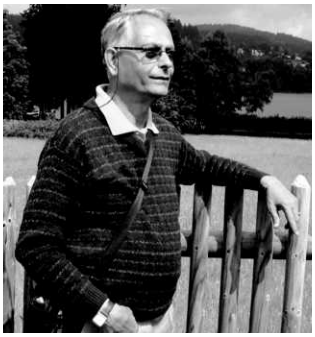
IC FABRICATION TECHNOLOGY

About the Author



Gouranga Bose received an MSc (Physics) degree in 1968 from Jabalpur University. He then obtained MTech and PhD degrees in 1971 and 1978 respectively in the area of optics from IIT Delhi. Thereafter, he joined Microelectronics Group, Centre for Applied Research in Electronics (CARE), IIT Delhi, in 1981 as a faculty. He retired from IIT Delhi in 2006 and continued as IRD Fellow at IIT Delhi till 2009. While in IIT Delhi, he was actively involved in teaching and research in CARE and the Electrical Engineering Department. At present, he is Dean of Postgraduate Studies, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, and faculty in Electronics and Instrumentation Engineering, Institute of Technical Education and Research, SOA University, since 2010.

Dr Bose has published innumerable papers in several reputed journals. He has also presented several papers in multiple international and national conferences and many of his papers have been cited in reputed journals. He has guided many PhD, PG and UG students, and has worked in more than 50 sponsored research and development projects. He has contributed to various aspects of microelectronics technology and made numerous microelectronics devices. He has been a member of monitoring committees of many Government of India funded projects of different organizations.

His current research interest is synthesis and characterization of thin films for sensor applications, CMOS compatible sensors and actuator fabrication, and integration of sensors with signal-processing chips.

IC FABRICATION TECHNOLOGY

Gouranga Bose

Dean of Postgraduate Studies
Institute of Technical Education and Research (ITER)
Siksha 'O' Anusandhan University
Bhubaneswar, Odisha
Ex-IRD Fellow, IIT Delhi

McGraw Hill Education (India) Private Limited
NEW DELHI

McGraw Hill Education Offices
New Delhi New York St Louis San Francisco Auckland Bogotá Caracas
Kuala Lumpur Lisbon London Madrid Mexico City Milan Montreal
San Juan Santiago Singapore Sydney Tokyo Toronto

McGraw Hill Education (India) Private Limited

Published by McGraw Hill Education (India) Private Limited
P-24, Green Park Extension, New Delhi 110 016

IC Fabrication Technology

McGraw Hill Education (India) Edition 2014

ISBN 13: 9781259029585
ISBN 10: 1259029581

Vice President and Managing Director: *Ajay Shukla*

Head—Higher Education (Publishing and Marketing): *Vibha Mahajan*

Senior Publishing Manager (SEM & Tech. Ed.): *Shalini Jha*
Editorial Executive: *Koyel Ghosh*
Manager—Production Systems: *Satinder S Baveja*
Assistant Manager—Editorial Services: *Sohini Mukherjee*
Assistant Manager—Production: *Anjali Razdan*

Assistant General Manager (Marketing)—Higher Education: *Vijay Sarathi*
Senior Product Specialist: *Tina Jajoriya*
Senior Graphic Designer—Cover: *Meenu Raghav*

General Manager—Production: *Rajender P Ghansela*
Manager—Production: *Reji Kumar*

Information contained in this work has been obtained by McGraw Hill Education (India), from sources believed to be reliable. However, neither McGraw Hill Education (India) nor its authors guarantee the accuracy or completeness of any information published herein, and neither McGraw Hill Education (India) nor its authors shall be responsible for any errors, omissions, or damages arising out of use of this information. This work is published with the understanding that McGraw Hill Education (India) and its authors are supplying information but are not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought.

Typeset at The Composers, 260, C.A. Apt., Paschim Vihar, New Delhi 110 063 and printed at
A P Offset Pvt. Ltd., 25/487, Zulfe Bengal, Dilshad Garden, New Delhi 110 095

Cover Printer: A P Offset

RZXLCRDPDLZDC

To

My Parents

Dr Naresh Chandra Bose
Mrs Nihar Kana Bose

&

My Wife

Mitali Bose

Contents

Preface

Overview

Semiconductor device fabrication is the process used to create the integrated circuits that are present in everyday electrical and electronic devices. It is a multiple-step sequence of photolithographic and chemical processing during which electronic circuits are gradually created on a wafer—a thin slice of semiconductor material, such as a silicon crystal used in the fabrication of integrated circuits and other microdevices. The wafer—made of pure semiconducting material—serves as the substrate for microelectronic devices built in and over the wafer and undergoes many microfabrication process steps. Silicon is almost always used, but various compound semiconductors are used for specialized applications.

The revolution in microelectronics has led to a rapid growth in all disciplines, especially in the field of electronics by virtue of the "Integrated Circuit" (IC). This is because ICs consume low power, offer high reliability, higher operational speeds and increased signal-processing capability. This was made possible because ICs contain transistors that can be scaled down and fabricated. Furthermore, increase of transistor density in an IC chip can accommodate more functional blocks (sub-circuit), and it is on these lines that technological advancements will be seen in the years to come. In the proposed text, all essential steps to IC fabrication are covered sequentially. Prior to engaging in IC fabrication, a brief review of semiconductor physics in the context of IC fabrication is covered, so that students do not have difficulties in understanding the IC fabrication physics and process steps.

Target Audience

IC fabrication technology is a course offered to both undergraduate and graduate students of ECE and VLSI specialization respectively. For graduate studies, this is mostly offered as a compulsory paper while for UG students, this is often an elective subject. Other than students of Communication Engineering disciplines, this book will also be useful to students of Electrical and Electronics Engineering, Computer Science Engineering and Information Technology.

About the Book

The facets of IC fabrication technology are important for students of VLSI to help them understand the implementation of VLSI design in a better way. *IC Fabrication Technology* is aimed at the novice reader—to help him/her develop a practical appreciation of the subject area, especially the process for fabrication. In keeping with this ideology, the book has been written in a highly illustrative manner and a number of examples have been provided which reflect practical problems faced during the processes of fabrication. It provides an introduction to major fabrication process steps of silicon-based CMOS. Highly suitable for postgraduates, and particularly for undergraduate students, this textbook can be used for a full-semester course on integrated circuit fabrication. Furthermore, this book may help to generate a few experiments too. The prerequisite to study this book is a basic understanding of semiconductor physics and devices, and analog and digital circuit design. Students interested in further study can take up courses like MOS devices technology and MEMS technology.

Salient Features

- Meticulous detailing of all important and critical steps to silicon IC fabrication
- Discussions on practical problems encountered during fabrication
- Dedicated chapter on Mask fabrication
- Sneak peek into the emerging world of ULSI technology in the chapter on *Submicron Technology*
- Simplified, exemplified and illustrative presentation:
  - Diagrams: *154*
  - Solved Examples: *12*
  - Description Problems: *58*
  - Multiple-Choice Questions: *108*

Chapter Organization

Chapter 1 begins with the historical background, evolution of transistor technology, and goes on to explore the trend of integrated circuit overview of discrete silicon transistor and silicon semiconductor physics. **Chapter 2** gives important information and requirements of silicon wafers and wafer-fabrication techniques. **Chapter 3** describes the fabrication sequences, methodology and essential fabrication processes of MOS transistors in brief.

This chapter describes how a wafer is processed through logical sequence from one individual fabrication process step to the next individual process to realize MOS transistors.

The rest of the chapters are arranged on the basis of MOS fabrication process sequence. **Chapter 4** describes the oxidation mechanism, kinetics of oxidation, oxide-film-thickness measurement and many essential oxidation-related issues. **Chapter 5** covers the types and fabrication of masks. **Chapter 6** covers lithography steps, techniques of lithography and mask-to-wafer alignment. **Chapter 7** describes etching techniques of various films. This step follows the lithography process. **Chapters 8 and 9** deal with the doping mechanisms of dopant diffusion in silicon. High-temperature dopant diffusion is described in Chapter 8, whereas dopant lodging in silicon by ion-implantation is described in Chapter 9. **Chapter 10** covers the last process step of MOS fabrication and covers the various techniques of film deposition, etching of films, and advantages and disadvantages of films. **Chapter 11** briefly describes the latest trends and techniques of integrated circuit fabrication. It also deals with the fabrication issues and current research in brief.

Online Learning Centre

The book is accompanied by an online learning center, available at http://www.mhhe.com/bose/ic that offers PowerPoint slides with diagrams and notes for effective presentation.

Acknowledgements

First of all, I thank all the reviewers responsible for making the text acceptable to a large user community. Their names are given below:

Jaynendra Kumar Rai	*Amity School of Engineering and Technology* *Noida, Uttar Pradesh*
Sushanta Kumar Mandal	*School of Electronics Engineering* *Kalinga Institute of Industrial Technology (KIIT) University* *Bhubaneswar, Odisha*
Rajat Mahapatra	*National Institute of Technology, Durgapur* *Durgapur, West Bengal*
Kamla Kanta Mahapatra	*National Institute of Technology (NIT) Rourkela* *Rourkela, Odisha*

Partha Bhattacharya	*Bengal Engineering and Science University* *Shibpur, West Bengal*
Ananya Dastidar	*College of Engineering and Technology* *Bhubaneswar, Odisha*
Ramesha C K	*Birla Institute of Technology and Science (BITS) Pilani* *KK Birla Goa Campus, Goa*
Anurag Lakhlani	*Marwadi's Education Foundation Group of Institutions* *Rajkot, Gujarat*
Nilesh D Patel	*Laljibhai Chaturbhai Institute of Technology* *Ahmedabad, Gujarat*
Usha Mehta	*Institute of Technology, Nirma University* *Ahmedabad, Gujarat*
S Ramasamy	*Manipal Institute of Technology* *Manipal, Karnataka*
Shailendra Kumar Tiwari	*Manipal Institute of Technology* *Manipal, Karnataka*
Hanumantha Rao	*National Institute of Technology (NIT) Warangal* *Warangal, Karnataka*

I genuinely appreciate the efforts of Professor Arun Kumar, CARE IITD; Dr Dilip Ranganathan, Physics Department, IITD; and Jyoti Prakas Kar, Physics Department, NIT, Rourkela, in encouraging and helping me while writing this book.

I render my thanks to Dr Satyanarayan Bhuyan, Ms Farida A Ali and Satya Prakash Narayan Das, EIE Department; and Tara Prasanna Dash, ECI Department, ITER, Siksha 'O' Anusandhan University, for giving useful suggestions.

I sincerely express my gratitude to IIT Delhi, New Delhi, and Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, for extending me all kinds of logistic support for writing this book.

Thanks are also due to the student community and my teachers who made *me* a teacher. I thank the McGraw-Hill Education (India) team, who made my dream a reality—Shalini Jha, Koyel Ghosh, Sohini Mukherjee, Satinder Singh, and Anjali Razdan—for their constant support and excellent interaction. I am also grateful to my family members who supported me in all stages of preparation of this text.

Last, but certainly not the least, I thank my wife, Mitali, for her constant encouragement and patience during evenings, weekends and holidays while writing this book. Without her support, this book could not have been written.

Publisher's Note

McGraw-Hill Education (India) invites suggestions and comments from you, all of which can be sent to *tmh.elefeedback@gmail.com* (kindly mention the title and author name in the subject line).

Piracy-related issues may also be reported**.**

CHAPTER

1

Overview of Metal-Oxide-Semiconductor (MOS) Transistor

1.1 INTRODUCTION

The revolution in microelectronics has led to a rapid growth in all disciplines, especially in the field of military and communication electronics, since the past half century. This is because of the integration of tiny microelectronic devices such as transistors, resistors and capacitors in a particular fashion called **Integrated Circuit (IC)**. These ICs are fabricated on the surface of a circular, single-crystal semiconductor called the **substrate**. As the thickness of the circular substrate is thin with respect to its diameter, it is also called a **semiconductor wafer**. The fabricated area of the IC on the silicon wafer is referred to as **die**; and to reduce the cost of the IC, a large number of these dies are made on a single silicon wafer. When a single die is packaged, it is called a **chip**.

Generally, ICs are fabricated on a single-crystal silicon wafer. This is because silicon has excellent electrical and physical properties that are required for microelectronic devices; and they are considered more IC fabrication friendly than any other semiconductor material. Furthermore, silicon dioxide film, which is one of the vital components used in microelectronics, can be made easily with less silicon dioxide/silicon interface defects as compared to other semiconductor materials. These properties of silicon are also exploited to scale down the microelectronic devices that in turn reduce the operating power and cost, and increase device reliability. In addition, more functional ICs can be put onto a chip, increasing the operational speeds of signal processing phenomenally. These have with time, led semiconductor industries to scale down microelectronic devices for better products. The reduction of device size with time was first reported by Gordon Moore of Intel around 1970 and is known as **Moore's law**.

Generally, signal-processing ICs are made of MOS microelectronic devices, so, henceforth, microelectronic devices will be addressed as MOS transistors or simply MOS.

1.2 MOORE'S LAW

Moore observed that the MOS transistor density doubled roughly every two years; and his finding became the future guideline for the entire semiconductor industry. On the basis of Moore's law, in 1994, the Semiconductor Industry Association (SIA) of USA made guidelines for the semiconductor industry which were collectively known as the **National Technology Roadmap for Semiconductors** or NTRS, as shown in Fig. 1.1. The roadmap projects how the IC technology will advance in terms of downsizing the device size and increasing the chip size. SIA identified a few ICs such as Dynamic Random Access Memories (DRAM), Static Random Access Memories (SRAM), microprocessors, and Application Specific Integrated Circuits (ASIC) as the vehicles in terms of volume or specific market requirements. It is important to mention that the reduction in transistor size took place in a series of steps which is referred to as the **generation of IC technology**. Usually, gate length is defined by the generation of IC technology. Presently, the IC technological roadmap is projected from the year 1977 to the year 2012. In this span of 35 years, the transistor gate length has reduced from around 0.25 μm generation to 0.05 μm generation, as shown in Fig. 1.1.

