The junction occurs at \( N = N_a \), and \( N_0/N_a = 10^4 \).

Using Fig. 4.10a, the ratio of lateral diffusion to vertical diffusion is 2.4/2.75, or 0.87. The lateral junction depth is therefore 1.74 µm.

### 6.3 Junction-Depth Measurement

Silicon wafers are normally processed in parallel with the actual integrated-circuit wafers. A masking is done on the test wafer so that diffusion may take place across its full surface. The test wafer provides a large area for experimental characterization of junction depth.

Two methods are commonly used to measure the junction depth of diffused layers. The first, known as the *groove-and-stain* method, a cylindrical groove is mechanically etched into the surface of the wafer, as in Fig. 4.11. If the radius \( R \) of the grinding tool known, the junction depth \( x \) is easily found to be

\[
x = \sqrt{R^2 - b^2} - \sqrt{R^2 - a^2}
\]

\[(4.10)\]

\[\begin{align*}
(x + x_2) - x &= \sqrt{R^2 - b^2} - \sqrt{R^2 - a^2} = R \left( \sqrt{1 - \frac{b^2}{R^2}} - \sqrt{1 - \frac{a^2}{R^2}} \right) \\
&= R \left( \frac{a - b}{2R} \right) \\
x &= \frac{a^2 - b^2}{2R} = \frac{(a + b)(a - b)}{2R}
\end{align*}\]

\[\text{Fig. 4.11} \quad \text{Junction-depth measurement by the groove-and-stain technique. The distances \( a \) and \( b \) are measured through a microscope, and the junction depth is calculated using eq. (4.11).}\]

If the radius \( R \) is much larger than both distances \( a \) and \( b \), then the junction depth is given approximately by

\[
x = \frac{(a^2 - b^2)}{2R} = \frac{(a + b)(a - b)}{2R}
\]

\[(4.11)\]

After the grooving operation, the junction is delineated using a chemical etchant which stains the *pn* junction. Concentrated hydrofluoric acid with 0.1 to 0.5% nitric acid can be used as a stain which is enhanced through exposure to high-intensity light.\(^{12}\) The distances \( a \) and \( b \) are measured through a microscope, and the junction depth is calculated using eq. (4.11).

The second technique is the *angle-lap* method. A piece of the wafer is mounted on a special fixture which permits the edge of the wafer to be lapped at an angle between 1 and 5°, as depicted in Fig. 4.12. The junction depth is magnified so that the distance on the lapped surface is given by

\[
x = d \tan \theta = \frac{N \lambda}{2}
\]

\[(4.12)\]

where \( \theta \) is the angle of the fixture. An optically flat piece of glass is placed over the lapped region, and the test structure is illuminated with a collimated monochromatic light from sodium vapor lamp.
beam of light with wavelength $\lambda$, typically from a sodium vapor lamp. The resulting interference pattern has fringe lines which are approximately 0.29 $\mu$m apart. The number of fringes is counted through a microscope, and the junction depth may be found using eq. (4.12).

### 4.7 SHEET RESISTANCE

In diffused layers, resistivity is a strong function of depth. For circuit and device design, it is convenient to work with a new parameter, $R_s$, called sheet resistance, which eliminates the need to know the details of the diffused-layer profile.

#### 4.7.1 Sheet-Resistance Definition

Let us first consider the resistance $R$ of the rectangular block of uniformly doped material in Fig. 4.13. $R$ is given by

$$R = \rho L / A$$

where $\rho$ is the material's resistivity, and $L$ and $A$ represent the length and cross-sectional area of the block, respectively. Resistance is proportional to the material resistivity. If the length of the block is made longer, the resistance increases, and the resistance is inversely proportional to cross-sectional area.

Using $W$ as the width of the sample and $t$ as the thickness of the sample, the resistance may be rewritten as

$$R = (\rho / t) (L / W) = R_s (L / W)$$

where $R_s = (\rho / t)$ is called the sheet resistance of the layer of material. Given the sheet resistance $R_s$, a circuit designer need specify only the length and width of the resistor to define its value. Strictly speaking, the unit for sheet resistance is the ohm, since the ratio $L/W$ is unitless. To avoid confusion between $R$ and $R_s$, sheet resistance is given the special descriptive unit of ohms per square. The ratio $L/W$ can be interpreted as the number of unit squares of material in the resistor.

Figure 4.14 shows top and side views of two typical dumbbell-shaped resistors with two contacts at the ends. The body of each resistor is seven "squares" long. If the sheet resistance of the diffusion were 50 ohms per square, each resistor would have a resistance of 350 ohms. Each end of the resistor adds approximately 0.65 squares to the resistor, and the total resistance would be approximately 415 ohms. Figure 4.15 gives the number of squares contributed by various end and corner configurations.

#### 4.7.2 Irvin's Curves

From Section 4.2 we know that the impurity concentration resulting from a diffusion varies rapidly between the surface and the junction. Thus $\rho$ is a function of depth for diffused resistors. For diffused layers, we define the sheet resistance $R_s$ by

$$R_s = \rho / x_1 = \left[ \int_0^n \sigma(x) dx \right]^{-1}$$

Fig. 4.13 Resistance of a block of material having uniform resistivity. A uniform current distribution is entering the material perpendicular to the end of the block. The ratio of resistivity to thickness is called the sheet resistance of the material.

Fig. 4.14 Top and side views of two diffused resistors of different physical size having equal values of resistance. Each resistor has a ratio $L/W$ equal to 7 squares. Each end of the resistor contributes approximately 0.65 additional squares.
DIFUSION

Exemple 4.4
Find the sheet resistance of the base diffusion from Example 4.2.

which \( x \) is the junction depth, \( \mu \) is the majority-carrier mobility, and \( N(x) \) is the net impurity concentration. We neglect the depletion of charge carriers near the junction.

For a given diffusion profile, sheet resistance is uniquely related to the surface concentration of the dopant.

(4.15) was evaluated numerically by Ired, and a number of Ired's results have been combined into Figs. 4.16a-e. These figures plot surface concentration versus the

Fig. 4.15 Effective square contributions of various resistor and center configurations.

Fig. 4.16 Surface impurity concentration versus the sheet resistance-junction depth product for different silicon background concentrations at 300 K. (a) p-type Gaussian distribution. (b) n-type Gaussian distribution. (c) Mini A&L (this figure comes from pages 76, 77, and 78.)

Sheet resistance-junction depth product, \( R_x \) (ohm-cm)

Surface dopant concentration, \( N_0 \) (atoms/cm²)
**Solution:** From Example 4.2, the background concentration of the wafer is $3 \times 10^{19}$/cm$^3$, the surface concentration is $1.1 \times 10^{20}$/cm$^3$, and the junction depth is $2.77\ \mu$m. The diffusion resulted in a $p$-type Gaussian layer. Using Fig. 4.16d, the $R_x$ product is found to be approximately 800 ohm-$\mu$m. Dividing by a junction depth of $2.77\ \mu$m yields a sheet resistance of 289 ohms/square.

Sheet resistance is an electrical quantity which depends on the majority-carrier concentration. As shown in Fig. 4.6, the electrically active impurity concentration for phosphorus and arsenic is considerably less than the total impurity concentration at high doping levels. In order to use Irvin's curves at high doping levels, the vertical axis, which is labeled "surface dopant density,” should be interpreted to be the electrically active dopant concentration at the surface.

### 4.7.3 The Four-Point Probe

A special instrument called a *four-point probe* may be used to measure the bulk resistivity of starting wafers and the sheet resistance of shallow diffused layers. As shown
van der Pauw's Method

The sheet resistance of an arbitrarily shaped sample of material may be measured by placing four contacts on the periphery of the sample. A current is injected through one pair of the contacts, and the voltage is measured across another pair of contacts. Van der Pauw\(^{13,14}\) demonstrated that two of these measurements can be related by eq. (4.19) below.

\[
\exp\left(-\frac{\pi t R_{AB,CD}}{\rho}\right) + \exp\left(-\frac{\pi t R_{BC,DA}}{\rho}\right) = 1
\]  

where \(R_{AB,CD} = V_{CD}/I_{AB}\) and \(R_{BC,DA} = V_{DA}/I_{BC}\). For a symmetrical structure like a square or a circle,

\[
R_{AB,CD} = R_{BC,DA}
\]

and

\[
R_s = \frac{\rho}{t} = (\pi/\ln 2)V_{CD}/I_{AB}
\]
Specially designed sheet-resistance test structures are often included on wafers so that the sheet resistances of $n$-type and $p$-type diffusions can be measured after final processing of the wafer. A sample structure is shown in Fig. 4.19.

4.8 CONCENTRATION-DEPENDENT DIFFUSION

Diffusion follows the theory of Section 4.3 as long as the impurity concentration remains below the value of the intrinsic-carrier concentration $n_i$ at the diffusion temperature.

Above this concentration, the diffusion coefficient becomes concentration-dependent. Each of the common impurities exhibits a different behavior.

The diffusion equation can be solved analytically for linear, parabolic, and cubic dependencies of the diffusion coefficient on concentration. The results are presented in Fig. 4.20, in which $D_{av}$ represents the diffusion coefficient at the surface. In general, concentration-dependent diffusion results in a much more abrupt profile than for the case of a constant-diffusion coefficient.

Boron and arsenic can be modeled by the first-order dependence in Fig. 4.20, resulting in the analytical relations between junction depth, sheet resistance, total dose, and surface concentration given in Table 4.2.

High-concentration phosphorus diffusion results in a more complicated profile than that of boron or arsenic. Figure 4.21 depicts typical shallow phosphorus diffusion profiles. As phosphorus diffuses into the wafer, the diffusion coefficient becomes enhanced at concentrations below approximately $10^{19}/\text{cm}^3$, resulting in a distinct “kink” in the profile. The kink effect represents a practical limitation to the use of phosphorus for the fabrication of diffusions.

**Table 4.2** Properties of High-Concentration Arsenic and Boron Diffusions.

<table>
<thead>
<tr>
<th>Element</th>
<th>$x_i$(cm)</th>
<th>$D$(cm$^2$/sec)</th>
<th>$N_0$(cm$^{-3}$)</th>
<th>$Q$(cm$^{-2}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arsenic</td>
<td>2.29$\sqrt{N_0Dt/n_i}$</td>
<td>$22.9 \exp(-4.1/kT)$</td>
<td>$1.56 \times 10^{10}(R_s x_i)^{-1}$</td>
<td>0.55$N_0x_i$</td>
</tr>
<tr>
<td>Boron</td>
<td>2.45$\sqrt{N_0Dt/n_i}$</td>
<td>$3.17 \exp(-3.59/kT)$</td>
<td>$2.78 \times 10^{10}(R_s x_i)^{-1}$</td>
<td>0.67$N_0x_i$</td>
</tr>
</tbody>
</table>

*The value of $n_i$ must be calculated at the diffusion temperature.
source/drain diffusions and emitter diffusions of shallow MOS and bipolar devices. Most MOS and bipolar VLSI processes now use arsenic to avoid this problem. Complex mathematical models describing the diffusion of phosphorus may be found in refs. 10 and 11.