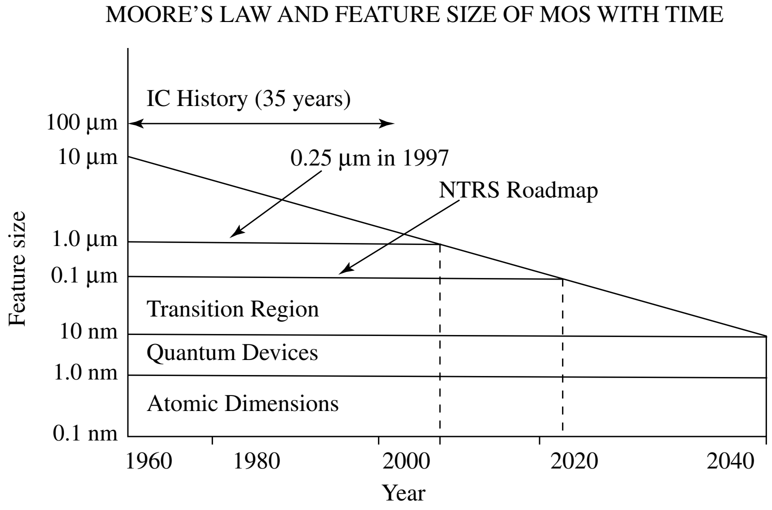


Fig. 1.1 National Technology Roadmap for Semiconductor (NTRS)

In recent days, ICs are made of Metal-Oxide-Semiconductor (MOS) transistors using the planar technology, especially for digital processing chips. These MOS transistors are made over or below the surface of the silicon wafer by localised selected elements diffusion (doping) and stacked with films of different materials over the wafer surface, as shown in Fig. 1.2. The detailed fabrication process is described in Chapter 3. MOS is made using planar technology that has an advantage in scaling down of MOS transistors. This NTRS projected target could not be achieved because of the lack of new-generation fabrication equipment, materials, fabrication process modelling, and simulation. To understand Moore's law and NTRS, it is essential to describe the structure of a MOS transistor, fabrication techniques, electrical characteristics, device physics, and scaling issues.

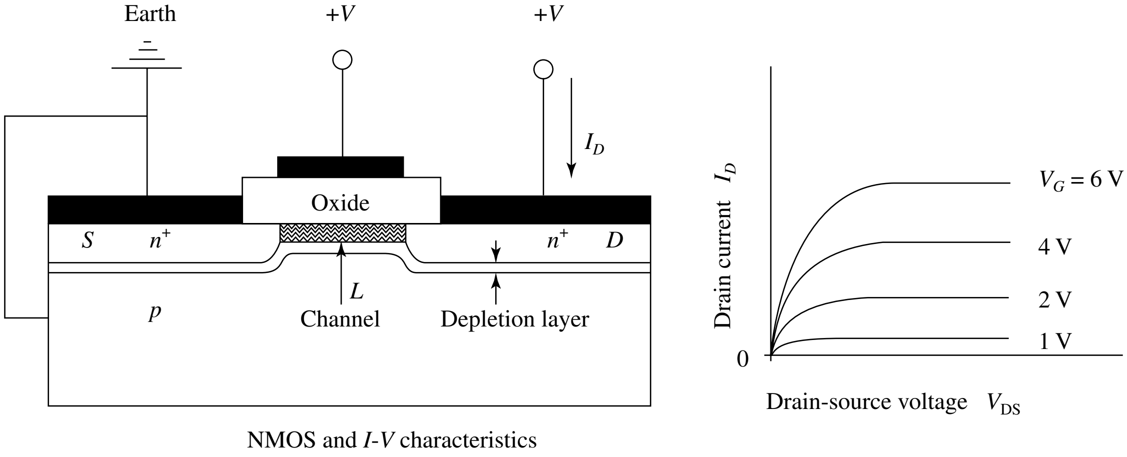


Fig. 1.2 Architecture and the electrical configuration of NMOS

1.2.1 Metal-Oxide-Semiconductor (MOS) Transistor

The MOS transistor is also called Insulated Gate Field Effect Transistor (IGFET) or simply Field Effect Transistor (FET). The MOS transistor is a surface device; hence its electrical characteristics are primarily determined by the silicon-surface properties and their interaction with the external electric field.

There are two types of MOS transistors, namely, the NMOS transistor and the PMOS transistor. In an NMOS transistor, electrons are the current carriers, whereas in a PMOS transistor, holes are the current carriers. The working principles of both these types of transistors are the same, only their electrical connections (polarities) are different. Apart from transistor biasing polarities, NMOS transistors are made on uniformly doped boron bulk silicon (*P*-type silicon) wafers, and PMOS transistors are made on uniformly doped phosphorous (*N*-type silicon) wafers. However, both types of transistors can also be made on the same silicon wafer, if localised *P*-type and *N*-type silicon are made (converted) into a single silicon wafer. Once the localised *P*-type and *N*-type silicon substrates are made, NMOS on *P*-type and PMOS on *N*-type transistors are made respectively. These localised *P*-type and *N*-type substrates are called ***P*-well** and ***N*-well** respectively. These two types of complementary transistors are made in a single wafer; hence it is called Complementary MOS or CMOS. The architecture and the electrical biasing of NMOS, PMOS and CMOS transistors are shown in Figs. 1.2, 1.3 and 1.4, respectively.

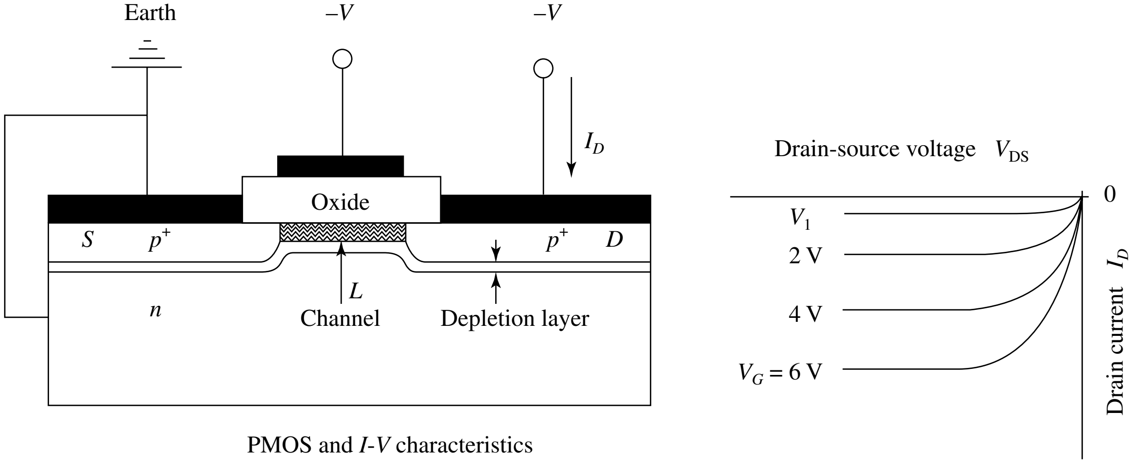


Fig. 1.3 Architecture and the electrical configuration of PMOS

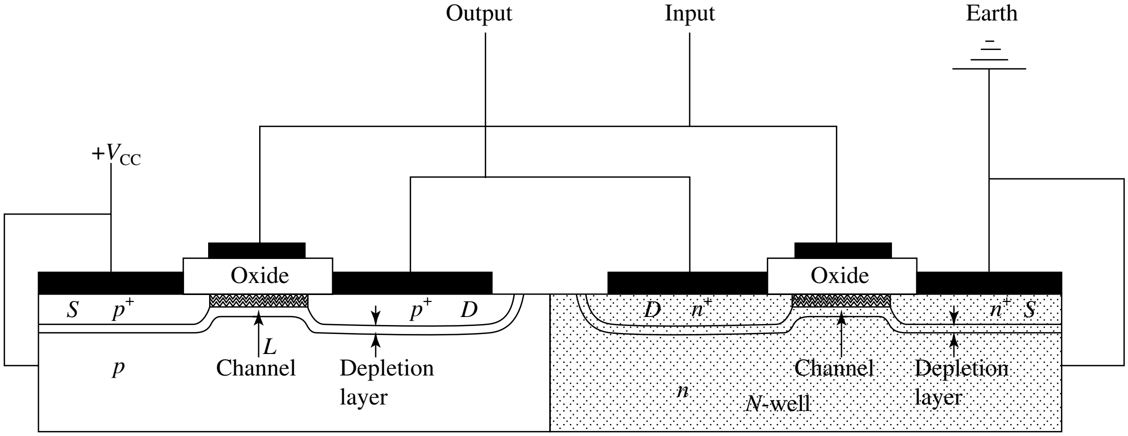


Fig. 1.4 Architecture and the electrical configuration of CMOS

The NMOS transistor is made on lightly doped *P*-type silicon wafer. Two heavily doped phosphorous (*N*-type) regions are made in the silicon. These heavily doped regions are called the source and the drain of a MOS transistor. In between the source and the drain, the gate of the MOS transistor is made. The gate of a MOS transistor is made of a conducting film over the oxide film. The conducting material can either be metal or heavily doped polysilicon. The first type of MOS transistor is called **Metal gate MOS transistor**, and the second type of MOS transistor is called **Polysilicon gate MOS transistor**. The potential applied on the gate is called **gate potential** or **gate voltage**. When a positive gate potential is applied to the electric field below the gate oxide, the electrons are attracted from the silicon wafer and a sheet of electrons is formed in the silicon under the gate. This condition is called **inversion** (because the silicon inverts from *P*-type to *N*-type). This sheet of negatively charged electrons (*N*-type) is around 50 Å in depth and is called ***N*-channel of MOS transistor**; hence this type of transistor is called **NMOS**. When the source, drain and gate are properly biased, a current flows from the source to the drain through the channel. A minimum voltage is needed at the gate to allow the flow of current from the source to the drain is called the **threshold voltage** of the MOS transistor, denoted by V_T. The channel formation between the source and the drain under the gate is called **channel length** denoted by L (or l), and the channel along the source and the drain is called **channel width** of the transistor, denoted by W (or w). The structure and electrical characteristic of NMOS are shown in Fig. 1.2. It is essential to mention that current does not flow through the insulated gate of the MOS transistor; thus, the input impedance is around 10^4 to 10^9 MΩ (in reality it should be infinite). Similarly, a PMOS transistor is made on the *N*-type silicon wafer and the source and the drain are made by boron doping. The PMOS structure, transistor biasing and electrical characteristic are shown in Fig. 1.3. The detailed fabrication process steps of NMOS, PMOS and CMOS transistors are described in Chapter 3.

Enhancement and Depletion Modes of MOS Transistor

There are two modes of a MOS transistor used in ICs, namely, **enhancement mode** and **depletion mode**. The one which is described above is called an **enhancement-mode transistor**, where the gate voltage is applied to create the MOS channel. If the MOS channel already exists without applying the gate voltage, it is called a **depletion-mode transistor**. Usually, positive charges generate in the gate oxide during MOS transistor fabrication, and they are sufficient to form a channel under the gate region in an NMOS transistor without applying gate voltage (unlike a PMOS transistor). To eliminate the existing channel, a negative gate voltage is applied. This type of transistor is called a **depletion transistor**. To make an NMOS enhancement-mode transistor, shallow boron doping (diffusing) into the silicon wafer is done to counter-balance the oxide charges. This type of doping is called **threshold voltage adjustment**. The enhancement and depletion modes of transistors are shown in Figs 1.5 and 1.6 respectively and the details of the threshold voltage adjustment are described in Chapters 3 and 9.

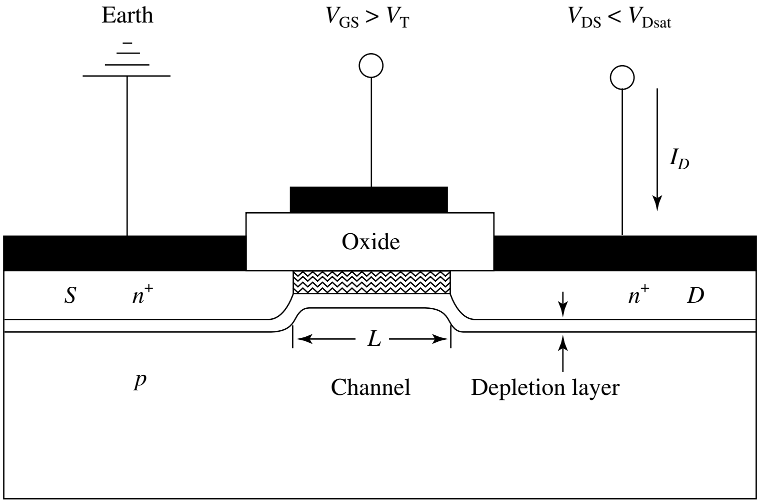


Fig. 1.5 Enhancement and depletion modes of a transistor

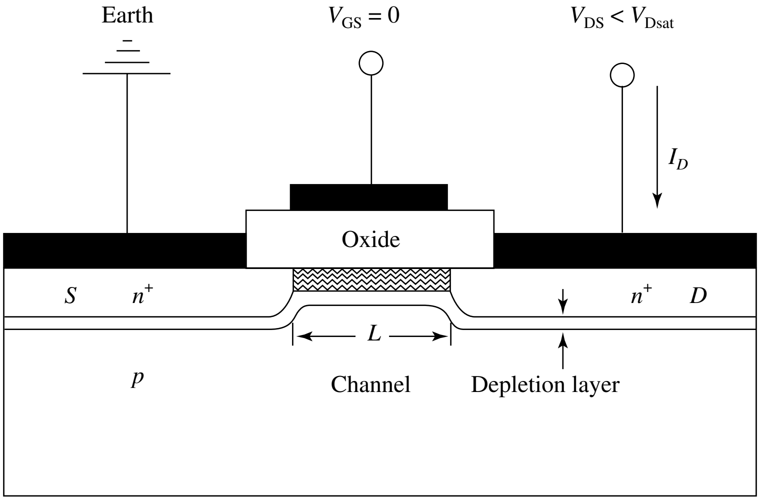


Fig. 1.6 Depletion modes of a transistor

1.3 FUTURE SIZES OF A TRANSISTOR AND A CHIP

To understand the future projections of the National Technology Roadmap for Semiconductors (NTRS), it is essential to mention a few important issues. The smallest dimension (minimum feature size) in an IC is called the **Critical Dimension (CD)** and generally, the smallest dimension (CD) in the MOS structure is the gate length (L). In NTRS, the CD of a transistor was around 250 nm in 1997, but the projected CD in 2012 is 50 nm!

Apart from the issue of reduction in CD, a large number of MOS transistors are required in logic circuits to store information (state) in terms of binary digits (bits). One way of increasing the bits in a chip is to reduce the smallest possible transistor size. Apart from reducing transistor size, functional blocks have to increase. Hence, to accommodate these functional blocks in a single chip, a bigger chip is required. All these requirements demand highly sophisticated equipment and complicated fabrication processes. For many applications, the chip size projected by the NTRS for the year 2012 is around 5.6 times, and for microprocessors, it is around 100 times the size that prevailed in 1977! Furthermore, the scaling down of transistor dimensions leads to many new material and electrical issues. In addition, as per the NTRS, the operating voltage of the device is expected to go down to 0.5 volts from 2.5 volts. If a MOS transistor is scaled by a factor of α^{-1} then the transistor's speed goes up by a factor of α, the density of the transistor increases by a factor of α^2, and the power density reduces by a factor of α^{-2}, all of which have major advantages. Unfortunately, this scaling down of the MOS looks simple, but in reality, it is extremely difficult. A significant variation in the threshold voltage and the subthreshold current takes place with scaling down of the transistor. Subthreshold current is always present in the device, even when the transistor is in the off state. This is due to the weak inversion region which always exists under the gate. There are significant implications on memory loss due to this leakage current. Furthermore, power dissipation, carrier velocity saturation and hot-electron injection are the other issues associated with the scaling down of transistor dimensions. In addition, increase in parasitic capacitance, contact resistance and device interconnection resistance have adverse effects on the transistor speed. Apart from these electrical issues, for the fabrication of a small-geometry MOS transistor, a large number of technical advancements and understanding of physics are imperative. Some of these issues are discussed in detail in Chapter 11.

1.4 PHYSICS OF SILICON

Silicon has three main forms in nature, single crystalline (Si), polycrystalline (poly-Si), and amorphous silicon dioxide (SiO_2). In a MOS transistor, all three forms are extensively used. All these forms are stable; this is one of the reasons why silicon has edged out most other semiconductor materials. Both silicon and poly-silicon change their electrical properties significantly when impurities are present. These impurity elements may be intentionally introduced (doped), or unwanted elements may be present in the silicon. Henceforth, the intentionally doped element will be called **dopant** and the unwanted elements present or diffused in the silicon will be called **impurities**.

It is essential to understand the electrical properties of both pure and doped silicon; especially near the wafer surface. Generally, simple visualisation and better understanding, the energy-band theory is used. All materials are made of atoms, and every atom, barring a few, is made of protons, neutrons and electrons. The protons and neutrons constitute the central nucleus, and the electrons revolve in orbits around the nucleus in several distinct shells. Depending on the nature of the inter-atomic forces, materials are usually found in one out of three forms: solid, liquid or gas. The solid state is further divided into three types: crystalline, polycrystalline and amorphous. If the atoms are aligned in a repetitive array in a particular form then the solid material is called **crystalline**. If the atoms are arranged in a completely random fashion then the solid material is called **amorphous**, and if the material is in between these two extreme forms, then it is called **poly-crystalline**. Poly-silicon consists of a large number of small crystalline grains at different orientations. The crystalline materials are further divided into subgroups depending on the nature of the array in which the atoms are arranged within the solid. One of these is a tetrahedron diamond structure, and silicon falls into this subgroup. Crystalline silicon atoms are lined up in an equidistant manner. The line-up of the atoms is called a **lattice**, an atom occupying the lattice is called a **site** and the specific distance between the silicon atoms is referred to as the **lattice constant**. The **lattice constant** is formed due to the forces of attraction and repulsion between the silicon atoms.

From the quantum mechanics theory, the pure silicon material, called **intrinsic silicon**, is stable if each atom of silicon possesses eight electrons in the outermost shell; but in reality, the silicon atom possesses only four electrons in its outermost shell. Hence, silicon atoms share four of their outermost electrons among themselves in a particular fashion, as if each atom is composed of eight electrons. The sharing (coupling) of electrons between neighbouring atoms is called a **covalent bond**, as shown in Fig. 1.7. Pauli's principle states

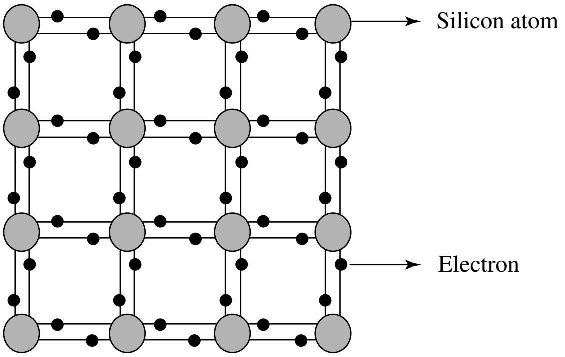


Fig. 1.7 Intrinsic silicon covalent-bond model at absolute temperature

that no two electrons can have the same energy state; thus, the shared electrons possess slightly different energies. As there are a very large number of atoms in a crystal, these energy states are very close to each other, and that in turn forms two distinct allowed energy-level bands, where these electrons can stay, as shown in Fig. 1.8. Each band has continuous allowed energy levels. This is the origin of the band structure in a crystal. Separation of bands is a typical characteristic of a material. The upper energy band is called the **conduction band**, and the lower energy band is called the **valence band**. The separation between the bands is called the "band gap", and for silicon, it is 1.1 eV. In the band gap, no electrons can stay in pure silicon at zero Kelvin; therefore, this band gap is also called the **forbidden gap**, as shown in Fig. 1.9. The energy-band diagram of the well-known diamond structure for crystalline carbon is shown in Fig. 1.8.

At higher temperatures, the electrons acquire enough energy to jump from the valence band into the conduction band, and they move in the conduction band as mobile charge carriers. When an electron jumps into the conduction band, it leaves behind a positive charge called a **hole** in the valence band. The generation of these electrons and holes in the material is called **carrier generation**. The generated hole is occupied by the nearby electron and a new hole is generated. This newly created hole is again occupied by another nearby electron and a new hole is created. The process of generation of holes is random in the valence band, but they move towards negative potential when it is electrically biased. The electrons in the conduction band and the holes in the valence band constitute the total current in the solid. The opposite process of carrier generation is when an electron from the conduction band comes down to the valence band and recombines with a hole. This process is known as **carrier recombination**. At any particular temperature, when the rates of carrier generation become equal to those of carrier recombination then the equilibrium state is reached. Thus, at equilibrium state, and at a particular temperature T, the number

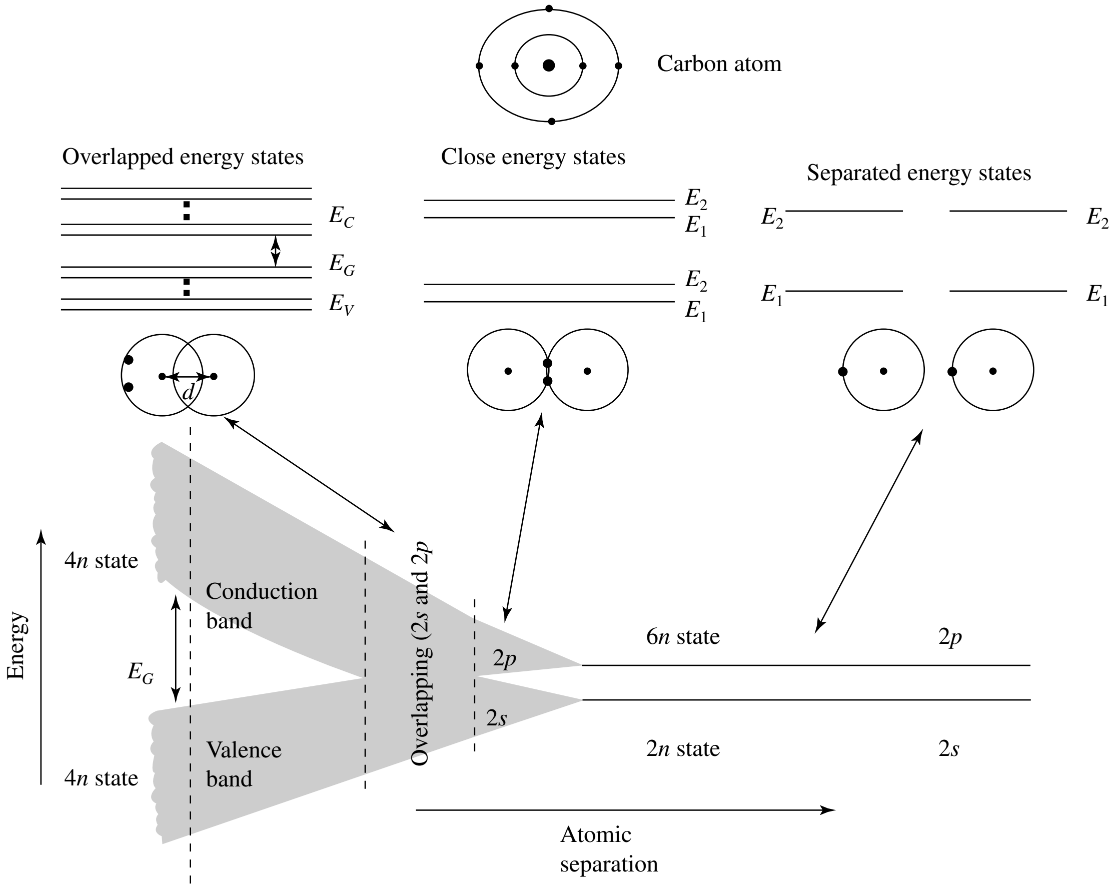


Fig. 1.8 Energy-band structure of crystalline carbon

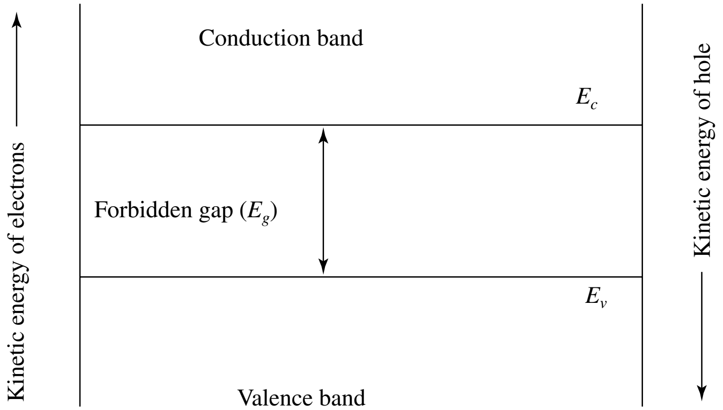


Fig. 1.9 Conduction band, valence band and forbidden gap

of free electrons n per cm^3 will be equal to the number of holes p per cm^3. The number of both electrons and holes can be calculated by the Boltzmann formula.

$$n = p = n_i = 3.9 \times 10^{16} T^{3/2} \exp\left(\frac{-E_G}{2kT}\right) \text{cm}^{-3} \tag{1}$$

Here,

T = temperature, K

n_i = intrinsic carrier concentration, cm^{-3}

E_G = band gap, eV, (1.21 eV for silicon at 0 K)

k = Boltzmann's constant

The equality of electrons and holes implies

$np = (n_i)^2$

The intrinsic carrier concentration based on temperature for some semiconductors is shown in Fig. 1.10.

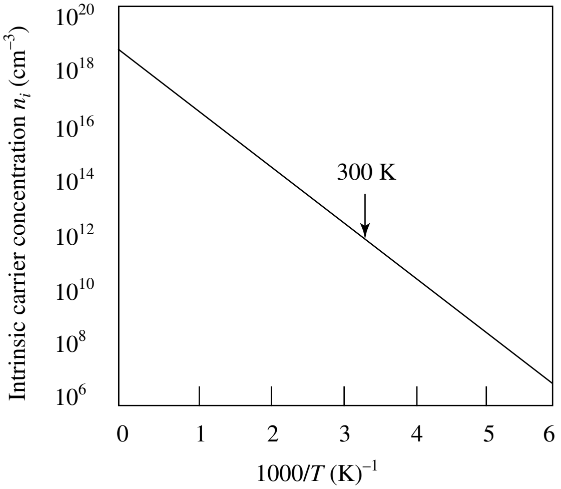


Fig. 1.10 Intrinsic carrier concentration as a function of reciprocal of temperature

Example 1.1

Calculate the values of the carrier concentration at T = 0 K, T = 300 K, T = 650 K, T = 950 K, T = 1000 K, T = 1100 K.

Answer To calculate the carrier concentration, the equation given below is used:

$$n = p = n_i = 3.9 \times 10^{16} T^{3/2} \exp\left(\frac{-E_G}{2kT}\right) \text{cm}^{-3}$$

where,

E_G = band gap, eV, (1.21 eV for silicon at 0 K);
k = Boltzmann's constant (8.617×10^{-5} eV/k).