9 PROCESS SIMULATION

As the scale of integrated circuits is reduced, accurate knowledge of one-, two-, and even three-dimensional impurity profiles is becoming more and more important. At the same time, experimental determination of profiles is becoming a very difficult and time-consuming task in VLSI fabrication processes.

Sophisticated computer programs which can predict the results of fabrication steps we become available. These programs not only numerically solve the generalized nonlinear diffusion equation in silicon, but also include the ability to simulate oxide growth with its attendant moving Si-SiO₂ boundary, impurity segregation during oxide growth, dopant evaporation from the surface, and ion implantation.

4.10 DIFFUSION SYSTEMS

One of the most widely used of these programs is SUPREM, the Stanford University Process Engineering Modeling program. The use of SUPREM requires specification of the process steps, including times, temperatures, and other ambient conditions for oxidation, diffusion, ion implantation, film deposition, and etching. The program calculates the impurity profile in the silicon substrate as well as in oxide and polysilicon layers. (See Problem 4.15.)

Simulation is growing in importance throughout the VLSI fabrication process. The detailed structures of recessed oxidation are being simulated, as are the photoresist and etching profiles resulting from processing at submicron dimensions. Ref. 20 contains a recent overview of process simulation.
the three-zone systems described in Chapter 3. For diffusion, wafers are placed in a quartz boat and positioned in the center zone of the furnace, where the wafers are heated to a high temperature. Impurities are transported to the silicon surface, where they diffuse into the wafer.

Most common silicon dopants can be applied using liquid spin-on sources. These spin-on dopants are versatile, safe, and easy to apply, but the uniformity is often poorer than with other impurity sources. To achieve good quality control, most production systems use other solid, liquid, or gaseous impurity sources.

In one type of solid-source system, carrier gases (usually N₂ or O₂) flow at a controlled rate over a source boat placed in the furnace tube. The carrier gas picks up the vapor from the source and transports it down the tube, where the dopant species is deposited on the surface of the wafer. The temperature of the source is controlled to maintain the desired vapor pressure. The source can be placed in a low-temperature section of the furnace or may be external to the furnace. Solid boron and phosphorus impurity sources are also available in wafer form and are placed in the boat between adjacent pairs of silicon wafers.

In liquid-source systems, a carrier gas passes through a bubbler where it picks up the vapor of the liquid source. The gas carries the vapor into the furnace tube where it reacts with the surface of the silicon wafer.

Gas-source systems supply the dopant species directly to the furnace tube in the gaseous state. The common gas sources are extremely toxic, and additional input purging and trapping systems are required to ensure that all the source gas is removed from the system before wafer entry or removal. In addition, most diffusion processes either do not use all of the source gas or produce undesirable reaction by-products. Therefore, the output of diffusion systems should be processed by burning or by chemical and/or water scrubbing before being exhausted into the atmosphere.

Boron is the only commonly used p-type dopant. The diffusion coefficients of aluminum and gallium are quite high in silicon dioxide, and these elements cannot be masked effectively by SiO₂. Indium is not used because it is a relatively deep-level acceptor (Eₐ - Eᵥ = 0.14 eV).

In contrast, antimony, phosphorus, and arsenic can all be masked by silicon dioxide and are all routinely used as n-type dopants in silicon processing.

### 4.10.1 Boron Diffusion

Boron has a high solubility in silicon and can achieve active surface concentrations as high as 4 × 10³⁴/cm³ (Fig. 4.6). Elemental boron is inert up to temperatures exceeding the melting point of silicon. A surface reaction with boron trioxide (B₂O₃) is used to introduce boron to the silicon surface:

\[
2B₂O₃ + 3Si ⇌ 4B + 3SiO₂
\]  (4.21)
An excess amount of boron trioxide can cause formation of a brown boron skin which is very difficult to remove with most acids. Boron skin formation can be minimized by performing the diffusions in an oxidizing atmosphere containing 3 to 10% oxygen. In a two-step diffusion, the boron predeposition step is commonly followed by a short wet-oxidation step to assist in removal of the boron skin prior to drive-in.

Common solid sources of boron include trimethylborate (TMB) and boron nitride wafers. TMB is a solid with high vapor pressure at room temperature. The TMB source is normally placed outside the diffusion furnace and cooled below room temperature during use. TMB vapor reacts in the furnace tube with oxygen to form boron trioxide, water, and carbon dioxide:

$$2(\text{CH}_3\text{O})_3\text{B} + 9\text{O}_2 \xrightarrow{900^\circ\text{C}} \text{B}_2\text{O}_3 + 6\text{CO}_2 + 9\text{H}_2\text{O}$$  \hspace{1cm} (4.22)

Any unreacted TMB should be scrubbed from the exhaust stream.

Boron nitride is a solid source available in wafer form. Activated wafers are placed in every third slot in the same quartz boat used to hold the silicon wafers. A silicon wafer faces each side of the oxidized boron nitride wafer, and boron trioxide is transferred directly to the surface of the silicon wafer during high-temperature diffusion. A small flow of inert gas such as nitrogen is used to keep contaminants out of the tube during diffusion.

The most common liquid source for boron is boron tribromide (BBr₃). The reaction is

$$4\text{BBr}_3 + 3\text{O}_2 \rightarrow 2\text{B}_2\text{O}_3 + 6\text{Br}_2$$  \hspace{1cm} (4.23)

Free bromine combines easily with metallic impurities and is useful in removing (gettering) metallic impurities during diffusion. Bromine, as well as unused boron tribromide, is in the exhaust stream, so the outlet gases should be carefully cleaned.

The primary gaseous source of boron is diborane (B₂H₆). Diborane is a highly poisonous and explosive gas. Table 4.3 summarizes the ACGIH recommendations for the maximum permissible exposure to the common gases used as diffusion sources. Extreme care must be taken in using these gases. In order to reduce the risk of handling, diborane is usually diluted with 99.9% argon or nitrogen by volume.

Diborane oxidizes in either oxygen or carbon dioxide to form boron trioxide:

$$\text{B}_2\text{H}_6 + 3\text{O}_2 \xrightarrow{900^\circ\text{C}} \text{B}_2\text{O}_3 + 3\text{H}_2\text{O}$$  \hspace{1cm} (4.24)

and

$$\text{B}_2\text{H}_6 + 6\text{CO}_2 \rightarrow \text{B}_2\text{O}_3 + 6\text{CO} + 3\text{H}_2\text{O}$$  \hspace{1cm} (4.25)

Both systems must provide a means for purging diborane from the input to the diffusion tube, and the output must be scrubbed to eliminate residual diborane and carbon monoxide.

### Table 4.3 Threshold Limit Recommendations for Common Gaseous Sources

<table>
<thead>
<tr>
<th>Source</th>
<th>8-h exposure level (ppm)</th>
<th>Life-threatening exposure</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diborane (B₂H₆)</td>
<td>0.10</td>
<td>160 ppm for 15 min</td>
<td>Colorless, sickly sweet, extremely toxic, flammable.</td>
</tr>
<tr>
<td>Phosphine (PH₃)</td>
<td>0.30</td>
<td>400 ppm for 30 min</td>
<td>Colorless, decaying fish odor, extremely toxic, flammable.</td>
</tr>
<tr>
<td>Arine (AsH₃)</td>
<td>0.05</td>
<td>6-15 ppm for 30 min</td>
<td>Colorless, garlic odor, extremely toxic.</td>
</tr>
<tr>
<td>Silane (SiH₄)</td>
<td>0.50</td>
<td>Unknown for 30 min</td>
<td>Repellent odor, burns in air, explosive, poorly understood.</td>
</tr>
<tr>
<td>Dichlorosilane (SiH₂Cl₂)</td>
<td>5.00</td>
<td>...</td>
<td>Colorless, flammable, toxic.</td>
</tr>
</tbody>
</table>

*Data from the 1979 American Conference of Governmental Hygienists (ACGIH).

#### 4.10.2 Phosphorus Diffusion

Phosphorus has a higher solubility in silicon than does boron, and surface concentrations in the low 10¹⁴/cm² range can be achieved during high-temperature diffusion. Phosphorus is introduced into silicon through the reaction of phosphorus pentoxide at the wafer surface:

$$2\text{P}_2\text{O}_5 \rightleftharpoons 4\text{P} + 5\text{SiO}_2$$  \hspace{1cm} (4.26)

Solid P₂O₅ wafers can be used as a solid source for phosphorus, as can ammonium monophosphate (NH₄H₂PO₄) and ammonium diphasoxide ([(NH₄)₂H₆PO₄] in wafer form. However, the most popular diffusion systems use either liquid or gaseous sources. Phosphorus oxychloride (POCl₃) is liquid at room temperature. A carrier gas is passed through a bubbler and brings the vapor into the diffusion furnace. The gas stream also contains oxygen, and P₂O₅ is deposited on the surface of the wafers:

$$4\text{POCl}_3 + 3\text{O}_2 \rightarrow 2\text{P}_2\text{O}_5 + 6\text{Cl}_2$$  \hspace{1cm} (4.27)

Liberated chlorine gas serves as a gettering agent, and Cl₂ and POCl₃ must be removed from the exhaust stream.
Phosphine, PH₃, is a highly toxic and explosive gas used as the gaseous source for phosphorus. It is also supplied in dilute form with 99.9% argon or nitrogen. Phosphine is oxidized with oxygen in the furnace:

\[ 2\text{PH}_3 + 4\text{O}_2 \rightarrow 3\text{P}_2\text{O}_5 + 3\text{H}_2\text{O} \]  
(4.28)

Unreacted phosphine must be cleaned from the exhaust gases, and the gas delivery system must be able to purge phosphine from the input to the tube.

4.10.3 Arsenic Diffusion

Arsenic has the highest solubility of any of the common dopants in silicon, with surface concentrations reaching \(2 \times 10^{21}/\text{cm}^2\). The surface reaction involves arsenic trioxide:

\[ 2\text{As}_2\text{O}_3 + 3\text{Si} \rightleftharpoons 3\text{SiO}_2 + 4\text{As} \]  
(4.29)

Oxide vapors can be carried into the furnace tube from a solid diffusion source by a nitrogen carrier gas. However, evaporation of arsenic from the surface limits surface concentrations to below \(3 \times 10^{21}/\text{cm}^2\). The exhaust must be carefully cleaned because of the presence of arsenic.

Arsine gas may be used as a source, but it is extremely toxic and also produces relatively low surface concentrations. The problems with arsenic deposition and safety delayed its widespread use in silicon processing until ion implantation was developed in the early 1970s. Ion implantation is now the preferred technique for introducing arsenic into silicon. (Chapter 5 is devoted to the subject of ion implantation.)