(i) $T = 0$ K $\quad n_i = 0$ cm^{-3}
(ii) $T = 300$ K $\quad n_i = 1.41 \times 10^{10}$ cm^{-3}
(iii) $T = 650$ K $\quad n_i = 1.32 \times 10^{16}$ cm^{-3}
(iv) $T = 950$ K $\quad n_i = 7.05 \times 10^{17}$ cm^{-3}
(v) $T = 1000$ K $\quad n_i = 1.10 \times 10^{18}$ cm^{-3}
(vi) $T = 1100$ K $\quad n_i = 2.41 \times 10^{18}$ cm^{-3}

A MOS transistor is made on lightly doped with boron or phosphorous on silicon wafer. The lightly doped silicon wafer is made during wafer fabrication as described in Chapter 2. The doped silicon material is called **extrinsic silicon**. The doped boron or phosphorous atoms replace silicon atoms from their lattice site; thus, conductivity of the silicon wafer changes. For instance, the four out of five electrons of the phosphorous atom are bonded with the silicon atom and the remaining one electron is loosely attached with the phosphorous atom. Similarly, when a boron atom replaces a silicon atom from its lattice site, the three electrons of the boron atom make bonds with the three silicon electrons. The deficit of one electron creates a loosely attached hole with the boron atom. At temperatures higher than absolute zero, these electrons and holes are detached from their parent atoms and the released electrons move to the conduction band and the holes move to the valence band. As the phosphorus atom donates an electron to the bulk silicon, it is called the **donor dopant** or the **donor**. It is denoted by N_D^+ and also called the *N*-type silicon; whereas, the boron atom has a deficit of an electron with respect to the bulk silicon, so it is called the **acceptor dopant** or the **acceptor**. It is denoted by N_A^- and also called the ***P*-type** silicon.

The presence of the dopant atoms in silicon creates the free carrier concentration in the silicon. These free carrier concentration govern the MOS electrical characteristics; thus, it is essential to tune the carrier concentration (dopant quantity) in the silicon during MOS fabrication to meet its characteristics. This carrier concentration is directly related with the resistivity (or conductivity) of the silicon wafer. For convenience, carrier concentration is measured by the resistivity method and thereafter, conductivity is calculated from the resistivity data. Furthermore, conductivity is related to the mobility of the carrier, as mentioned below.

$$\sigma_n = nq\mu_n + pq\mu_p, \quad (\Omega \text{ cm})^{-1} \tag{2}$$

or

$$\rho = \frac{1}{nq\mu_n + pq\mu_p}, \qquad \Omega \text{ cm} \tag{3}$$

where, σ is the bulk conductivity of the material, ρ is the bulk resistivity of the material, n is the electron concentration in cm^{-3}, p is the hole concentration in cm^{-3}, q is the electron charge, and μ is the electron mobility in cm^2/V s.

In general, silicon is doped with only one type of dopant; hence, the equations for *N*-type and *P*-type silicon can be written as

***N*-type silicon**	***P*-type silicon**
$\sigma_n = nq\mu_n$, $(\Omega\ \mathrm{cm})^{-1}$	$\sigma_p = nq\mu_p$, $(\Omega\ \mathrm{cm})^{-1}$

or,

$$\rho_n = \frac{1}{nq\mu_n} \quad \Omega\,\mathrm{cm} \qquad\qquad \rho_p = \frac{1}{nq\mu_p} \quad \Omega\,\mathrm{cm}$$

It is found that the maximum mobility of electrons and holes for low doping and small electric field are around 1500 cm^2/V s and 500 cm^2/V s respectively.

Furthermore, the mobility (velocity) described above is directly proportional to the applied electric field.

$$v = \mu E$$

where v is the carrier (electron or hole) velocity and E is the applied electric field.

Generally, the doped thin film is different than the bulk material and it is defined as the sheet material. Furthermore, the dopant atoms are not uniformly doped with the diffused depth (deep inside the wafer) and the sheet resistivity also changes with temperature. The process of measurement of sheet resistivity is described in detail in Chapter 8.

It is mentioned earlier that electrons and holes are generated in the silicon after doping with phosphorous and boron atoms. Apart from the dopant-induced free electrons (holes), the covalent bonds of the silicon atoms also break and generate electrons and holes. Electrons move to the conduction band at higher (above zero degrees Kelvin) temperature. In a similar fashion, holes move into the valence band.

When a dopant is introduced at higher temperature, energy levels are created in the forbidden gap. It is found that a dopant like phosphorous (*N*-type) introduces allowed energy levels close to the conduction band in the forbidden gap and electrons occupy those allowed energy levels. Since these electrons are much closer to the conduction band, they need very less energy to reach the conduction state. At high doping concentrations, the allowed levels in the forbidden gap further shift towards the conduction band.

Similarly, the holes occupy the introduced energy levels closer to the valence band in the forbidden gap when the silicon is doped with a *P*-type dopant like boron. As these holes are very close to the valence band, they need little energy to jump into the valence band. At room temperature, in the *P*-type silicon, almost all loosely bonded holes are in the valence band. Similarly, at room temperature, in the *N*-type silicon, almost all loosely bonded electrons are in the conduction band. The number of electrons (holes) is much more than the thermally generated electrons (holes) in the silicon wafer. For example, doping of the order of 10^{14} cm^{-3} (considered to be light doping) gives carriers of the order of four times the magnitude of the intrinsic carrier concentration. Under these circumstances, equations for electron and hole carrier concentration at room temperature are expressed in terms of dopant concentration as

$$n = N_{D+} \tag{4}$$

and

$$p = \frac{n_i^2}{n} \tag{5}$$

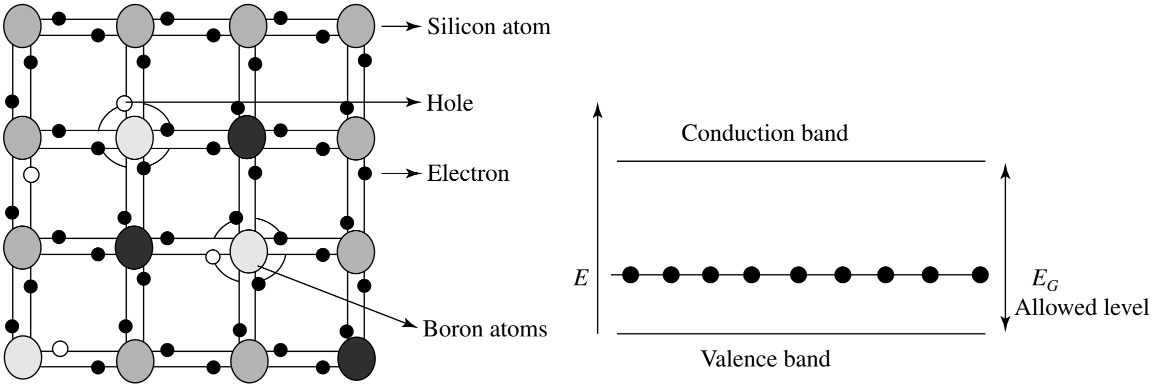
The actual number of electrons occupying a particular energy level can be calculated at any temperature by applying the Fermi–Dirac distribution function, which is given by

$$F(E) = \frac{1}{1 + \exp\left(\frac{E - E_F}{kT}\right)} \tag{6}$$

Here, $F(E)$ is the probability of occupation of a particular level E at temperature T of the semiconductor expressed in K, E_F is the Fermi energy, k is the Boltzmann's constant which is equal to 8.62×10^{-5} eV/K (1.38×10^{-23} J/K) and kT is typically of the order of 0.026 eV at 300 K.

The Fermi energy is the value of energy at which the probability $F(E_F)$ is equal to 0.5 at absolute zero temperature. For an intrinsic semiconductor, this is about midway between the valence and conduction bands at absolute zero. At any finite temperature, the energy level at which this probability falls to half is called **Fermi level**. The Fermi energy and the Fermi level are the same at absolute zero, and very often the same symbol E_F is used for the Fermi level too.

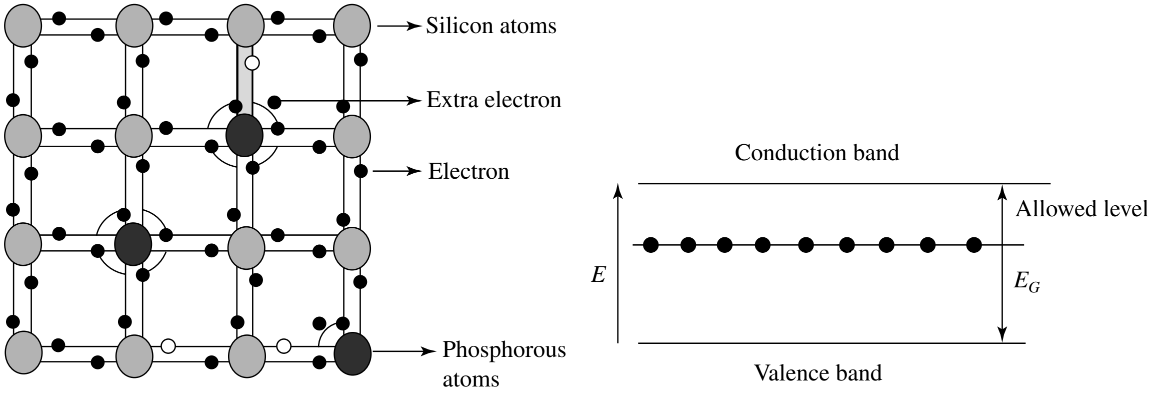
The Fermi level of the donor (phosphorous) dopant level is very close to the conduction band. Similarly, the Fermi level of an acceptor (boron) dopant is close to the valence band at any finite temperature. Thus, it is the exponential term in the distribution which is denoted by $F(E_C) = \exp\left(\frac{E_C - E_F}{kT}\right)$. The Fermi level of a *P*-doped (*P*-type) silicon and an *N*-doped (*N*-type) silicon is shown in Figs 1.11 and Fig. 1.12 respectively.



(a) Boron-doped silicon

(b) Boron-doped silicon energy-band diagram

Fig. 1.11 Fermi level for intrinsic, *n*-doped silicon



(a) Phosphorous-doped silicon

(b) Phosphorous-doped silicon energy-band diagram

Fig. 1.12 Fermi level for intrinsic, *p*-doped silicon

To find the actual number of electrons in the conduction band and holes in the valence band, we must multiply this by the number of states actually available at energy *E*. The number of states available in the conduction band or in the valence band is called **density of states**. Close to the edge of the conduction band, we treat the electrons as free particles with some effective mass m_e*, hence the density of states in the conduction band must follow the equation:

$$N(E) = \frac{4\pi}{h^3}(2m_e^*)^{\frac{3}{2}}(E - E_C)^{\frac{1}{2}}, \qquad E > E_c \tag{7}$$

Similarly, the density of states in the valence band is

$$N(E) = \frac{4\pi}{h^3}(2m_h^*)^{\frac{3}{2}}(E_V - E)^{\frac{1}{2}}, \qquad E < E_v \tag{8}$$

Here, m_h^* is the effective mass of the holes and h is the Planck's constant. Thus, the carrier concentration in the conduction band is

$$n = \int_{E_C}^{\infty} N(E)F(E)dE = N_C \exp\left(-\frac{E_C - E_F}{kT}\right) \tag{9}$$

and the carrier concentration in the valence band is

$$p = \int_{-\infty}^{E_V} N(E)(1 - F(E))dE = N_V \exp\left(-\frac{E_F - E_V}{kT}\right) \tag{10}$$

where

$$N_C = 2\left[\frac{2\pi m_e^* kT}{h^2}\right]^{\frac{3}{2}}, \quad \text{and} \quad N_V = 2\left[\frac{2\pi m_h^* kT}{h^2}\right]^{\frac{3}{2}} \tag{11}$$

where N_c and N_v are called the **effective density of states** for the conduction band and the valence band respectively.

Typically, in silicon at room temperature, $N_C = 2.8 \times 10^{19}$ cm^{-3}, $N_V = 1.04 \times 10^{19}$ cm^{-3}. An interesting consequence of this is that the Fermi level does not occur midway between the valence and conduction bands at any finite temperature as the effective masses are different. If we denote the middle of the forbidden gap by E_i then we can write the carrier concentration in terms of the effective density of states N_c as follows:

$$n_i = N_C \exp\left[-\frac{E_C - E_F}{kT}\right] \tag{12}$$

and

$$p_i = N_V \exp\left[-\frac{E_F - E_V}{kT}\right] \tag{13}$$

On substituting these two formulae for the two constants, the actual electron and hole concentrations can be expressed as

$$n = n_i \exp\left[\frac{E_F - E_i}{kT}\right], \tag{14}$$

and

$$p = p_i \exp\left[\frac{E_i - E_F}{kT}\right], \tag{15}$$

where n and p are the densities of the electron and the hole respectively.

These formulae relate the concentration of electrons and holes to the position of the Fermi level relative to the middle of the forbidden gap. If the Fermi level rises above the middle of the gap then the number of electrons in the conduction band increases exponentially. Similarly, if the Fermi level falls below the middle of the gap then the number of holes in the valence band increases exponentially. This is shown in Fig. 1.12.

Intrinsic carrier concentration is calculated using the effective density of states function of silicon from Table 1 given below.

Table 1.1 Values—effective density of states function and effective mass of silicon (at T = 300 K)

N_c	N_v	m_n^*/m_0	m_p^*/m_0
2.8×10^{19}	1.04×10^{19}	1.08	0.56

It is important to note that the intrinsic carrier concentration at 300 K is experimentally reported as 1.5×10^{10} cm^{-3}. Whereas in the above calculation, the intrinsic carrier concentration at 300 K comes out to be 7.023×10^{10} cm^{-3}. This discrepancy of a factor of 2 is due to the facts that the effective mass changes with temperature, and the theoretical equations of the density of states are not accurate.

Example 1.2

Find out the effective density of states in the valence band and in the conduction band at 300 K temperature.

$$N_C = 2\left[\frac{2\pi m_e^* kT}{h^2}\right]^{\frac{3}{2}} \quad \text{and} \quad N_V = 2\left[\frac{2\pi m_h^* kT}{h^2}\right]^{\frac{3}{2}}$$

where m_e and m_h* are the effective masses of electrons and holes respectively and h is the Planck's constant.*

Answer

$K = 8.62 \times 10^{-5}$ eV/K

$h = 4.135 \times 10^{-15}$ eV/s

$m_e^* = 1.08\, m_0$ kg

$m_h^* = 0.56\, m_0$ kg

$m_0 = 9.11 \times 10^{-31}$ kg

The effective density of states in the valence band:

$$N_C = 2\left[\frac{2\pi m_e^* kT}{h^2}\right]^{\frac{3}{2}}$$

$$N_C = 2\left[\frac{2 \times 3.14 \times 1.08 \times 9.11 \times 10^{-23} \times 300}{(6.625 \times 10^{-34})^2}\right]^{\frac{3}{2}}$$

$$N_C = 2.816 \times 10^{25}\ \mathrm{m}^{-3}$$

The effective density of states in the conduction band:

$$N_V = 2\left[\frac{2\pi m_h^* kT}{h^2}\right]^{\frac{3}{2}}$$

$$N_V = 2\left[\frac{2 \times 3.14 \times 0.56 \times 9.11 \times 10^{-31} \times 1.38 \times 10^{-23} \times 300}{(6.625 \times 10^{-34})^2}\right]^{\frac{3}{2}}$$

$$N_V = 1.051 \times 10^{25}\ \mathrm{m}^{-3}$$

1.5 SILICON DEVICES

The electrical parameters of a MOS transistor are governed by doping parameters such as doping concentration, junction depth, etc. These electrical parameters can be understood by the energy-band diagram. For convenience and simplicity, the energy-band diagram of one junction diode device is taken as the vehicle.

1.5.1 Diode

As an ideal diode, let us take a *P*-type doped region sharply separated from the *N*-type silicon; though this is not possible in the fabrication process as the dopant always diffuses from a high-dopant-concentration region to a low-dopant-concentration region that makes the junction, graded. Therefore, in reality, the junction of a diode is always graded. Let us take the ideal case where the junction of a diode is abrupt and no gradation of dopant exists on either side of the junction.

As discussed before, in the *N*-type silicon, electrons are the majority carriers, and they are mostly in the conduction band in a nearly free state; while the holes are in a minority. The opposite is true in the case of *P*-type silicon, where holes are the majority carriers, and electrons are in a minority. When these two types of silicon are bonded, the high concentration of electrons from the *N*-type silicon diffuses into the *P*-type silicon. Similarly, the high concentration of holes from the *P*-type silicon migrates to the *N*-type silicon. This phenomenon is called **carrier diffusion current** or simply **diffusion current**. When the electrons migrate from the *N*-type to the *P*-type, they leave behind the positively charged immobile phosphorous ionized atoms in the *N*-type silicon near the junction which leads to a build-up of positive charge. The electric field produced by this built-up charge opposes further diffusion of holes (positive charged) into the *N*-type silicon. Similarly, the electric field produced by the built-up charge will oppose further diffusion of electrons (positive charged) into the *P*-type silicon. Initially, the migration of carriers continues for some time, and thereafter, no diffusion of current takes place; and finally a state of equilibrium is reached. The region that contains the immobile charge at both sides of the junction does not contain any types of charge carriers; hence this region is called **depletion charges**, as shown in Fig. 1.13 and Fig. 1.14.

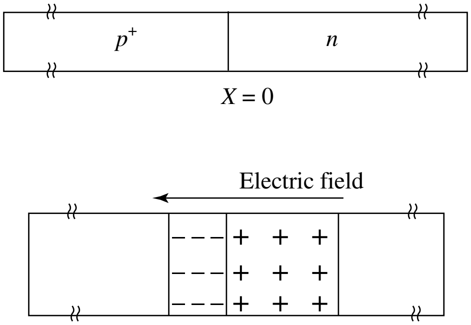


Fig. 1.13 Diode and depletion width

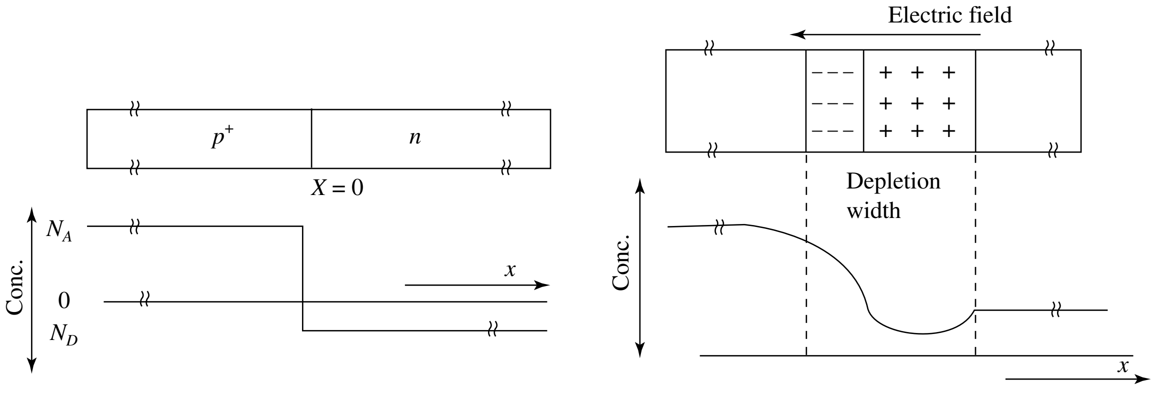


Fig. 1.14 Diffusion current is opposite the built-in potential at the equilibrium state

It was discussed in the energy-band theory that in the *N*-type silicon, Fermi level is close to the conduction band and in the *P*-type silicon, it is close to the valence band. When these two types of silicon are bonded together, it forms an abrupt junction; and the Fermi levels of the depletion region of *N*-type and the depletion region of *P*-type silicon are aligned into a straight line, as shown in Fig. 1.15. This results in the build-up of a voltage and can be expressed as,

$$V = \phi_{F_p} + \phi_{F_n} \tag{16}$$

where ϕ_{F_p} and ϕ_{F_n} denote the Fermi-level potential of the depletion layers at both sides of the junction.

As $q\phi_{F_n}$ is the energy of the electrons, we have,

$$\phi_{F_n} = \frac{E_F - E_{in}}{q} = \frac{kT}{q}\ln\left(\frac{N_D}{n_i}\right) \tag{17}$$

In a similar fashion,

$$\phi_{F_p} = \frac{E_F - E_{ip}}{q} = \frac{kT}{q}\ln\left(\frac{N_A}{n_i}\right) \tag{18}$$

Thus, the electrical potential across the junction can be calculated from the material and process parameters.

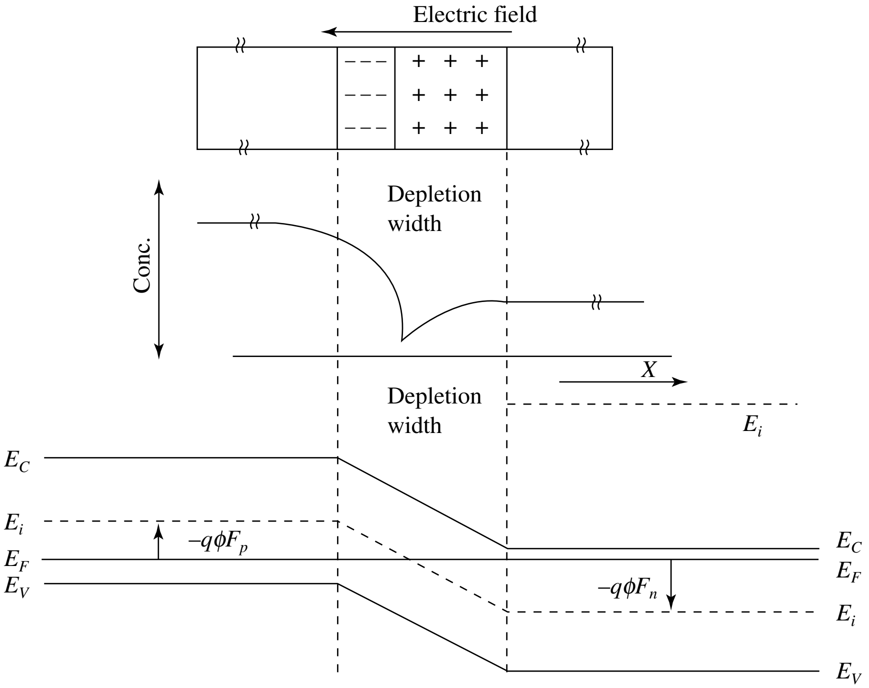


Fig. 1.15 Fermi levels at abrupt junctions of *N*-type and *P*-type silicon

Example 1.3

Find the carrier concentration and Fermi level, when 10^{15} atoms/cm^3 of phosphorus is doped into an intrinsic silicon wafer and the same quantity of boron is doped into another silicon wafer at a temperature of 300° K. Take the N_c value from Example 1.2.

Answer

At 300 K almost all dopant atoms are ionised.

Hence, N_D and N_A = 10^{15} atoms/cm^3

From Eq. (5), one can write

Hole carrier concentration, $p \approx \dfrac{n_i^2}{N_D} = \dfrac{(1.45\times10^{10})^2}{10^{15}} = 2.1\times10^5\ \text{cm}^{-3}$

Electron carrier concentration, $n \approx \dfrac{n_i^2}{N_A} = \dfrac{(1.45\times10^{10})^2}{10^{15}} = 2.1\times10^5\ \text{cm}^{-3}$

$$\begin{aligned} E_C - E_F &= kT\ \ln\left[\frac{N_C}{N_D}\right] \\ &= 0.0259\ \ln\left[\frac{2.816\times10^{31}}{10^{15}}\right] \\ &= 0.981\ \text{e V} \end{aligned}$$

$$\begin{aligned} E_F - E_V &= kT\ \ln\left[\frac{N_C}{N_D}\right] \\ &= 0.0259\ \ln\left[\frac{1.015\times10^{31}}{10^{15}}\right] \\ &= 0.955\ \text{eV} \end{aligned}$$

Built-in Voltage, Electric Field and Depletion Width of a Junction

A diode is made of one type of dopant highly doped into another type of lightly doped silicon substrate (wafer). The relation between the electric field, the potential and the depletion width of the two doped regions of the junction can be represented by the Poisson equation.

As the junction is assumed to be a plane boundary, one-dimensional Poisson equation is needed, as the field can vary only in the direction perpendicular to the junction. So the Poisson equation can be written as

$$\frac{dE(x)}{dx} = \frac{\rho(x)}{4\pi\varepsilon_s\varepsilon_0} \tag{19}$$

Here, ρ is the charge density (not to be confused with the resistivity), ε_0 is the permittivity of free space (8.86×10^{-14} F/cm) and ε_s is the dielectric constant (11.9) of silicon.

As an illustration, consider a junction formed between a heavily doped *P*-type and a lightly doped *N*-type silicon. The depletion width *dx* is much greater in the *N*-type silicon as the number of free carriers per unit volume is less in this region. So, integrating the Poisson equation, the maximum electric field at the junction can be written as

$$E_{\max} = \int_{-x_D}^{0} \frac{qN_D}{\varepsilon_0 \varepsilon_s} dx = \frac{qN_D}{\varepsilon_0 \varepsilon_s} x_d \tag{20}$$

Here, N_D is the concentration of the *N*-type dopant and is taken as a constant. Integrating the electric field over the depletion width, the **built in potential** $|V_B|$ can be expressed as

$$V_B = -\frac{qN_D}{2\varepsilon_0 \varepsilon_s}(x_d)^2 \tag{21}$$

Rewriting this, we can find the depletion width in terms of the silicon material parameters as

$$x_d = \sqrt{\frac{2\varepsilon_0 \varepsilon_s}{qN_D} |V_B|} \tag{22}$$

In the presence of an external potential V_A, the depletion width can be written as

$$x_d = \sqrt{\frac{2\varepsilon_0 \varepsilon_s}{qN_D} |V_B + V_A|} \tag{23}$$

Example 1.4

Find the built-in potential and depletion width of an abrupt junction, doped with boron atoms 10^{16}, 10^{18} and 10^{20} cm^{-3} in uniformly doped 10^{15} cm^{-3} phosphorous atoms in a silicon wafer at 300 K temperature.