4.10.4 Antimony Diffusion

Antimony, like arsenic, has a low diffusion coefficient and has been used for a long time for buried layers in bipolar processes. Antimony trioxide is a solid source which is placed in a two-zone furnace in which the source is maintained at a temperature of 600 to 650 °C. Antimony is introduced at the silicon surface as in the other cases:

\[ 2\text{Sb}_2\text{O}_3 + 3\text{Si} \rightleftharpoons 3\text{SiO}_2 + 4\text{Sb} \]  
(4.30)

A liquid source, antimony pentachloride (Sb₅Cl₅), has been successfully used with oxygen as a carrier gas passing through a bubbler. The gas stabine (SbH₅) is unstable and cannot be used for antimony diffusion.

4.11 SUMMARY

In Chapter 4 we have discussed the formation of pn junctions using high-temperature diffusion. Mathematical models for diffusion have been presented, and the behavior of common n- and p-type dopants in silicon has been discussed. A key parameter governing the diffusion process is the diffusion coefficient, which is highly temperature-dependent, following an Arrhenius relationship. Boron, phosphorus, antimony, and arsenic all have reasonable diffusion coefficients in silicon at temperatures between 900 and 1200 °C, and they can be conveniently masked by a barrier layer of silicon dioxide. Gallium and aluminium are not easily masked by SiO₂ and are seldom used, and indium is not used because of its large activation energy. At high concentrations, diffusion coefficients become concentration-dependent, causing diffused profiles to differ substantially from predictions of simple theories.

Two types of diffusions are most often used. If the surface concentration is maintained constant throughout the diffusion process, then a complementary error function (erfc) distribution is obtained. In the erfc case, the surface concentration is usually set by the solid-solubility limit of the impurity in silicon. If a fixed dose of impurity is diffused into silicon, a Gaussian diffusion profile is achieved. These two cases are often combined in a two-step process to obtain lower surface concentrations than those achievable with a solid-solubility-limited diffusion. As the complexity of fabrication processes grows, simulation with process modeling programs such as SUPREM is becoming ever more important.

The concept of sheet resistance has been introduced, and Irvin’s curves have been used to relate the sheet resistance, junction depth, and surface concentration of diffused layers. Techniques for calculating and measuring junction depth have also been presented. Resistor fabrication has been discussed, including end and corner effects, as well as the effects of lateral diffusion under the edges of diffusion barriers.

High-temperature open-furnace diffusion systems are routinely used for diffusion with solid, liquid, and gaseous impurity sources. Boron, phosphorus, and antimony are all easily introduced into silicon using high-temperature diffusion. However, arsenic deposition by diffusion is much more difficult, and today it is usually accomplished using ion implantation (see Chapter 5). As with many chemicals used in integrated-circuit fabrication, some of the sources used for diffusion are extremely toxic and must be handled with great care.

REFERENCES


PROBLEMS

4.1 A phosphorus diffusion has a surface concentration of $5 \times 10^{18}$/cm$^2$, and the background concentration of the p-type wafer is $1 \times 10^{19}$/cm$^2$. The $D_t$ product for the diffusion is $10^{-8}$ cm$^2$/s.
(a) Find the junction depth for a Gaussian distribution.
(b) Find the junction depth for an erfc profile.
(c) What is the sheet resistance of the two diffusions?
(d) Draw a graph of the two profiles.

4.2 A 5-hr boron diffusion is to be performed at 1100 °C.
(a) What thickness of silicon dioxide is required to mask this diffusion?
(b) Repeat part (a) for phosphorus.

4.3 A boron diffusion into a 1-ohm-cm n-type wafer results in a Gaussian profile with a surface concentration of $5 \times 10^{16}$/cm$^2$ and a junction depth of 4 μm.
(a) How long did the diffusion take if the diffusion temperature was 1100 °C?
(b) What was the sheet resistance of the layer?
(c) What is the dose in the layer?
(d) The boron dose was deposited by a solid-solubility-limited diffusion. Design a diffusion schedule (temperature and time) for this predeposition step.

4.4 The boron diffusion in Problem 4.3 is followed by a solid-solubility-limited phosphorus diffusion for 30 min at 950 °C. Assume that the boron profile does not change during the phosphorus diffusion.
(a) Find the junction depth of the new phosphorus layer. Assume an erfc profile.
(b) Find the junction depth based on the concentration-dependent diffusion data presented in Fig. 4.21.

4.5 The p-well in a CMOS process is to be formed by a two-step boron diffusion into a 5-ohm-cm n-type substrate. The sheet resistance of the well is 1000 ohms per square and the junction depth is 7.5 μm.
(a) Design a reasonable diffusion schedule for the drive-in step which produces the p-well.
(b) What is the final surface concentration in the p-well?
(c) What is the dose required to form the well?
(d) Can this dose be achieved using a solid-solubility-limited diffusion with diffusion temperatures of 900 °C or above? Discuss.

4.6 The channel length of a metal-gate NMOS transistor is the spacing between the source and drain diffusions as shown in Fig. P4.6a. The spacing between the source and drain diffusion openings is 6 µm on the masking oxide used to make the transistor. The source/drain junctions are diffused to a depth of 1.5 µm using a constant-source diffusion. The surface concentration is $1 \times 10^{22}$/cm² and the wafer has a concentration of $1 \times 10^{18}$/cm³. What is the channel length in the actual device after the diffusion is completed?

![Fig. P4.6](image)

4.7 (a) What is the total number of squares in the resistor shown in Fig. P4.7 assuming that its geometry is specified precisely by the mask dimensions?

(b) The resistor is actually formed from a p-type base diffusion with a 6-µm junction depth. What is the actual number of squares in this resistor, assuming that the lateral diffusion under the edge of the mask is 5 µm?

(c) What would be the resistance of the resistors in parts (a) and (b) if the surface concentration of the base diffusion was $5 \times 10^{18}$ boron atoms/cm², the bulk concentration $10^{19}$/cm³, and the junction depth 6 µm.

![Fig. P4.7](image)

4.8 In practice, wafers are slowly pushed into and pulled out of the furnace, or the furnace temperature may be changed with time. Assume that the furnace temperature is being ramped down with time: $T = T_o - Rt$, where $T_o$ is the initial temperature and $R$ is the temperature change per second. Show that the effective $Dt$ product defined by

$$\langle Dt \rangle_{eff} = \int_0^t D(t) \, dt$$

where $t_o$ is the ramp-down time is given by

$$\langle Dt \rangle_{eff} = D(T_o) \left( \frac{kT_o^2}{2} \right)$$

where

$$D(T_o) = D_0 \exp\left(-\frac{E_s}{kT_o}\right)$$

4.9 Determine the sensitivity of junction depth to changes in furnace temperature by calculating $(\Delta T_j)/(\Delta T/k)$ for a Gaussian diffusion profile. What fractional change in junction depth will occur at 1100 °C if the furnace temperature is in error by 10 °C?

4.10 Rework Example 4.2 using the concentration-dependent boron diffusion expressions for the preposition calculations. Find the new surface concentration and junction depth following the drive-in step and compare the results with those presented in the example. At 900 °C, $n_i = 4 \times 10^{16}$/cm³.

4.11 Derive the expressions for the Gaussian and complementary-error-function solutions to the diffusion equation.

4.12 What is the minimum sheet resistance to be expected from shallow arsenic- and boron-doped regions if the regions are 1 µm deep? 0.25 µm deep? Make use of Fig. 4.6 and your knowledge of the dependence of mobility on doping (Volume I). Assume that the region is uniformly doped. Compare your results to the equations presented in Section 4.8.

4.13 Gold is diffused into a silicon wafer using a constant-source diffusion with a surface concentration of $10^{19}$/cm³. How long does it take the gold to diffuse completely through a silicon wafer 400 µm thick with a background concentration of $10^{10}$/cm³ at a temperature of 1000 °C?

4.14 A gas cylinder contains 100 ft³ of a mixture of diborane and argon. The diborane represents 0.1% by volume. An accident occurs and the complete cylinder is released into a room measuring 10 x 12 x 8 ft.

(a) What will be the equilibrium concentration of diborane in the room in ppm?

(b) Compare this level with the toxic level based on Table 4.3.

(c) Would your answer to part (b) change if the gas cylinder contained arginine?

4.15 (a) Use SUPREM to simulate the diffusion profile of Example 4.2. Compare the simulation results with those given in the example.

(b) Follow the boron diffusion by the growth of a 500-nm layer of oxide in wet oxygen at 1100 °C. Discuss what has happened to the boron concentration at the Si-SiO₂ interface.

(c) Add a 30-min solid-solubility-limited phosphorus diffusion at 1000 °C.

(d) The phosphorus diffusion created a new pn junction. Update the hand calculations for the boron impurity profile of Example 4.2 and estimate the location of both pn junctions with the aid of Fig. 4.21. Compare your results to those of SUPREM in part (c).
5 / Ion Implantation

Ion implantation offers many advantages over diffusion for the introduction of impurity atoms into the silicon wafer and has become a workhorse technology in modern integrated-circuit fabrication. In this chapter we will first discuss ion implantation technology and mathematical modeling of the impurity distributions obtained with ion implantation. We will subsequently explore deviations from the model caused by nonideal behavior and will discuss annealing techniques used to remove crystal damage caused by the implantation process.

5.1 IMPLANTATION TECHNOLOGY

An ion implanter is a high-voltage particle accelerator producing a high-velocity beam of impurity ions which can penetrate the surface of silicon target wafers. The basic parts of the system, shown schematically in Fig. 5.1, are described in detail below, beginning with the impurity-source end of the system.

1. **Ion Source.** The ion source operates at a high voltage (25 kV) and produces a plasma containing the desired impurity as well as other undesired species. Arsine, phosphine, and diborane, as well as other gases, can be used in the source. Solids can be sputtered in special ion sources, and this technique offers a wide degree of flexibility in the choice of impurity.

2. **Mass Spectrometer.** An analyzer magnet bends the ion beam through a right angle to select the desired impurity ion. The selected ion passes through an aperture slit into the main accelerator column.

3. **High-Voltage Accelerator.** The accelerator column adds energy to the beam (up to 175 keV) and accelerates the ions to their final velocity. Both the accelerator column and the ion source are operated at a high voltage relative to the target. For protection from high voltage and possible X-ray emission, the ion source and accelerator are mounted within a protective shield.
5.2 MATHEMATICAL MODEL FOR ION IMPLANTATION

The silicon wafer is maintained in good electrical contact with the target holder, so electrons can readily flow to or from the wafer to neutralize the implanted ions. This electron current is integrated over time to measure the total dose $Q$ from the implanter given by

$$Q = \int_0^\infty I \, dt / nqA$$

(5.3)

where $I$ is the beam current in amperes, $A$ is the wafer area, $n = 1$ for singly ionized ions and 2 for doubly ionized species, and $T$ is the implantation time. The use of a doubly ionized species increases the energy capability of the machine by a factor of 2 since $E = nqV$.

The target wafers can be maintained at relatively low temperatures during the implantation. Low-temperature processing prevents undesired spreading of impurities by diffusion, which is very important in VLSI fabrication. Another advantage of ion implantation is the ability to use a much wider range of impurity species than possible with diffusion. In principle, any element that can be ionized can be introduced into the wafer using implantation.