Answer

$$V_b = \frac{KT}{q} \text{in} \frac{N_A N_D}{n_i^2}$$

$$V_b = \frac{8.616 \times 10^{-5} \times 300}{1.6 \times 10^{-19}} \ln \frac{10^{16} \times 10^{15}}{(1.5 \times 10^{10})^2} = 0.025875 \times 24.5175 = 0.63439 \text{ eV}$$

$$V_b = \frac{8.616 \times 10^{-5} \times 300}{1.6 \times 10^{-19}} \ln \frac{10^{18} \times 10^{15}}{(1.5 \times 10^{10})^2} = 0.025875 \times 29.12267 = 0.75354 \text{ eV}$$

$$V_b = \frac{8.616 \times 10^{-5} \times 300}{1.6 \times 10^{-19}} \ln \frac{10^{20} \times 10^{15}}{(1.5 \times 10^{10})^2} = 0.025875 \times 33.7278 = 0.87270 \text{ eV}$$

Depletion width for 10^{16} cm^{-3},

$$x_d = \sqrt{\frac{2 \times 11.7 \times 8.85 \times 10^{-14} \times 0.6349}{1.6 \times 10^{-19} \times 10^{15}} |V_B|}$$

$$x_d = 9.06 \times 10^{-5} \text{ cm}$$

Depletion width for 10^{18} cm^{-3}, $x_d = \sqrt{\frac{2\varepsilon_s \varepsilon_0}{qN_D} |V_B|}$

$$x_d = 9.88 \times 10^{-5} \text{ cm}$$

Depletion width for 10^{20} cm^{-3}, $x_d = \sqrt{\frac{2\varepsilon_s \varepsilon_0}{qN_D} |V_B|}$

$$x_d = 10.63 \times 10^{-5} \text{cm}$$

$$x_d = 0.1063 \text{ mm}$$

When a positive voltage is applied to an *N*-type silicon and a negative voltage is applied to a *P*-type silicon, the depletion width increases and that in turn increases the built-in potential; as a consequence, electrons and holes are not able to cross the junction. This electrical biasing condition creates a **reverse-biased junction**. On the other hand, when a negative voltage is applied to an *N*-type silicon and a positive voltage is applied to a *P*-type silicon, the depletion width decreases and that in turn decreases the built-in potential; as a consequence, electrons and holes can easily cross the junction. This electrical biasing condition creates a **forward-biased junction**. Note that the depletion width is inversely proportional to the square root of the applied voltage. The electric field at the junction, in the presence of the applied voltage, can be written as

$$E_{\max} = \frac{V_A + V_B}{x_d} = \sqrt{\frac{qN_D}{2\varepsilon_0 \varepsilon_s} |V_A + V_B|} \tag{24}$$

As there are no charge carriers in the depletion width, it behaves like a capacitor. This capacitance plays an important role in high-frequency devices.

The effective charge density per unit area in the *N*-type silicon at the end of the depletion width is

$$Q = qN_D x_d \tag{25}$$

or,

$$\frac{dQ}{dx_d} = qN_D \tag{26}$$

while the voltage across the junction is

$$|V_A + V_B| = V = \frac{qN_D}{2\varepsilon_0\varepsilon_s}(x_d)^2 \tag{27}$$

$$\frac{dV}{dx} = \frac{qN_D}{\varepsilon_0\varepsilon_s}x_d \tag{28}$$

As charge does not increase linearly with voltage, capacitance can be defined as

$$C = \frac{dQ}{dV} = \frac{dQ/dx_d}{dV/dx_d} = \frac{\varepsilon_0\varepsilon_s}{x_d} = \sqrt{\frac{qN_D\varepsilon_0\varepsilon_s}{2\,|V_A + V_B|}} \tag{29}$$

1.6 MOS TRANSISTOR

The energy-band diagram can be extended to study the electrical properties of a MOS transistor. A MOS transistor is made of three regions; namely, the source, the drain and the gate. The most important part of a MOS transistor is the gate region, where transistor action takes place. When gate voltage is applied, accumulation, depletion and then inversion (conducting channel) can be formed below the gate in the silicon; thus, the state of a MOS transistor can be changed according to the gate voltage. Hence, the energy-band diagram is studied from the gate to the underneath silicon substrate. To maintain consistency with the energy-band diagram, and for better visualisation, the MOS structure is taken on the vertical axis and the origin x is taken as the oxide and silicon interface (oxide-silicon interface), as shown in Fig. 1.16. When there is no voltage at the gate and no oxide charges present in the oxide and at the oxide/silicon interface then the energy levels will be perfectly flat (straight line). This condition is called **flat-band condition**. When voltage is applied to the gate and the silicon substrate is grounded then the energy band bends. The nature and magnitude of the energy-band bending will depend on the polarity and magnitude of voltage applied to the gate with respect to grounded silicon substrate. For example, if the P-type silicon substrate is grounded and negative voltage is applied to the gate then the mobile holes in the silicon wafer get attracted under the gate to balance the gate potential and get accumulated there, as shown in Fig. 1.17. This condition is called the **accumulation stage**. When negative gate voltage is gradually changed to positive voltage, the mobile holes get repelled from under the gate and a depletion layer (intrinsic silicon, i.e. $E_F = E_i$) is formed, as shown in Fig. 1.18. With further increase in positive voltage, the increase in depletion width will begin to cease and electrons will start collecting under the gate, to balance the gate voltage. With an even further increase in the gate voltage, the depletion

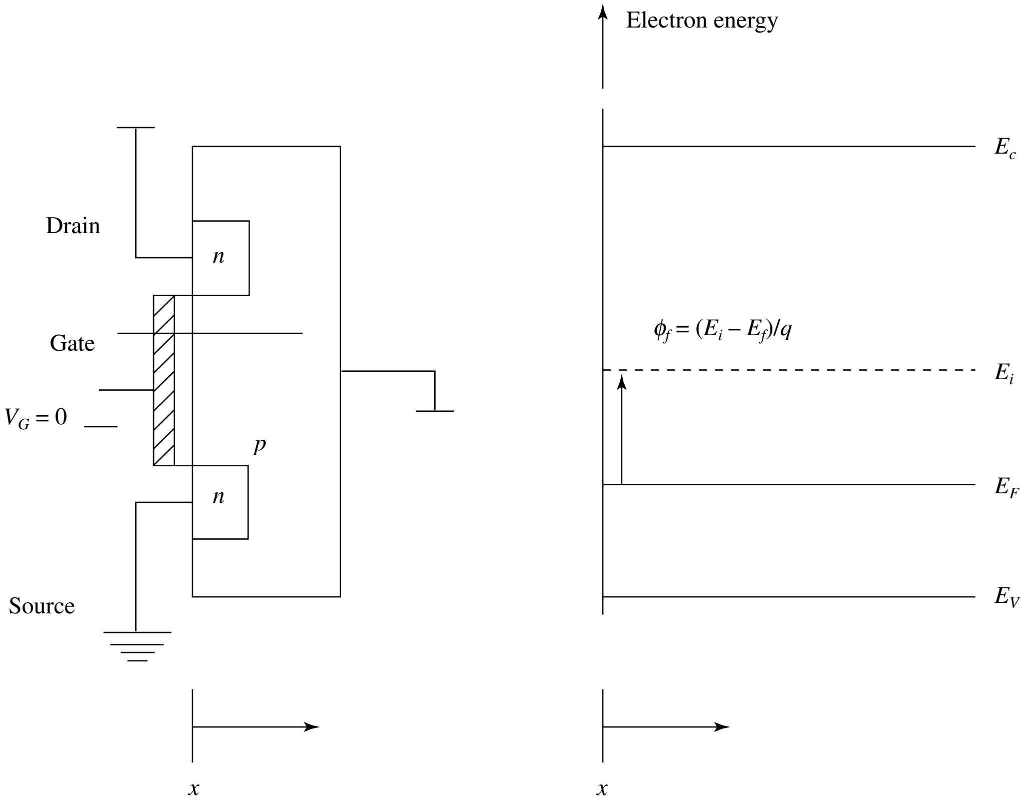


Fig 1.16 Energy-band diagram of MOS transistor

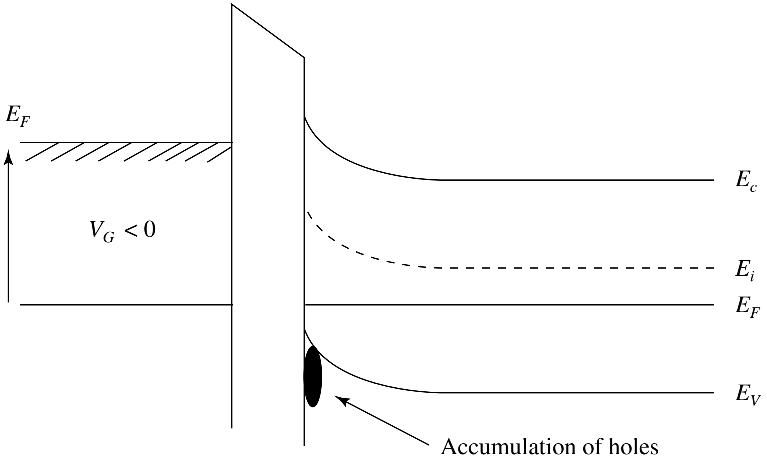


Fig 1.17 Inversion condition and channel formation for PMOS

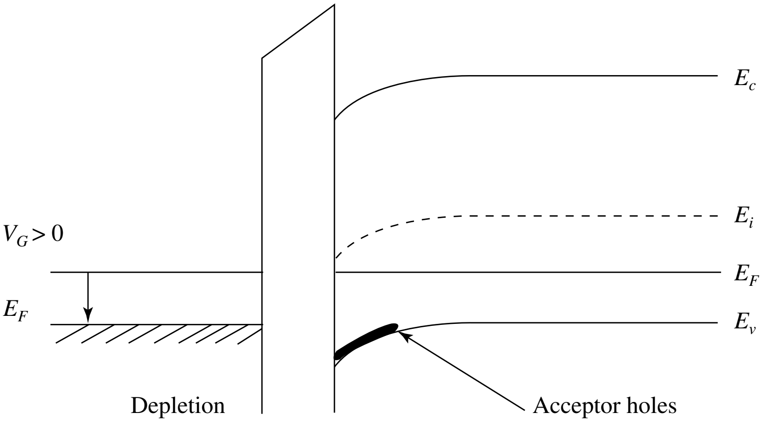


Fig 1.18 Inversion condition and channel formation for PMOS

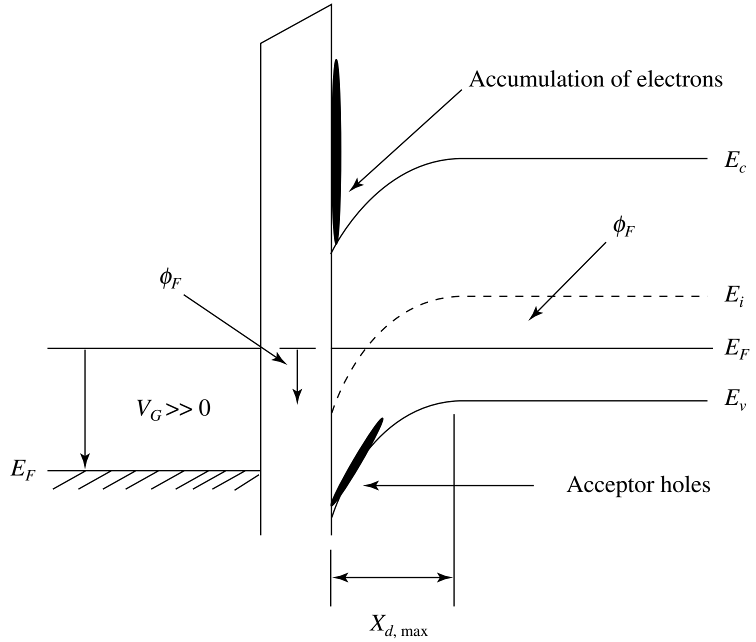


Fig 1.19 Inversion condition and channel formation for PMOS

width will cease increasing (widening) and electrons will start getting collected under the gate resulting in channel formation. In MOS, this condition is called **inversion condition** and the gate voltage that creates the channel is called **threshold voltage** (V_T), as shown in Fig. 1.19. At this inversion stage, the silicon surface potential is equal to the amount of band bending.

For the potential calculations, the intrinsic silicon Fermi level E_i is taken as the reference potential, and the equation can be expressed as

$$\phi_s \frac{-(E_i \text{ at surface} - E_i \text{ at bulk})}{q} = 2\phi_f \tag{30}$$

where ϕ is the Fermi potential and is defined as

$$2\phi_f = \frac{-(E_F - E_i)}{q} \tag{31}$$

Similarly, when positive gate voltage is applied with respect to the *N*-type silicon substrate, there is accumulation of electrons, and when the gate potential is changed gradually from positive to negative, a depletion layer and finally an inversion condition is reached, as shown in Fig. 1.19.

When gate voltage is swept from one polarity to another polarity, the capacitance below the gate also changes. In accumulation and inversion conditions, the capacitance is due to the gate oxide dielectric, but in the depletion condition, the gate-to-silicon capacitance is equal to the gate oxide plus the depletion width capacitance. In between one potential to another, the gate capacitance changes with the gate voltage. The measurement of capacitance with gate voltage is called the **C-V measurement**. The *C-V* measurement reveals a great deal about MOS electrical characteristics and MOS processing qualities and parameters. For instance, one can find the type, quantity and nature of the charges (generally positive charges), during silicon oxidation. These positive charges have great effect on the MOS transistor threshold voltage. Details of *C-V* measurement are described in Chapter 4. To counter the effects of these charges, and bring back a flat band condition, the dopant concentration is introduced under the gate during MOS fabrication process. This fabrication step is called **threshold voltage adjustment**. The detail of the threshold voltage adjustment is described in Chapter 9 under ion implantation.

1.6.1 Threshold Voltage for the MOS Transistor

Case 1: Threshold for the Ideal Case

To derive the threshold voltage for a MOS transistor, for simplicity, let us first take the ideal case where no oxide and oxide-silicon interface charges are present.