A production-level ion implanter may cost from $1$ to $2$ million, and cost is its greatest disadvantage. However, the advantages of flexibility and process control have far outweighed the disadvantage of cost, and ion implantation is now used routinely throughout bipolar and MOS integrated-circuit fabrication.

5.2 MATHEMATICAL MODEL FOR ION IMPLANTATION

As an ion enters the surface of the wafer, it collides with atoms in the lattice and interacts with electrons in the crystal. Each nuclear or electronic interaction reduces the energy of the ion until it finally comes to rest within the target. Interaction with the crystal is a statistical process, and the implanted impurity profile can be approximated by the Gaussian distribution function illustrated in Fig. 5.2. The distribution is described mathematically by

$$N(x) = N_p \exp\left(-\frac{(x - R_p)^2}{2 \Delta R_p^2}\right)$$

(5.4)

$R_p$ is called the projected range and is equal to the average distance an ion travels before it stops. The peak concentration $N_p$ occurs at $x = R_p$. Because of the statistical nature of the process, some ions will be “lucky” and will penetrate beyond the projected range $R_p$, and some will be “unlucky” and will not make it as far as $R_p$. The spread of the distribution is characterized by the standard deviation, $\Delta R_p$, called the straggig.

The area under the impurity distribution curve is the implanted dose $Q$, defined by

$$Q = \int_0^\infty N(x) \, dx$$

(5.5)
This theory assumes that the implantation goes into an amorphous material in which the atoms of the target material are randomly positioned. Figure 5.3 displays the results of LSS calculations for the projected range and straggles for antimony, boron, phosphorus, and arsenic in amorphous silicon and silicon dioxide. For the moment we will assume that these results are also valid for crystalline silicon. Deviations from the LSS theory will be discussed in Section 5.5.

Range and straggles are roughly proportional to ion energy over a wide range, although some nonlinear behavior is clearly evident in Fig. 5.3. For a given energy, the lighter elements strike the silicon wafer with a higher velocity and penetrate more deeply into the wafer. The result indicates that the projected ranges in Si and SiO2 are essentially the same, and we will assume that the stopping power of silicon dioxide is equal to that of silicon. Figure 5.3 also gives values for the transverse straggles $\Delta R_x$, which will be discussed in the next section.

![Graph showing projected range and straggles for various elements](image-url)

**Fig. 5.3** Projected range and straggles calculations based on LSS theory. (a) Projected range $R_p$ for boron, phosphorus, arsenic, and antimony in amorphous silicon. Results from SiO2 and for silicon are virtually identical. (b) (On page 94) Vertical $\Delta R_y$ and transverse $\Delta R_x$ straggles for boron, phosphorus, arsenic, and antimony. Reprinted with permission from ref. [2]. (Copyright Van Nostrand Reinhold Company, Inc.)
Example 5.1: Phosphorus with an energy of 100 keV is implanted into a silicon wafer. (a) What are the range and straggle associated with this implantation? (b) What should the implanted dose be if a peak concentration of $1 \times 10^{19}/\text{cm}^2$ is desired?

Solution: Using Fig. 5.3, the range and straggle are 0.12 $\mu$m and 0.045 $\mu$m, respectively. The dose and peak concentration are related by eq. (5.6). Note that this is an approximation, since the peak is only a little over $2\Delta R_p$ below the silicon surface.

$$Q = \sqrt{2\pi N_p \Delta R_p} = \sqrt{2\pi \left(1 \times 10^{19}/\text{cm}^2\right)(4.5 \times 10^{-6} \text{ cm})} = 1.13 \times 10^{19}/\text{cm}^2$$

5.3 SELECTIVE IMPLANTATION

In most cases we desire to implant impurities only in selected areas of the wafer. Windows are opened in a barrier material wherever impurity penetration is desired. In the center of the window the impurity distribution is described by eq. (5.4), but near the edges the distribution decreases and actually extends under the edge of the window, as shown in Fig. 5.4. The overall distribution can be modeled by:

$$N(x, y) = N(x)F(y)$$

$$F(y) = 0.5[\text{erfc}\left(\frac{y-a}{\sqrt{2}\Delta R_p}\right) - \text{erfc}\left(\frac{y+a}{\sqrt{2}\Delta R_p}\right)]$$

(5.7)

where $N(x)$ is given by eq. (5.4). The parameter $\Delta R_p$ is called the transverse straggle and characterizes the behavior of the distribution near the edge of the window. Figure 5.4 shows normalized impurity distributions near the barrier edge calculated with eq. (5.4). Figure 5.3b gives values of both normal straggle $\Delta R_p$ and transverse straggle $\Delta R_L$.

In order to mask the ion implantation, it is necessary to prevent the implanted impurity from changing the doping level in the silicon beneath the barrier layer. Figure 5.5 shows a silicon wafer with a layer of silicon dioxide on the surface. An implanted species is shown.

![Fig. 5.4](image)

Contours of equal ion concentration for an implantation into silicon through a 1-$\mu$m window. The profiles are symmetrical about the $x$-axis and were calculated using eq. (5.7), which is taken from ref. [4].
Implanted impurity profile with implant peak in the oxide. The barrier material must be tough to ensure that the concentration is the tail of the distribution is much less than $N_b$.

Silicon dioxide and silicon nitride are routinely used as barrier materials during implantation. Since implantation is a low-temperature process, additional materials such as photoresist and aluminum, which cannot withstand high-temperature diffusion, may be used as barrier materials during the implantation.

Silicon nitride is more effective than silicon dioxide in stopping ions, and a silicon nitride barrier layer need only be 85% of the thickness of an SiO$_2$ barrier layer. On the other hand, photoresist is less effective in stopping ions, and a photoresist barrier layer should be 1.8 times the thickness of an SiO$_2$ layer under the same implantation conditions. Metals are of such a high density that even a very thin layer will mask most implantations.

**Example 5.2:** A boron implantation is to be performed through a 50-nm gate oxide so that the peak of the distribution is at the Si-SiO$_2$ interface. The dose of the implant in silicon is to be $1 \times 10^{15}$/cm$^2$. (a) What are the energy of the implant and the peak concentration at the interface? (b) How thick should the SiO$_2$ layer be in areas which are not to be implanted, if the background concentration is $1 \times 10^{15}$/cm$^2$? (c) Suppose the oxide is 50 nm thick everywhere. How much photoresist is required on top of the oxide to completely mask the ion implantation?

**Solution:** The projected range needs to be 0.05 μm in order to place the peak of the distribution at the Si-SiO$_2$ interface. Using Fig. 5-3a, the $R_p$ of 0.05 μm requires an energy of 15 keV. Since the peak of the implant is at the interface, the total dose will be twice the dose needed in silicon. The peak concentration is

$$N_p = \frac{Q}{\Delta R_p \sqrt{2\pi}} = 2 \times 10^{12}/(2.3 \times 10^{-6} \sqrt{2\pi}) = 3.5 \times 10^{16}/\text{cm}^2$$

where the straggles was found using Fig. 5-3b. In order to completely mask the implantation, the tail of the distribution must be less than the background concentration at the interface. The minimum oxide thickness is found using eq. (5.9):

$$X_o = 0.05 + 0.023\sqrt{2 \ln(10N_p/N_b)} \mu \text{m} = 0.14 \mu \text{m}$$

Since the oxide is 0.05 μm thick, the photoresist must provide a thickness equivalent to 0.09 μm. The resist thickness must be 1.8 times the needed thickness of SiO$_2$ to provide
JUNCTION DEPTH AND SHEET RESISTANCE

Implantation is often used to form shallow pn junctions for various device applications. The implanted profile approximates a Gaussian distribution, and the junction depth \( x_j \) can be found by equating the implanted distribution to the background concentration, Chapter 4.

\[
N_p \exp\left(-\frac{(x_j - R_p)^2}{2 \Delta R_p^2}\right) = N_B \\
x_j = R_p \pm \Delta R_p \sqrt{2 \ln(N_D/N_B)}
\]  
(5.10)

These roots may be meaningful, as indicated in Fig. 5.6, in which a deep subsurface junction has junctions occurring at two different depths, \( x_{1j} \) and \( x_{2j} \).

Example 5.3: Boron is implanted into an n-type silicon wafer to a depth of 0.3 \( \mu \)m. Find the location of the junction if the peak concentration is \( 1 \times 10^{19} \) cm\(^{-3} \) and the doping of the silicon is \( 3 \times 10^{15} \) cm\(^{-3} \).

Solution: From Fig. 5.3a, the implant energy is 100 keV. From Fig. 5.3b, the straggle is 0.07 \( \mu \)m. Equating the Gaussian distribution to the background concentration yields

\[
3 \times 10^{19} = 10^{19} \exp\left(-\frac{(x_j - R_p)^2}{2 \Delta R_p^2}\right)
\]

\[
x_j = R_p \pm 2.65 \Delta R_p
\]

is yields junction depths of 0.12 \( \mu \)m and 0.49 \( \mu \)m.

The peak of an implantation is often positioned at the silicon surface. For this special case, we may use the curves for Gaussian distributions to find the sheet resistance of implanted layer, as discussed in Chapter 4. These curves may also be used to find the sheet resistance of a layer which is completely below the surface (see Problem 5.4), that an implanted Gaussian impurity distribution will remain Gaussian through any quent high-temperature processing steps.

If the peak energy is not deep enough to implant the entire wafer, the implanted Gaussian impurity distribution will remain Gaussian through any subsequent high-temperature processing steps. If the peak energy is deep enough to implant the entire wafer, the implanted Gaussian impurity distribution will remain Gaussian through any subsequent high-temperature processing steps.

The LSS results of Section 5.2 are based on the assumption that the target material is amorphous, having a completely random order. This assumption is true of thermal SiO\(_2\), deposited Si\(_3\)N\(_4\) and SiO\(_2\), and many thin metal films, but it is not valid for a crystalline substrate. The regular arrangement of atoms in the crystal leaves a large amount of open space in the crystal. Figure 5.7 shows a view through the silicon lattice in the (110) direction. If the incoming ion flux is improperly oriented with respect to the crystal planes, the ions will tend to miss the silicon atoms in the lattice and will "channel" much more deeply into the material than the LSS theory predicts. However, electronic interactions will eventually stop the ions.

The effects of channeling are demonstrated in Fig. 5.8. Phosphorus has been implanted at an energy of 40 keV into a silicon target with several orientations of the ion beam relative to the (100) silicon surface. The appearance of a random target can be achieved by tilting (100) silicon approximately 7° relative to the incoming beam. The results are represented by the x's in Fig. 5.8. The range for this case compares well with...
the LSS calculations presented in Fig. 5.3. The open circles represent the boron profile implanted perpendicular to the (100) surface. Note that the range for the "channeled" case is almost twice that predicted by the LSS theory. Results for two other angles of incidence are given in Fig. 5.8, showing progressively less channeling as the angle is increased.