Initially, consider the situation where the applied voltage is just about to cause the device to enter the inversion condition (channel formation)

$$\phi_s = 2\phi_F \tag{32}$$

where, the gate voltage can be equated to the drop across the gate oxide plus the silicon surface potential. If E_{OX} is the electric field across the gate oxide, x_d is the thickness of the gate oxide, and ϕ_S is the surface potential of the silicon then

$$V_G = E_{ox}x_d + \phi_s \tag{33}$$

From the boundary condition on the electric field at the oxide-silicon interface,

$$\varepsilon_{Si}E_{Si} = \varepsilon_{Ox}E_{Ox} \tag{34}$$

While from Gauss' law,

$$E_{Si} = \frac{Q_s}{\varepsilon_{Si}\varepsilon_0} \tag{35}$$

Here, Q_s is the surface charge, E_{si} is the electric field in the silicon, ε_{Si} is the relative permittivity of the silicon, and ε_0 is the permittivity of free space; a similar notation is used for the oxide. If the depletion width is x_d then the corresponding voltage is

$$V = -\frac{Q_s t_{Ox}}{\varepsilon_{Si}\varepsilon_0} = -\frac{Q_s}{C_0} \tag{36}$$

where C_0 is the capacitance per unit area of the gate and t_{ox} is the oxide thickness. The gate voltage is thus given by

$$V_G = -\frac{Q_s}{C_0} + \phi_s \tag{37}$$

The depletion Q_S can be expressed as the depletion layer charge Q_B as

$$Q_S = Q_B = -qN_A x_D = -\sqrt{2\varepsilon_{Si}\varepsilon_0 N_A \phi_S} \tag{38}$$

where x_d is the depletion width. We now substitute the value of Q_S in our equation for gate voltage and use the relation $\varepsilon_S = 2\varepsilon_F$ to obtain

$$V_G = \frac{\sqrt{2\varepsilon_{Si}\varepsilon_0 N_A (2\phi_F)}}{C_0} + 2\phi_F \tag{39}$$

When an increased gate voltage causes a strong inversion condition under the gate, an inversion charge Q_{in} also has to be added to the equation, hence it now becomes $Q_S = Q_B + Q_{\text{in}}$. So the threshold condition at which the MOS transistor starts working is

$$V_T = -\frac{Q_B}{C_0} - \frac{Q_{in}}{C_0} + 2\phi_F \tag{40}$$

For an *n*-channel transistor, the threshold voltage is V_{Tn} and the inversion charge is Q_n.

Substituting these in our equation, we obtain

$$V_{Tn} = -\frac{Q_B}{C_0} - \frac{Q_n}{C_0} + 2\phi_F \tag{41}$$

or,

$$V_{Tn} = -\frac{Q_n}{C_0} + V_G \tag{42}$$

or,

$$Q_n = C_0(V_G - V_{Tn}). \tag{43}$$

For the p-channel transistor, the threshold voltage is V_{Tp} and the inversion charge density is due to the holes, so

$$V_{Tp} = -\frac{\sqrt{2\varepsilon_{Si}\varepsilon_0 N_A 2\,|\phi_F|}}{C_0} + 2\phi_F \tag{44}$$

or,

$$V_{Tp} = -\frac{Q_p}{C_0} + V_G \tag{45}$$

or,

$$Q_p = -C_0(V_G - V_{Tp}) \tag{46}$$

Thus, by optimisation of doping concentration, the voltage characteristics can be changed as desired.

Case 2: Threshold for the Non-ideal Case

In the previous section, threshold voltage is derived for the ideal case where no charges are present in the oxide as well as in the oxide-silicon interface; but in reality it never happens and charges always come up during oxidation. Generally, four types of charges are generated during MOS fabrication. The first type of charge is generated on the silicon surface due to broken bonds when oxidation is suddenly stopped. The second type of charge is generated because of the impurities such as sodium and potassium atoms in the oxide. The third type of charge is generated because of radiation exposure during MOS fabrication, and the fourth type of charge is generated due to dangling bonds and a different material used to make the gate of the MOS transistor. The first three types of charges are generated at the MOS fabrication stage and the fourth type of charge is generated due to dangling bonds created due to the electronic property of material called **work function**. The details of these charges are covered in the chapter on oxidation (Chapter 4).

The work function can be calculated using the energy-band concept. For example, one type of MOS is made of aluminium metal as a gate electrode, where the work functions of silicon, silicon dioxide and aluminium are different. To find the effects of these work functions, the Fermi energy of silicon dioxide is taken as the reference energy level, as shown in Fig. 1.20. When the three materials come into intimate contact, the Fermi levels align themselves, which results in band bending. If the work function ϕ_{Si} of silicon is greater than that of the metal ϕ_M then the work function at the gate is given by

$$\phi_{MSi} = \phi_M - \phi_{Si} \tag{47}$$

This band bending (or interface work function) results in a charge at the gate of the MOS transistor. Similarly, a poly-silicon gate MOS transistor has a different work function, and hence different charges at the gate than the metal-gate MOS transistor.

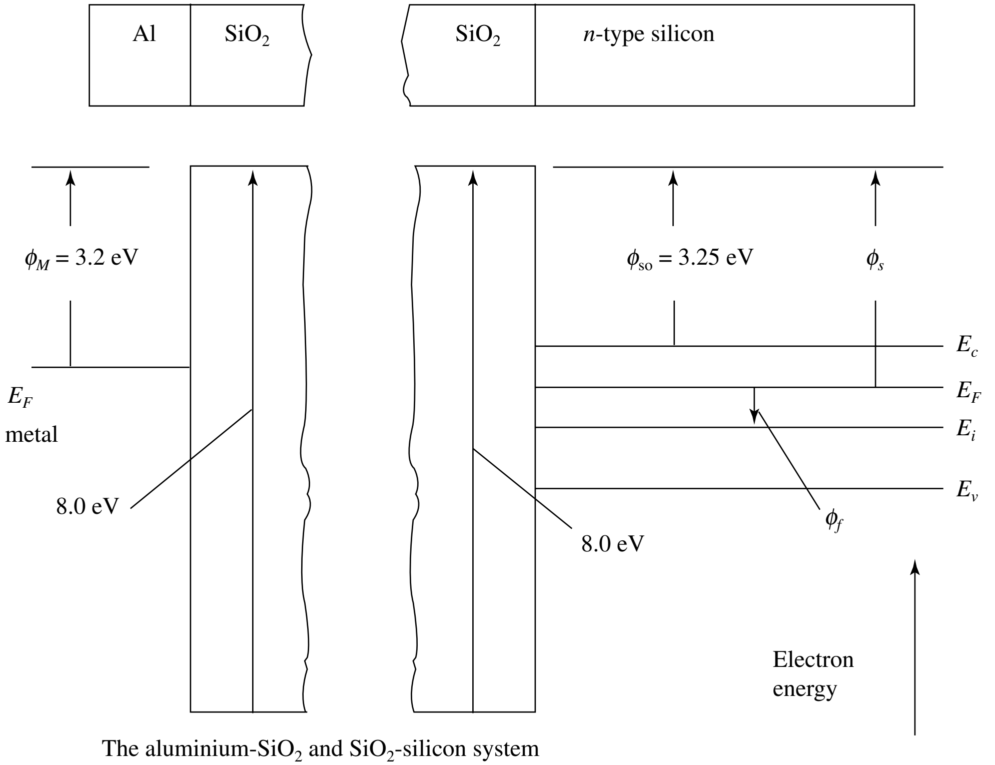


Fig. 1.20 Work-function difference of silicon and aluminium metals

1.6.2 Derivation of the Idealised Current-Voltage (I_D–V_D) Relationship of a MOS Transistor

It is essential to know the basic current-voltage relationship of a MOS transistor for an integrated circuit. Consider an element, with the local cross section of the NMOS channel having length dy in the direction of the channel. The voltage drop dV across this element is given by

$$dV = IdR_{dy} \tag{48}$$

where I is the current flowing through the element and dR_{dy} is the resistance of the element. As channel thickness varies with position as shown in Fig. 1.21, the resistance can be written as

$$dR_{dy} = \frac{\rho dy}{Wx(y)} \tag{49}$$

where W is the width of the channel, ρ is the resistivity of the material and $x(y)$ is the channel depth over the region dy on the y-axis. The resistivity is given by

$$\rho = \frac{1}{qn\mu} \tag{50}$$

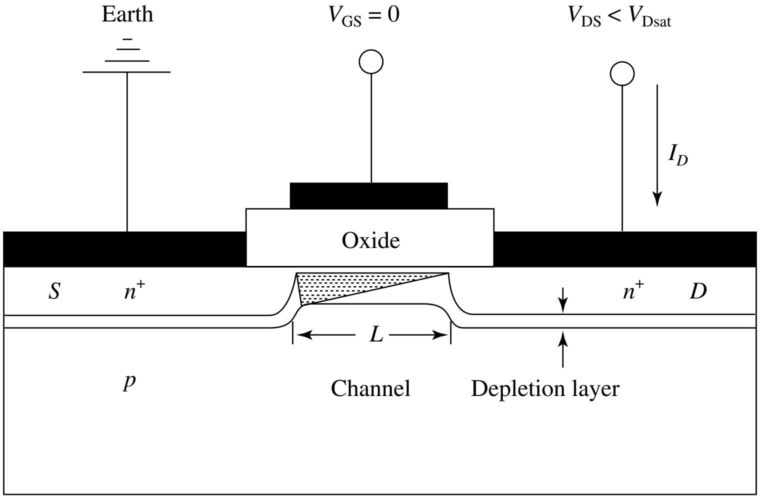


Fig. 1.21 Channel-thickness variation with position

From equations, (49), (50), and (51), the voltage drop across the element can be expressed as

$$dV = \frac{I_{dy}\,dy}{qn\mu x(y)W} = \frac{I_{dy}\,dy}{\mu Q_n(y)W} \tag{51}$$

where $Q_n(y)$ is the charge due to the free electrons (in general, the free carriers) in the element dy.

Prior to deriving the current-voltage relationship, the following notations for the various voltages of the MOS transistor should be known. The gate-to-source, drain-to-source, bulk-to-source, and gate-to-bulk voltages of the MOS transistor are denoted as V_{GS}, V_{DS}, V_{BS} and V_{GB} respectively. For an NMOS transistor operation, the drain is connected to a higher potential and the source to the minimum potential. This results in a potential gradient along the channel.

The total free charge in the element *dy* is

$$Q_n(y) = (V_{GS} - V_T - V(y))C_0 \tag{52}$$

But

$$Idy = Q_n(y)W\mu dV \tag{53}$$

Hence,

$$Idy = (V_{GS} - V_T - V(y))C_0 W\mu dV \tag{54}$$

We can integrate this equation along the length of the channel to obtain

$$I\int_0^L dy = C_0 W\mu \int_0^{V_{DS}} (V_{GS} - V_T - V(y)dV \tag{55}$$

or,

$$I = C_0\mu\frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}(V_{DS})^2\right] \tag{56}$$

or,

$$I = \beta\left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}(V_{DS})^2\right] \tag{57}$$

where $\beta = \beta_0\dfrac{W}{L}$ and (58)

$$\beta_0 = C_0\mu = \frac{\varepsilon_{OX}\varepsilon_0}{t_{OX}}\mu \tag{59}$$

The above equations reveal that the drain current is inversely related to the gate-oxide thickness, t_{OX}. Hence, the oxide quality as well as the geometry of the gate is of utmost importance. To get good quality of the gate oxide, the oxidation process (recipe) is carried out with extreme care. The oxidation processes are explained in Chapter 4. Furthermore, transistor current depends on the transistor width and length; so the dimensions of the transistor have to be perfect to attain the desired characteristics of the MOS transistor.

The current voltage (I-V) characteristics of an NMOS transistor are shown in Fig. 1.22. The current increases with drain voltage for all values of the gate voltage. The I-V characteristics of the transistor can be divided into three regions. In the initial part of the I-V curve, the current increases linearly with the drain voltage. In the second phase, the current approaches the saturation region, and finally in the last phase, the current increases drastically when the device enters the breakdown region. All the three cases are discussed below.

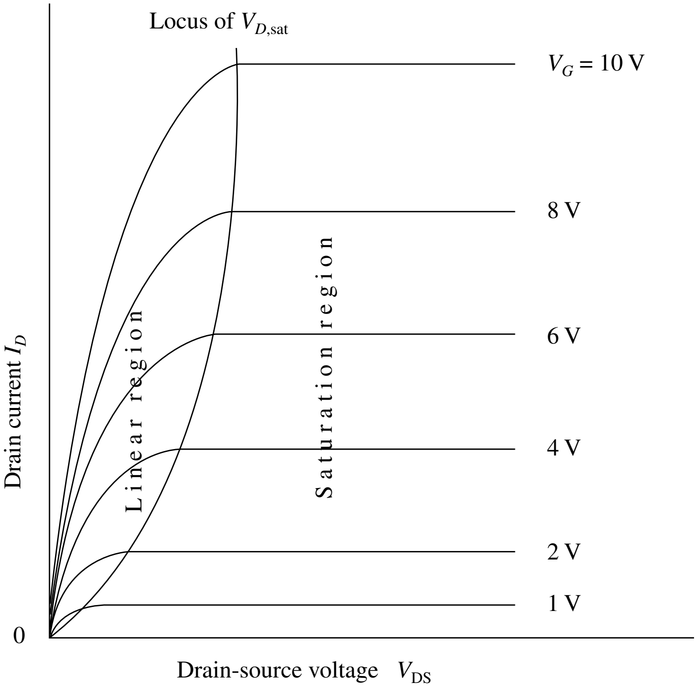


Fig. 1.22 *I–V* characteristic of MOS with gate voltage

Linear Region (Triode Region)

When the drain voltage is greater than the threshold voltage ($V_{GS} > V_T$), a constant channel width is formed from the source to the drain and current flows through that channel. At this region, the channel behaves like a resistor; current increases linearly with source to drain (V_{DS}) voltage as per Ohm's law. Therefore, this region is referred to as the **linear or triode region**, as depicted in Fig. 1.23. Further increase of drain voltage leads to the saturation region.