5.5.2 Lattice Damage and Annealing

During the implantation process, ion impact can knock atoms out of the silicon lattice, damaging the implanted region of the crystal. If the dose is high enough, the implanted layer will become amorphous. Figure 5.9 gives the dose required to produce an amorphous silicon layer for various impurities as a function of substrate temperature. The heavier the impurity, the lower the dose required to create an amorphous layer. At sufficiently high temperatures, an amorphous layer can no longer be formed.

Implantation damage can be removed by an "annealing" step. Following implantation, the wafer is heated to a temperature between 800 and 1000 °C for approximately 30 min. At this temperature, silicon atoms can move back into lattice sites, and impurity atoms can enter substitutional sites in the lattice. After the annealing cycle,
nearly all of the implanted dose becomes electrically active, except for impurity concentrations exceeding $10^{19}/\text{cm}^3$.

Annealing cycles for 30 min at temperatures approaching 1000 °C can cause considerable spreading of the implant by diffusion. It has been found that truly amorphous layers can actually be annealed at lower temperatures through the process of solid-phase epitaxy. The crystalline substrate seeds recrystallization of the amorphous layer, and epitaxial growth can proceed as rapidly as 500 Å/min at 600 °C. During solid-phase epitaxy, impurity atoms are incorporated into substitutional sites, and full activation is achieved at low temperatures.

Low-energy arsenic implantations produce shallow amorphous layers that can be annealed using solid-phase epitaxy to yield shallow, abrupt junctions that are ideal for VLSI structures. Boron, however, is so light that it does not produce an amorphous layer even at relatively high doses, unless the substrate is deliberately cooled (see Fig. 5.9). Today, boron is often implanted using ions of the heavier BF$_2$ molecule. The lower-velocity implant results in shallow layers that can be annealed under solid-phase-epitaxy conditions.

5.5.3 Deviations from the Gaussian Theory

If we take a detailed look at the shape of the implanted impurity distribution, we find deviations from the ideal Gaussian profile. When light ions, such as boron, impact atoms of the silicon target, they experience a relatively large amount of backward scattering and fill in the distribution on the surface side of the peak, as in Fig. 5.10. Heavy atoms, such as arsenic, experience a larger amount of forward scattering and tend to fill in the profile on the substrate side of the peak. A number of methods have been proposed for mathematical modeling of this behavior, such as the use of Pearson Type-IV distributions. However, for common implant energies below 200 keV, the Gaussian theory provides a useful model of the impurity distribution. This is particularly true since the forward and backward scattering tend to align the tails of the distribution where the concentration is well below the peak value.

5.6 SUMMARY

Ion implantation uses a high-voltage accelerator to introduce impurity atoms into the surface of the silicon wafer, and it offers many advantages over deposition by high-temperature diffusion. Ion implantation is a low-temperature process minimizing impurity movement by diffusion, which has become very important to VLSI fabrication. Low-temperature processing also permits the use of a wide variety of materials as barrier layers to mask the implantation. Photore sist, oxide, nitride, aluminum, and other metal films can all be used, adding important increased flexibility to process design.

Ion implantation also permits the use of a much wider range of impurity species than diffusion. In principle, any element that can be ionized can be introduced into the wafer using implantation. Implantation offers much tighter control of the dose introduced into the wafer, and a much wider range of doses can be reproducibly achieved than possible with diffusion.

Diffused profiles almost always have the maximum impurity concentration at the surface. Ion-implantation techniques can be used to produce new profiles with subsurface peaks or retrograde profiles which decrease toward the wafer surface. Implantation can introduce impurities into very shallow layers near the surface, again a significant advantage for VLSI structures.

The main disadvantage of ion implantation is the cost of the equipment, which may exceed $2 million for a production machine. Also, the ion implanter has trouble achieving high doses (>10$^{19}/\text{cm}^3$) in a time reasonable for high-volume production. New high-current machines are being developed to overcome this latter problem. Overall, the flexibility and process control achievable with ion implantation have far outweighed the disadvantage of cost, and ion implantation is used routinely for state-of-the-art bipolar and MOS integrated-circuit fabrication.

Ion implantation results in profiles which can be modeled by a Gaussian distribution. The depth and width of the distribution depend on both the ion species and the energy of the implantation. In order to prevent channeling, implantation is normally performed at an angle of approximately 7° off the normal to the surface.

The implantation process damages the surface, and an annealing step is required to remove the effects of the damage. Low doses may result in the need for annealing at 800 to 1000 °C for 30 min. However, if the surface layer has become amorphous, annealing can be achieved through solid-phase epitaxy at temperatures of only 600 °C.

Fig. 5.10 Measured boron impurity distributions compared with four-moment (Pearson IV) distribution functions. The boron was implanted into amorphous silicon without annealing. Reprinted with permission from Philips Journal of Research.®
REFERENCES


PROBLEMS

5.1 Boron is implanted with an energy of 60 keV through a 0.25-μm layer of silicon dioxide. The implanted dose is 1 × 10¹⁸/cm².

(a) Find the boron concentration at the silicon-silicon dioxide interface.

(b) Find the dose in silicon.

(c) Determine the junction depth if the background concentration is 3 × 10¹⁵/cm³.

5.2 An arsenic dose of 1 × 10¹⁴/cm² is implanted through a 50-nm layer of silicon dioxide with the peak of the distribution at the Si-SiO₂ interface. A silicon nitride film on top of the silicon dioxide is to be used as a barrier material in the regions where arsenic is not desired. How thick should the nitride layer be if the background concentration is 1 × 10¹⁵/cm³?

5.3 An implantation will be used for the predeposition step for a boron base diffusion. The final layer is to be 5 μm deep with a sheet resistance of 125 ohms per square (Rᵦ = 10⁵ ohms).

(a) What is the dose required from the ion implanter if the boron will be implanted through a thin silicon dioxide layer so that the peak of the implanted distribution is at the silicon-silicon dioxide interface?

(b) What drive-in time is required to produce the final base layer at a temperature of 1100 °C?

5.4 (a) Use Irvin’s curves to find the sheet resistance of a boron layer implanted completely below the surface in n-type silicon (Nₑ = 10¹⁵/cm³). Assume the layer has a peak concentration of 1 × 10¹⁴/cm², and the range and straggle are 1.0 μm and 0.11 μm, respectively. (Hint: Think about conductors in parallel.)

(b) What is the dose of this implantation?

(c) What was the energy used for this ion implantation?

(d) At what depths are pn junctions located?

5.5 The source and drain regions of a self-aligned n-channel polysilicon-gate MOS transistor are to be formed using arsenic implantation. The dimensions of a cross section of the device are given in Fig. P5.5. Calculate the channel shrinkage caused by lateral straggling if the peak concentration if the implantation is 10¹⁴/cm² and the substrate doping is 10¹⁰/cm³. Assume that the channel region is in the silicon immediately below the oxide. Use Rᵦ = 0.1 μm, ΔRₑ = 0.04 μm, and ΔRᵦ = 0.022 μm.

![Fig. P5.5](image-url)
5.6 An implanted profile is formed by two boron implantations. The first uses an energy of 100 keV and the second an energy of 200 keV. The peak concentration of each distribution is $5 \times 10^{19}$/cm$^3$. Draw a graph of the composite profile and find the junction depth(s) if the phosphorus background concentration is $10^{10}$/cm$^3$. What are the dose rates of the two implant steps?

5.7 A high energy (5 MeV) beam is used to implant oxygen deep below the silicon surface in order to form a buried SiO$_2$ layer. Assume that the desired SiO$_2$ layer is to be 0.2 $\mu$m thick.

(a) What is the oxygen dose required to be implanted in silicon?
(b) What beam current is required if a 125-mm-diameter wafer is to be implanted in 15 min?
(c) How much power is being supplied to the ion beam? Discuss what effects this implantation may have on the wafer.

5.8 The threshold voltage of a NMOS transistor may be increased by ion implantation of boron into the channel region. For shallow implantations, the voltage shift is given approximately by $\Delta V_T = qQ/Cox$ where $Q$ is the boron dose and $Cox = \varepsilon_0/X_0$. $X_0$ is the oxide thickness and $\varepsilon_0$ is the permittivity of silicon dioxide: $3.9 \times (8.854 \times 10^{-14}$ F/cm). What boron dose is required to shift the threshold by 0.75 V if the oxide thickness is 40 nm?

5.9 An ion implanter has a beam current of 10 $\mu$A. How long does it take to implant a boron dose of $10^{10}$/cm$^2$ into a wafer with a diameter of 125 mm?

5.10 Write a computer program to calculate the sheet resistance of an arbitrary Gaussian layer in silicon.

6 / Film Deposition

Fabrication processes involve many steps in which thin films of various materials are deposited on the surface of the wafer. This chapter presents a survey of deposition processes, including evaporation, chemical vapor deposition, and sputtering, which are used to deposit metals, silicon and polysilicon, and dielectrics such as silicon dioxide and silicon nitride. Evaporation and sputtering require vacuum systems operating at low pressure, whereas chemical vapor deposition and epitaxy can be performed at either reduced or atmospheric pressure. An overview of vacuum systems and some results from the theory of ideal gases are also presented in this chapter.

6.1 EVAPORATION

Physical evaporation is one of the oldest methods of depositing metal films. Aluminum and gold are heated to the point of vaporization, and then evaporate to form a thin film covering the surface of the silicon wafer. In order to control the composition of the deposited material, evaporation is performed under vacuum conditions.

Figure 6.1 shows a basic vacuum deposition system consisting of a vacuum chamber, a mechanical roughing pump, a diffusion pump or turbomolecular pump, valves, vacuum gauges, and other instrumentation. In operation, the roughing valve is opened first, and the mechanical pump lowers the vacuum chamber pressure to an intermediate vacuum level of approximately 1 Pascal (Pa$^1$). If a higher vacuum level is needed, the roughing valve is closed, and the foreline and high-vacuum valves are opened. The roughing pump now maintains a vacuum on the output of the diffusion pump. A liquid-nitrogen (77 K) cold trap is used with the diffusion pump to reduce the pressure in the vacuum chamber to approximately $10^{-4}$ Pa. Ion and thermocouple gauges are used to monitor the pressure at a number of points in the vacuum system, and several other valves are used as vents to return the system to atmospheric pressure.

1 atm = 760 mm Hg = 760 torr = 1.013 x 10^5 Pa. 1 Pa = 1 N/m^2 = 0.0075 torr.
molecules/mole). The concentration of gas molecules is given by

\[ n = N_w/V = P/kT \]  

(6.2)

In some systems, the surface of the substrate must be kept extremely clean prior to deposition. The presence of even a small amount of oxygen or other elements will result in formation of a contamination layer on the surface of the substrate. The rate of formation of this layer is determined from the impingement rate of gas molecules hitting the substrate surface and is related to the pressure by

\[ \Phi = P/\sqrt{2\pi mkT} \text{ (molecules/cm}^2\text{-sec)} \]  

(6.3)

where \( m \) is the mass of the molecule. This can be reduced to

\[ \Phi = 2.53 \times 10^{28} P/\sqrt{MT} \text{ (molecules/cm}^2\text{-sec)} \]  

(6.4)

where \( P \) is the pressure in Pa and \( M \) is the molecular weight (e.g., \( M = 32 \) for oxygen molecules). If we assume that each molecule sticks as it contacts the surface, then the time required to form a monolayer on the surface is given by

\[ t = N_v/\Phi = N_v\sqrt{2\pi mkT}/P \]  

(6.5)

where \( N_v \) is the number of molecules/cm² in the layer.

**Example:** Suppose the residual pressure of oxygen in the vacuum system is 1 Pa. How long does it take to deposit one atomic layer of oxygen on the surface of the wafer at 300 K?

**Solution:** The radius of an oxygen molecule is approximately 3.6 Å. If we assume close packing of the molecules on the surface, there will be approximately \( 2.2 \times 10^{14} \) molecules/cm². At 300 K and 1 Pa, the impingement rate for oxygen is \( 2.7 \times 10^{16} \) molecules/cm²·sec. One monolayer is deposited in 82 μsec.

Pressure and temperature also determine another important film-deposition parameter called the mean free path, \( \lambda \). The mean free path of a gas molecule is the average distance the molecule travels before it collides with another molecule. \( \lambda \) is given by

\[ \lambda = kT/\sqrt{2\pi P d^2} \]  

(6.6)

in which \( d \) is the diameter of the gas molecule and is in the range of 2 to 5 Å. Evaporation is usually done at a background pressure near \( 10^{-4} \) Pa. At this pressure, a 4 Å molecule has a mean free path of approximately 60 m. Thus, during aluminum evaporation, for example, aluminum molecules do not interact with the background gases and tend to travel in a straight line from the evaporation source to the deposition target.
On the other hand, sputtering, which will be discussed in Section 6.4, uses argon gas at a pressure of approximately 100 Pa. Using the same radius results in a mean free path only 60 μm. Thus the material being deposited tends to scatter often with the argons and arrives at the target from random directions.

1.2 Filament Evaporation

The simplest evaporator consists of a vacuum system containing a filament which can be heated to high temperature. In Fig. 6.2a, small loops of a metal such as aluminum are formed from a filament formed of a refractory (high-temperature) metal such as tungsten. Evaporation is accomplished by gradually increasing the temperature of the filament until the metal melts and wets the filament. Filament temperature is then raised to evaporate the aluminum from the filament. The wafers are mounted near the filament and covered by a thin film of the evaporating material.

Although filament evaporation systems are easy to set up, contamination levels can be high, particularly from the filament material. In addition, evaporation of composite materials cannot be easily controlled using a filament evaporator. The material with the lowest melting point tends to evaporate first, and the deposited film will not have the same composition as the source material. Thick films are difficult to achieve since a limited supply of material is contained in the metal loops.

6.1.3 Electron-Beam Evaporation

In electron-beam (E-beam) evaporation systems (see Fig. 6.2b), the high-temperature filament is replaced with an electron beam. A high-intensity beam of electrons, with an energy up to 15 keV, is focused on a source target containing the material to be evaporated. The energy from the electron beam melts a region of the target. Material evaporates from the source and covers the silicon wafers with a thin layer.

The growth rate using a small planar source is given by

\[ G = \frac{n}{\pi \rho \cos \phi \cos \theta} \text{ (cm/sec)} \]  

(6.7)

for the geometrical setup in Fig. 6.3. \( \phi \) is the angle measured from the normal to the plane of the source, and \( \theta \) is the angle of the substrate relative to the vapor stream. \( \rho \) and \( m \) are the density (g/cm³) and mass evaporation rate (g/sec), respectively, of the material being deposited.

For batch deposition, a planetary substrate holder (Fig. 6.4) consisting of rotating sections of a sphere is used. Each substrate is positioned tangential to the surface of the sphere with radius \( r_o \), as in Fig. 6.3. Application of some geometry yields

\[ \cos \theta = \cos \phi = r/2r_o \]  

(6.8)

For the planetary substrate holder, \( G \) becomes independent of substrate position:

\[ G = m/4\pi \rho r_o^2 \]  

(6.9)

The wafers are mounted above the source and are typically rotated around the source during deposition to ensure uniform coverage. The wafers are also often radially heated to improve adhesion and uniformity of the evaporated material. The source material sits in a water-cooled crucible, and its surface only comes in contact with the electron beam during the evaporation process. Purity is controlled by the purity of the original source material. The relatively large size of the source provides a virtually unlimited supply of material for evaporation, and the deposition rate is easily controlled by changing the current and energy of the electron beam.

One method of monitoring the deposition rate uses a quartz crystal which is covered by the evaporating material during deposition. The resonant frequency of the crystal shifts in proportion to the thickness of the deposited film. By monitoring the resonant frequency of the crystal, the deposition rate may be measured with an accuracy of better than 1 Å/sec. Dual electron beams with dual targets may be used to coevaporate composite materials in E-beam evaporation systems.
3.1.4 Flash Evaporation

Flash evaporation uses a fine wire as the source material, and a high-temperature ceramic bar is used to evaporate the wire. The wire is fed continuously and evaporates on contact with the ceramic bar. Flash evaporation can produce relatively thick films, as in an E-beam system, without problems associated with radiation damage.

X-ray radiation can be generated in an electron-beam system for acceleration voltages exceeding 5 to 10 keV. Substrates may suffer some radiation damage from both energetic electrons and X-rays. The damage can usually be annealed out during subsequent process steps. However, the radiation effects are of great concern to MOS process designers, and so sputtering has replaced electron-beam evaporation in many steps in manufacturing processes.

6.1.5 Shadowing and Step Coverage

Because of the large mean free paths of gas molecules at low pressure, evaporation techniques tend to be directional in nature, and shadowing of patterns and poor step coverage can occur during deposition. Figure 6.5 illustrates the shadowing phenomenon which can occur with closely spaced features on the surface of an integrated circuit. In the fully shadowed region, there will be little deposition. In the partially shadowed region, there will be variation in film thickness. To minimize these effects, the planetary substrate holder of the electron-beam system continuously rotates the wafers during the film deposition.

6.2 SPUTTERING

Sputtering is achieved by bombarding a target with energetic ions, typically Ar⁺. Atoms at the surface of the target are knocked loose and transported to the substrate, where deposition occurs. Electrically conductive materials such as Al, W, and Ti can use a dc power source, in which the target acts as the cathode in a diode system. Sputtering of dielectrics such as silicon dioxide or aluminum oxide requires an RF power source to supply energy to the argon atoms. A diagram of a sputtering system is shown in Fig. 6.6.
Fig. 6.7 Sputtering yield versus ion energy for a dc sputtering system using argon.

Sputtering can be used to deposit a broad range of materials. In addition, alloys may be deposited in which the film has the same composition as the target. An example is the Al-Cu-Si alloy commonly used for metatllization in integrated circuits. (We will discuss this alloy in Chapter 7.) As one might expect, sputtering results in the incorporation of some argon into the film, and heating of the substrate up to 350 °C can occur during the deposition process. Sputtering can also give excellent coverage of the sharp topologies often encountered in integrated circuits.

Sputter etching (a reversal of the sputter deposition process) can be used to clean the substrate prior to film deposition, and the sputter etching process is often used to clean contact windows prior to metal deposition. Etching removes any residual oxide from the window and improves the contact between the metal and the underlying material.

6.3 CHEMICAL VAPOR DEPOSITION

Chemical vapor deposition (CVD) forms thin films on the surface of a substrate by thermal decomposition and/or reaction of gaseous compounds. The desired material is deposited directly from the gas phase onto the surface of the substrate. Polysilicon, silicon dioxide, and silicon nitride are routinely deposited using CVD techniques. In addition, refractory metals such as tungsten (W) can also be deposited using CVD.

Chemical vapor deposition can be performed at pressures for which the mean free path for gas molecules is quite small, and the use of relatively high temperatures can result in excellent conformal step coverage over a broad range of topological profiles.
6.3.1 CVD Reactors

Several different types of CVD reactor systems are shown in Fig. 6.8. In Fig. 6.8a, a continuous atmospheric-pressure reactor is shown. This type of reactor is often used for deposition of the silicon dioxide passivation layer as one of the last steps in integrated-circuit processing. The reactant gases flow through the center section of the reactor and are contained by nitrogen curtains at the ends. The substrates can be fed continuously through the system, and large-diameter wafers are easily handled. However, high gas-flow rates are required by the atmospheric-pressure reactor.

The hot-wall, low-pressure system of Fig. 6.8b is commonly used to deposit polysilicon, silicon dioxide, and silicon nitride, and is referred to as an LPCVD (low-pressure CVD) system. The reactant gases are introduced into one end of a three-zone furnace and are pumped out the other end. Temperatures range from 300 to 1150 °C, and the pressure is typically 30 to 250 Pa. Excellent uniformity can be obtained with LPCVD systems, and several hundred wafers may be processed in a single run. Hot-wall systems have the disadvantage that the deposited film simultaneously coats the inside of the tube. The tube must be periodically cleaned or replaced to minimize problems with particulate matter. In spite of this problem, hot-wall LPCVD systems are in widespread use throughout the semiconductor industry.

CVD reactions can also take place in a plasma reactor, as shown in Fig. 6.8c. Formation of the plasma permits the reaction to take place at low temperatures, which is a primary advantage of plasma-enhanced CVD (PECVD) processes. In the parallel-plate system, the wafers lie on a grounded aluminum plate which serves as the bottom electrode for establishing the plasma. The wafers can be heated up to 400 °C using high-intensity lamps or resistance heaters. The top electrode is a second aluminum plate placed in close proximity to the wafer surface. Gases are introduced along the outside of the system, flow radially across the wafers, and are pumped through an exhaust in the center. An RF signal is applied to the top plate to establish the plasma. The capacity of this type of system is limited, and wafers must be loaded manually. A major problem in VLSI fabrication is particulate matter that may fall from the upper plate onto the wafers.

The furnace-plasma system in Fig. 6.8d can handle a large number of wafers at one time. A special electrode assembly holds the wafers parallel to the gas flow. The plasma is established by alternating groups of electrodes supporting the wafers.

6.3.2 Polysilicon Deposition

Silicon is deposited in an LPCVD system using thermal decomposition of silane:

\[ \text{SiH}_4 \xrightarrow{600\, ^\circ\text{C}} \text{Si} + 2 \text{H}_2 \]  

(6.10)

Low-pressure systems (25 to 150 Pa) use either 100% silane or 20 to 30% silane diluted with nitrogen. A temperature between 600 and 650 °C results in deposition of polysilicon material at a rate of 100 to 200 Å/min. A less commonly used deposition occurs between
850 and 1050 °C in a hydrogen atmosphere. The higher temperature overcomes a reduction in deposition rate caused by the hydrogen carrier gas.

Polysilicon can be doped by diffusion or ion implantation or during deposition in situ by the addition of dopant gases such as phosphine, arsine, or diborane. The addition of diborane greatly increases the deposition rate, whereas the addition of phosphine or arsine substantially reduces the deposition rate.

Polysilicon is often deposited as undoped material and is then doped by diffusion. High-temperature diffusion occurs much more rapidly in polysilicon than in single-crystal silicon, and the polysilicon film is typically saturated with the dopant to achieve as low a resistivity as possible for interconnection purposes. Resistivities of 0.01 to 0.001 ohm-cm can be achieved in diffusion-doped polysilicon. Ion implantation typically yields a lower active-impurity density in the polysilicon film, and ion-implanted polysilicon exhibits a resistivity about ten times higher than that achieved by high-temperature diffusion.

### 6.3.3 Silicon Dioxide Deposition

Silicon dioxide films can be deposited using a variety of reactions and temperature ranges, and the films can be doped or undoped. Phosphorus-doped oxide can be used as a passivation layer over a completed integrated circuit or as the insulating medium in multilevel metal processes (which will be discussed in the next chapter). Silicon dioxide containing 6 to 8% phosphorus by weight will soften and flow at temperatures between 1000 and 1100 °C. This "P-glass reflow" process is often used to improve step coverage and provide a smoother topography for later process steps. SiO₂ with lower concentrations of phosphorus will not reflow properly, and higher concentrations can corrode aluminum if moisture is present. Oxide doped with 5 to 15% by weight of various dopants can also be used as a diffusion source.

Deposition of silicon dioxide over aluminum must occur at a temperature below the silicon-aluminum eutectic point of 577 °C (see Chapter 7). A reaction between silane and oxygen is commonly used between 300 and 500 °C:

\[
\text{SiH}_4 + O_2 \rightarrow \text{SiO}_2 + 2\text{H}_2 \quad (6.11)
\]

The oxide may be doped with phosphorus using phosphine:

\[
4\text{PH}_3 + 5\text{O}_2 \rightarrow 2\text{P}_2\text{O}_5 + 6\text{H}_2 \quad (6.12)
\]

Oxide passivation layers can be deposited at atmospheric pressure using the continuous reactor of Fig. 6.8a, or they can be deposited at reduced pressure in an LPCVD system, as in Fig. 6.8b.

Deposition of SiO₂ films prior to metallization can be performed at higher temperatures, which gives a wider choice of reactions and results in better uniformity and step coverage. For example, a dichlorosilane reaction with nitrous oxide is an LPCVD system at approximately 900 °C,

\[
\text{SiCl}_2\text{H}_2 + 2\text{N}_2\text{O} \rightarrow \text{SiO}_2 + 2\text{N}_2 + 2\text{HCl} \quad (6.13)
\]

used to deposit insulating layers of SiO₂ on wafer surfaces.

Decomposition of the vapor produced from a liquid source, tetraethylorthosilicate (TEOS), can also be used in an LPCVD system between 650 and 750 °C.

\[
\text{Si(OC}_2\text{H}_5)_4 \rightarrow \text{SiO}_2 + \text{by-products} \quad (6.14)
\]

Deposition based on the decomposition of TEOS provides excellent uniformity and step coverage. Oxide doping may be accomplished in the LPCVD systems by adding phosphine, arsine, or diborane.

A comparison of some of the properties of various CVD oxides is given in Table 6.1.

### 6.3.4 Silicon Nitride Deposition

As discussed in Chapter 3, silicon nitride is used as an oxidation mask in recessed oxide processes. Silicon nitride is also used as a final passivation layer because it provides an excellent barrier to both moisture and sodium contamination. Composite films of oxide and nitride are being investigated for use as very thin gate insulators in scaled VLSI devices, and they are also used as the gate dielectric in electrically programmable memory devices.

Both silane and dichlorosilane will react with ammonia to produce silicon nitride. The slane reaction occurs between 700 and 900 °C at atmospheric pressure:

\[
3\text{SiH}_4 + 4\text{NH}_3 \rightarrow \text{Si}_3\text{N}_4 + 12\text{H}_2 \quad (6.15)
\]

Dichlorosilane is used in an LPCVD system between 700 and 800 °C:

\[
3\text{SiCl}_2\text{H}_2 + 4\text{NH}_3 \rightarrow \text{Si}_3\text{N}_4 + 6\text{HCl} + 6\text{H}_2 \quad (6.16)
\]

| Table 6.1 Properties of Various Deposited Oxides. (After ref. [12].) |
|-------------------|-------------------|-------------------|-------------------|
| Source            | Deposition Temperature (°C) | Composition | Conformal Step Coverage | Dielectric Strength (MV/cm) | Etch Rate (Å/min) [100:1 H₂O:HF] |
| Silane            | 450               | SiO₂(H)        | No                 | 8                 | 60                          |
| Dichlorosilane    | 900               | SiO₂(Cl)       | Yes                | 10                | 30                          |
| TEOS              | 700               | SiO₂           | Yes                | 10                | 30                          |
| Plasma            | 200               | SiO₂(H)        | No                 | 5                 | 400                         |
Thermal growth of silicon nitride is also possible but not very practical. Silicon nitride will form when silicon is exposed to ammonia at temperatures between 1000 and 1100 °C, but the growth rate is very low.

Plasma systems may also be used for the deposition of silicon nitride. Silane will react with a nitrogen discharge to form plasma nitride (SiN):

$$2\text{SiH}_4 + \text{N}_2 \rightarrow 2\text{SiNH} + 3\text{H}_2 \quad (6.17)$$

Silane will also react with ammonia in an argon plasma:

$$\text{SiH}_4 + \text{NH}_3 \rightarrow \text{SiNH} + 3\text{H}_2 \quad (6.18)$$

LPCVD films are hydrogen-rich, containing up to 8% hydrogen. Plasma deposition does not produce stoichiometric silicon nitride films. Instead, the films contain as much as 20 to 25% hydrogen. LPCVD films have high internal tensile stresses, and films thicker than 2000 Å may crack because of this stress. On the other hand, plasma-deposited films have much lower tensile stresses.

The resistivity (10⁶ ohm-cm) and dielectric strength (10 MV/cm) of the LPCVD nitride film are better than those of most plasma films. Resistivity of plasma nitride can range from 10⁶ to 10⁸ ohm-cm, depending on the amount of nitrogen in the film, while the dielectric strength ranges between 1 and 5 MV/cm.

### 6.3.5 CVD Metal Deposition

Many metals can be deposited by CVD processes. Molybdenum (Mo), tantalum (Ta), titanium (Ti), and tungsten (W) are all of interest in today’s processes because of their low resistivity and their ability to form silicides with silicon (see Chapter 7). Aluminum can be deposited from a metalorganic compound such as tri-isobutyl aluminum, but this technique has not been commonly used because many other excellent methods of aluminum deposition are available.

Tungsten can be deposited by thermal, plasma, or optically assisted decomposition of WF₆:

$$\text{WF}_6 \rightarrow \text{W} + 3\text{F}_3 \quad (6.19)$$

or through reduction with hydrogen:

$$\text{WF}_6 + 3\text{H}_2 \rightarrow \text{W} + 6\text{HF} \quad (6.20)$$

Mo, Ta, and Ti can be deposited in an LPCVD system through reaction with hydrogen. The reaction is the same for all three metals:

$$2\text{MCl}_5 + 5\text{H}_2 \rightarrow 2\text{M} + 10\text{HCl} \quad (6.21)$$

where M stands for any one of the three metals mentioned above.

### 6.4 EPITAXY

Chemical vapor deposition processes can be used to deposit silicon onto the surface of a silicon wafer. Under appropriate conditions, the silicon wafer acts as a seed crystal, and a single-crystal silicon layer is grown on the surface of the wafer. The growth of a crystalline silicon layer from the vapor phase is called vapor-phase epitaxy (VPE), and it is the most common form of epitaxy used in silicon processing. In addition, liquid-phase epitaxy (LPE) and molecular-beam epitaxy (MBE) are being used widely in GaAs technology.

Epitaxial growth was first used in integrated-circuit processing to grow single-crystal p-type layers on p-type substrates for use in standard buried-collector bipolar processing. More recently, it has been introduced into CMOS VLSI processes where lightly doped layers are grown on heavily doped substrates of the same type (n on n⁺ or p on p⁺) to help suppress a circuit-failure mode called latchup.

#### 6.4.1 Vapor-Phase Epitaxy

Silicon epitaxial layers are commonly grown with silicon deposited from the gas phase. A basic model for the process is given in Fig. 6.9. At the silicon surface, the flux \(J_s\) of gas molecules is determined by

$$J_s = k_s N_e \quad (6.22)$$

in which \(k_s\) is the surface-reaction rate constant and \(N_e\) is the surface concentration of the molecule involved in the reaction. In the steady state, this flux must equal the flux \(J_e\) of molecules diffusing in from the gas stream. The flux \(J_e\) may be approximated by

$$J_e = \frac{(D_e/\delta)(N_e - N_s)}{h_e (N_e - N_s)} \quad (6.23)$$

![Fig. 6.9 Schematic model for the epitaxial growth process.](image)
in which $\overline{D}^*_g$ is an effective diffusion constant for the gas molecule and $\delta$ is the distance over which the diffusion is taking place. The ratio $\frac{\overline{D}^*_g}{\delta}$ is called the vapor-phase mass-transfer coefficient, $h_g$. Equating $J_x$ and $J_y$ yields the flux impinging on the surface of the wafer. The growth rate is equal to the flux divided by the number $N$ of molecules incorporated per unit volume of film.

$$v = \frac{J_x}{N} = \frac{k_i h_x N_x}{k_s + h_x N}$$  \hspace{1cm} (6.24)$$

If $k_i \gg h_x$, then growth is said to be mass-transfer-limited, and

$$v = \frac{h_x N_x}{N}$$  \hspace{1cm} (6.25)$$

If $h_x \gg k_i$, then growth is said to be surface-reaction-limited, and

$$v = \frac{k_i N_x}{N}$$  \hspace{1cm} (6.26)$$

Figure 6.10 shows epitaxial growth rate as a function of temperature. Chemical reactions at the surface tend to follow an Arrhenius relationship characterized by an activation energy $E_A$, whereas the mass-transfer process tends to be independent of temperature. These two regions show up clearly in this figure. At low temperatures, the growth rate follows an Arrhenius relationship with an activation energy of approximately 1.5 eV. At higher temperatures, the growth rate becomes independent of temperature. In order to have good growth-rate control and to minimize sensitivity to variations in temperature, epitaxial growth conditions are usually chosen to yield a mass-transfer-limited growth rate.

Three common types of VPE reactors, the horizontal, vertical, and barrel systems, are shown in Fig. 6.11. The susceptor that supports the wafers is made of graphite and is heated by RF induction in the horizontal and vertical reactors and by radiant heating in the barrel reactor.

Silicon tetrachloride ($\text{SiCl}_4$), silane ($\text{SiH}_4$), dichlorosilane ($\text{SiH}_2\text{Cl}_2$), and trichlorosilane ($\text{SiHCl}_3$) have all been used for silicon VPE. Silicon tetrachloride has been widely used in industrial processing:

$$\text{SiCl}_4\,\text{(gas)} + 2\text{H}_2\,\text{(gas)} \leftrightarrow \text{Si}\,\text{(solid)} + 4\text{HCl}\,\text{(gas)}$$  \hspace{1cm} (6.27)$$

This reaction takes place at approximately 1200 °C and is reversible. If the carrier gas coming into the reactor contains hydrochloric acid, etching of the surface of the silicon wafer can occur. This in situ etching process can be used to clean the wafer prior to the start of epitaxial deposition.
A second reaction competes with the epitaxial deposition process:

\[
\text{SiCl}_4(\text{gas}) + \text{Si(solid)} \leftrightarrow 2\text{SiCl}_2(\text{gas})
\]  

\((6.28)\)

This second reaction also etches the silicon from the wafer surface. If the concentration of \(\text{SiCl}_4\) is too high, etching of the wafer surface will take place rather than epitaxial deposition. Figure 6.12 shows the effect of \(\text{SiCl}_4\) concentration on the growth of epitaxial silicon. The growth rate initially increases with increasing \(\text{SiCl}_4\) concentration, peaks, then decreases. Eventually, growth stops and the etching process becomes dominant. If the growth rate is too high, a polysilicon layer is deposited rather than a layer of single-crystal silicon.

Epitaxial growth can also be achieved by the pyrolytic decomposition of silane:

\[
\text{SiH}_4 \rightarrow_{600^\circ C} \text{Si} + 2\text{H}_2
\]  

\((6.29)\)

The reaction is not reversible and takes place at low temperatures. In addition, it avoids the formation of HCl gas as a reaction by-product. However, careful control of the reactor is needed to prevent formation of polysilicon rather than single-crystal silicon layers. The presence of any oxidizing species in the reactor can also lead to contamination of the epitaxial layer by silica dust.

![Graph showing silicon epitaxial growth rate as a function of SiCl4 concentration.](image)

**Fig. 6.12** Silicon epitaxial growth rate as a function of SiCl4 concentration. Polysilicon deposition occurs for growth rates exceeding 2 \(\mu\text{m/min}\). Etching of the surface will occur for mole fraction concentrations exceeding 28%. Copyright, 1985, John Wiley & Sons, Inc., with permission from ref. [1].

### 6.4.2 Doping of Epitaxial Layers

Epitaxial layers may be doped during the growth process by adding impurities to the gas used for deposition. Arsenic, diborane, and phosphine are the most convenient sources of the common impurities. The resistivity of the epitaxial layer is controlled by varying the partial pressure of the dopant species in the gas supplied to the reactor. The addition of arsine or phosphine tends to slow down the rate of epitaxial growth, while the addition of diborane tends to enhance the epitaxial growth rate.

Lightly doped epitaxial layers are often grown on more heavily doped substrates, and "autodoping" of the epitaxial layer can occur during growth. Impurities can evaporate from the wafer or may be liberated by chlorine etching of the surface during deposition. The impurities are incorporated into the gas stream, resulting in doping of the growing layer. As the epitaxial layer grows, less dopant is released from the wafer into the gas stream, and the impurity profile eventually reaches a constant level determined by the doping in the gas stream.

During deposition, the substrate also acts as a source of impurities which diffuse into the epitaxial layer. This "out-diffusion" will be discussed more fully in the next section. Both autodoping and out-diffusion cause the transition from the doping level of the substrate to that of the epitaxial layer to be less abrupt than desired. The effects of autodoping and out-diffusion are illustrated in Fig. 6.13.

### 6.4.3 Buried Layers

Out-diffusion is a common problem that occurs with the buried layer in bipolar transistors. In order to reduce the resistance in series with the collector of the bipolar transistor, heavily doped \(n\)-type regions are diffused into the substrate prior to the growth of an \(n\)-type epitaxial layer. During epitaxy, impurities diffuse upward from the heavily doped buried-layer regions.

Diffusion of impurities from the substrate during epitaxial growth is modeled by the diffusion equation with a moving boundary,\(^{14}\) as in Fig. 6.14:

\[
D \frac{\partial^2 N}{\partial x^2} = \frac{\partial N}{\partial t} + v_x \frac{\partial N}{\partial x}
\]

\((6.30)\)

in which \(v_x\) is the rate of growth of the epitaxial layer.

Two specific solutions of eq. \((6.30)\) are applicable to epitaxial layer growth. The first case is the growth of an undoped epitaxial layer on a uniformly doped substrate. The boundary conditions are \(N(x, 0) = N_s = N(\infty, t)\), and the flux \(J_x = (h + v_x)N(0, t)\) where \(h\) is the mass-transfer coefficient which characterizes the escape rate of dopant atoms from the silicon into the gas. Normally, \(h \ll v_x\). A change of variables from \(x\) to \(x' = x - v_x t\) simplifies eq. \((6.30)\) and gives an approximate solution for \(N(x, t)\):

\[
N(x, t) = \frac{N_s}{2} \left[ 1 + \text{erf} \frac{x - x_{eq}}{2\sqrt{D_n t}} \right]
\]

\((6.31)\)
These two cases gives a good approximation to diffusion which occurs during epilaxial growth:

\[ N(x,t) = N_i(x,t) + N_c(x,t) \]  

(6.33)

\( N_i \) represents the doping in the substrate, and \( N_c \) is the doping intentionally introduced into the epilaxial layer. \( D_i \) and \( D_c \) represent the diffusion coefficients of the impurity species in the substrate and epilaxial layer, respectively. Figure 6.13 shows diffusion profiles for a phosphorus-doped epilaxial layer grown at various rates on an antimony-doped substrate. The curves were produced using eq. (6.33).

An additional problem occurs during epilaxial growth. The oxidation and lithographic processing steps used during formation of a buried layer result in a step of as much as 0.2 \( \mu \)m around the perimeter of the buried layer. Epilaxial growth on this nonplanar surface causes the pattern to shift during growth, as illustrated in Fig. 6.15. Pattern shift is difficult to predict, may be as large as the epilaxial layer thickness, and must be accounted for during the design of subsequent mask levels.

### 6.4.4 Liquid-Phase and Molecular-Beam Epitaxy

In liquid-phase epitaxy, the substrate is brought into contact with a solution containing the material to be deposited in liquid form. The substrate acts as a seed for material crystallizing directly from the solute. Growth rates typically range between 0.1 and 1 \( \mu \)m/min.

In the molecular-beam epitaxy process, the crystalline layer is formed by deposition from a thermal beam of atoms or molecules. Deposition is performed in ultrahigh-
Sputtering uses energetic ions such as argon to bombard a target material and dislodge atoms from the surface of the target. The dislodged atoms are deposited on the surface of the wafer. Direct-current sputtering systems can be used to deposit conductive materials, and RF sputtering can be used to deposit insulators. Sputtering can be used to deposit composite materials in which the deposited film maintains the same composition as the source material. Sputtering also uses higher pressures than evaporation. The much shorter mean free paths which result yield a deposition with freedom from shadowing and much better step coverage.

Low-pressure and atmospheric chemical vapor deposition (CVD) systems deposit films from chemical reactions taking place in a gas stream passing over the wafer. Polysilicon, silicon dioxide, silicon nitride, and metals can all be deposited using CVD techniques. A special type of CVD deposition called epitaxy results in the growth of single-crystal silicon films on the surface of silicon wafers. Out-diffusion and autodoping cause problems with impurity profile control during epitaxial layer growth.

In a modern bipolar or MOS fabrication process, one can expect to find evaporation, sputtering, and chemical vapor deposition techniques all used somewhere in the process flow.

REFERENCES


FURTHER READING


PROBLEMS

6.1 A silicon wafer sits on a bench in the laboratory at a temperature of 300 K and a pressure of 1 atm. Assume that the air consists of 100% oxygen. How long does it take to deposit one atomic layer of oxygen on the wafer surface, assuming 100% adhesion?

6.2 Calculate the impingement rate and mean free path for oxygen molecules \((M = 32)\) at 300 K and a pressure of \(10^{-4} \text{ Pa}\). What is this pressure in torr?

6.3 An ultrahigh vacuum system operates at a pressure of \(10^{-8} \text{ Pa}\). What is the concentration of residual air molecules in the chamber at 300 K?

6.4 The partial pressure of a material being deposited in a vacuum system must be well above the residual background gas pressure if reasonable deposition rates are to be achieved. What must the partial pressure of aluminum be to achieve a deposition rate of 100 nm/min? Assume close packing of spheres with a diameter of 5 Å, 100% adhesion of the impinging aluminum and 300K.

6.5 A wafer 100 mm in diameter is mounted in an electron-beam evaporation system in which the spherical radius is 40 cm. Use eq. (6.7) to estimate the worst-case variation in film thickness between the center and edges of the wafer for an evaporated aluminum film 1 µm thick.

6.6 A MBE system must operate under ultrahigh vacuum conditions to prevent the formation of undesired atomic layers on the surface of the substrate. What pressure of oxygen can be permitted at 300K if formation of a monolayer of contamination can be permitted after the sample has been in the chamber for no less than 4 hr?

6.7 (a) Calculate the growth rate of a silicon layer from a SiCl₄ source at 1200 °C. Use \(h_s = 1 \text{ cm/sec}, k_s = 2 \times 10^6 \exp(-1.9/KT) \text{ cm/sec, and } N_s = 3 \times 10^{25} \text{ atoms/cm}^2\).

(b) What is the change in growth rate if the temperature is increased by 25 °C?

(c) At what temperature does \(k_s = h_s\)? What is the growth rate at this temperature?

(d) What is the value of \(E_A\) in Fig. 6.10?

6.8 Use eqs. (6.31) and (6.32) to model the case of a 10-µm n-type epitaxial layer \((N_E = 1 \times 10^{16} /\text{cm}^3)\) grown on a p-type substrate \((N_s = 1 \times 10^{18} /\text{cm}^3)\). Plot the impurity profile in the epitaxial layer and substrate assuming that the layer was grown at a rate of 0.2 µm/min at a temperature of 1200 °C. Assume boron and phosphorus are the impurities. Find the location of the pn junction.

6.9 Compare and discuss the advantages and disadvantages of evaporation, sputtering, and chemical vapor deposition.

6.10 A 1-kg source of aluminum is used in an E-beam evaporation system. How many 100-mm wafers can be coated with a 1-µm Al film before the source material is exhausted? Assume that 15% of the evaporated aluminum actually coats a wafer. (The rest is deposited on the inside of the electron-beam system.)

6.11 A silicon wafer 75 mm in diameter is centered 100 mm above a small planar evaporation source. Calculate the ratio of thickness between the center and edges of the wafer using eq. (6.7), following a 1-µm film deposition.