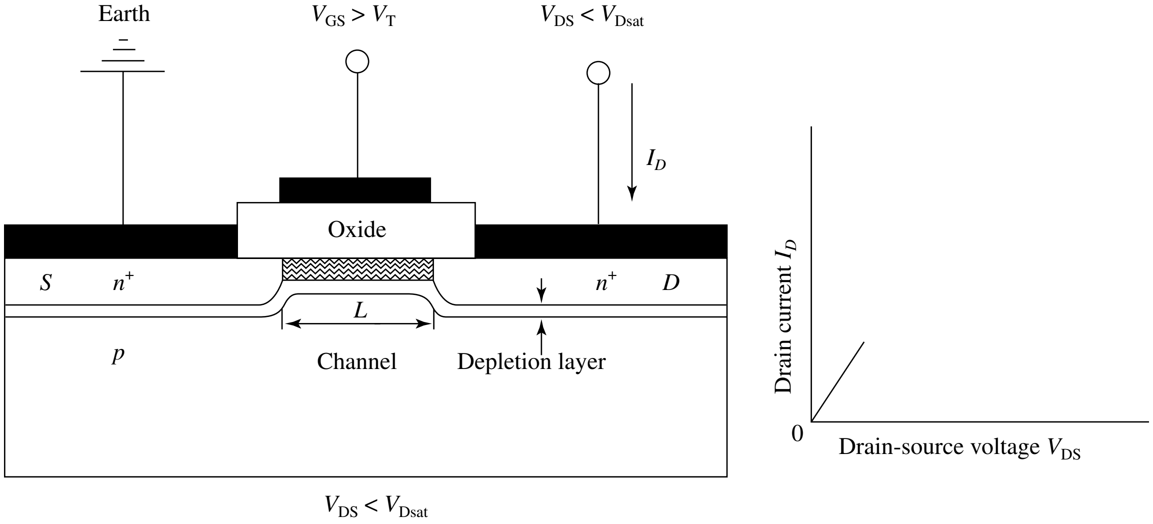


Fig. 1.23 *I–V* characteristics of MOS with increasing drain voltage

Saturation Region

Increase in the drain to source voltage V_{DS} leads to an increase in the electric field at the drain end of the channel. This electric field influences the below gate at the drain end, that causes the channel length to decrease, and subsequently the depletion width increases, as shown at the point P in Fig. 1.24. The movement of P with the increase of drain voltage

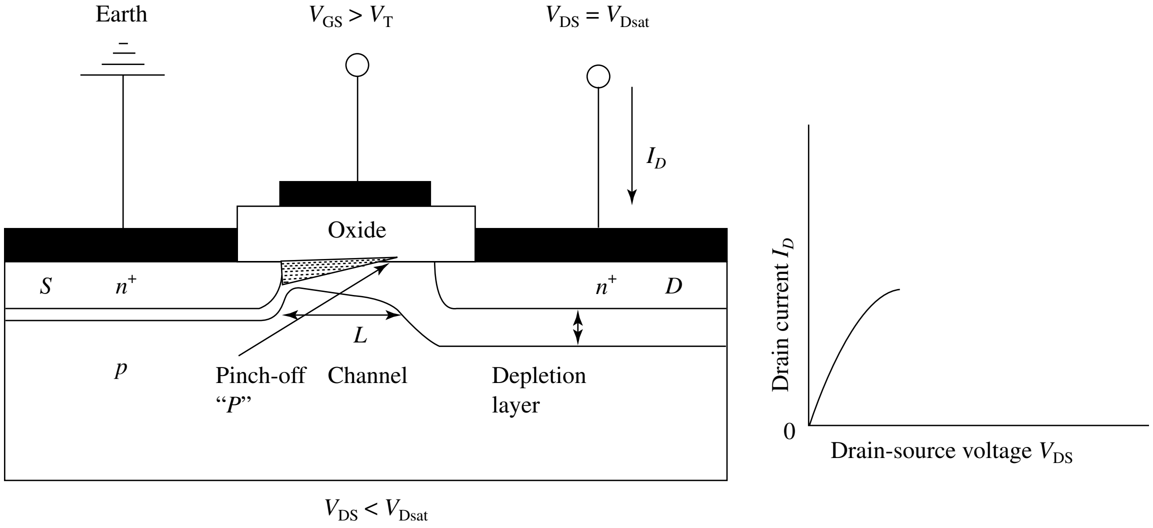


Fig. 1.24 *I–V* characteristics of MOS with increasing drain voltage

is called **channel-length modulation**. Further increase of drain voltage leads to a zero inversion (no channel) at the drain side, called **pinch-off**. This pinch-off condition is defined in terms of the voltages as

$$V_{D,SAT} = V_{GS} - V_T \tag{60}$$

With the increase of V_{DS}, the pinch-off point P moves towards the source and the depletion width increases from the drain to the source end. However, as $V_{\mathrm{D,SAT}}$ remains the same, the current remains the same and a saturation condition is reached. With further increase of the drain voltage V_{DS}, the source electrons are swept away by the high electric field of the drain, and the drain current becomes

$$I_D = \frac{1}{2}\beta(V_{GS} - V_T)^2 \tag{61}$$

It is essential to mention here that the above equation is not quite correct and the current increases very slowly with V_{DS}.

Avalanche Region

When the drain voltage is further increased, the drain depletion width also increases till it touches the source depletion width. At this juncture, the gate loses its control over the transistor and the current increases abruptly, as shown in Fig. 1.25. This phenomenon is

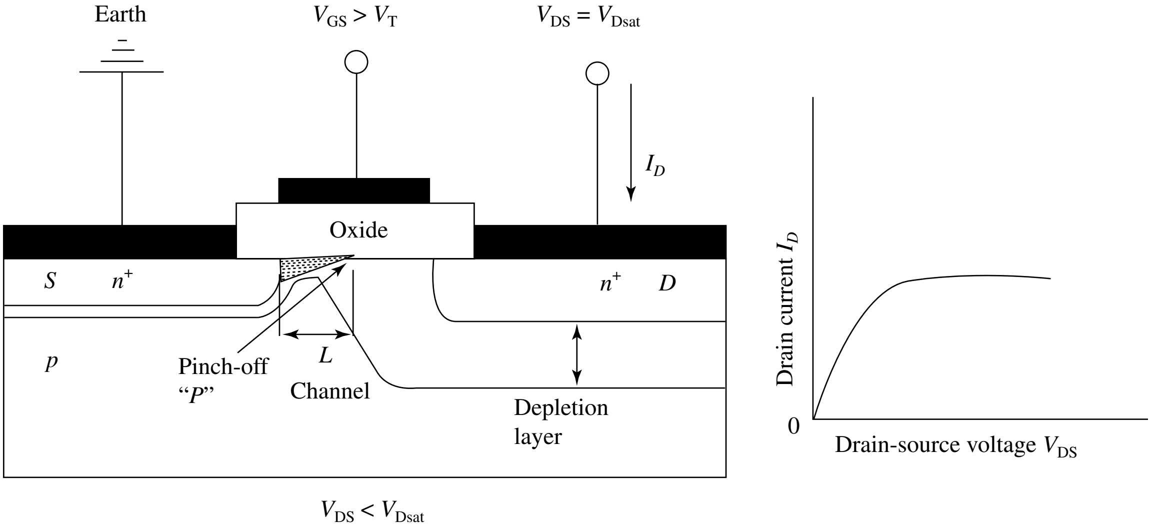

Earth
$V_{\mathrm{GS}} > V_{\mathrm{T}}$
$V_{\mathrm{DS}} = V_{\mathrm{Dsat}}$
I_D
Oxide
S
n^+
n^+
D
p
Pinch-off
"P"
L
Channel
Depletion
layer
$V_{\mathrm{DS}} < V_{\mathrm{Dsat}}$
Drain current I_D
0
Drain-source voltage V_{DS}

Fig. 1.25 *I–V* characteristics of MOS with increasing drain voltage

called **transistor breakdown**. This condition is generally encountered with a very short channel and/or low substrate doping of the MOS transistor. It is found that the breakdown due to the electric field occurs in the range of 2 volts to a few hundred volts for junction doping between 10^{19} cm^{-3} and 10^{14} cm^{-3}. Generally, transistor breakdown takes place where the local electric field exceeds 5 × 10^5 V/m. Furthermore, breakdown is lowered

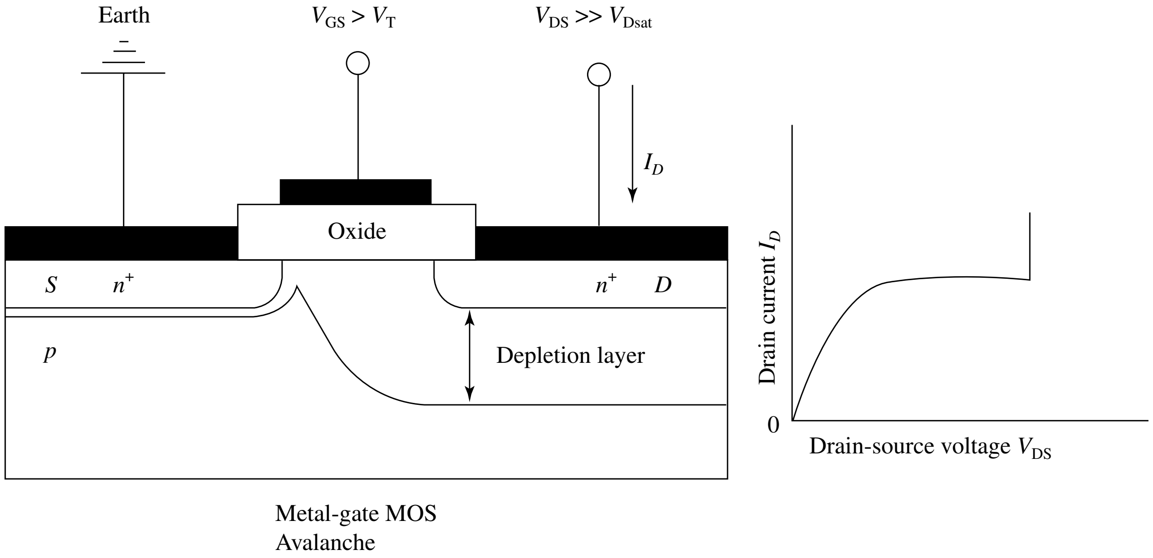


Fig. 1.26 *I–V* characteristics of MOS with drain voltage

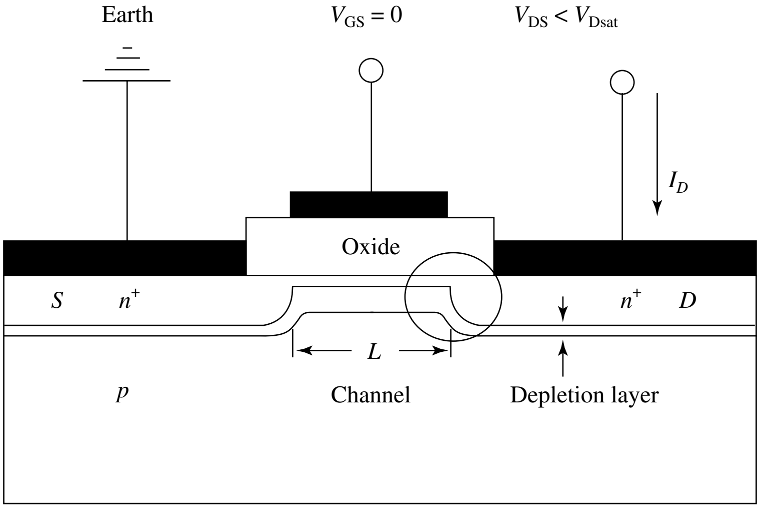


Fig. 1.27 Circle shown in the drain end of MOS transistor is more prone to breakdown

if the junction is graded compared to the perfect step junction, as shown by the circle in Fig. 1.27. This is because the gate electrode has a high field at the curved region. However, to make a perfect step junction is a very challenging issue. In addition, as the dopants are diffused in localised areas where the oxide layer (film) is not present, this results in curved junctions at the corners. At these curved surfaces, the electrical field becomes high; hence the electrical breakdown takes place earlier. The technique of diffusion and the diffusion profile are covered in Chapters 8 and 9 on diffusion and implantation respectively.

Summary

In this chapter, the evaluation of transistor, MOS scaling, physics of semiconductor, energy-band diagram, and the electrical characteristics of MOS transistor have been covered. It was found by Moore that every two years, the complexity of ICs doubles and that leads to new generations of ICs. Generally, these generations are based on the gate length of the MOS transistor, which happens to be the minimum dimension of a MOS transistor. A roadmap called the *National Technology Roadmap for Semiconductors (NTRS)* was drawn by the Semiconductor Industry Association (SIA) to capture the decrease in MOS size (scaling down) from the year 1997 to the year 2012. It was projected that the gate length of around 250 nm in 1997 shall be reduced to around 50 nm in the year 2012, i.e, in the span of 35 years! This leads to technology generations of ICs, namely, Small-Scale Integration (SSI), Medium-Scale Integration (MSI), Large-Scale Integration (LSI), Very Large-Scale Integration (VLSI) and Ultra Large-Scale Integration (ULSI) circuits.

For convenience and simplicity, the energy-band diagram of a silicon semiconductor has been used to predict the electric field, barrier potential, accumulation, depletion, inversion, and channel formation; and the working principle of MOS transistors is described.

The high-frequency and low frequency capacitance-gate voltage (*C-V*) measurement is generally used for MOS capacitor to know the device parameters, and especially for oxide charges described in this chapter. These measurements are essential to characterise (unit process) prior to the fabrication of the ICs so that the bulk produced ICs behave as desired. The current-voltage (*I-V*) characteristics for the three regions of a MOS transistor namely, linear, saturation and breakdown are described.

References

- J D Plummer, M Deal and P B Griffin; *Silicon Fundamental Technology: Fundamentals, Practice and Modelling*, Prentice Hall, 2000
- S M Sze; *Semiconductor Devices Physics and Technology*, John Wiley and Sons, 1985
- D G Ong; *Modern MOS Technology: Process, Devices, and Design*, McGraw-Hill, 1984
- S M Sze, *Physics of Semiconductor Devices*, Wiley-India and Sons, 2005

Multiple-Choice Questions

1.1 Transistor density is roughly increasing every two years by
(a) two times (b) three times
(c) four times

1.2 In intrinsic silicon, Fermi level lies in the forbidden
(a) centre (b) just above centre
(c) just below centre

1.3 How is the transistor threshold voltage related to the silicon substrate doping concentration? Is it
(a) proportional to the square root of substrate doping
(b) proportional to the square of substrate doping
(c) not dependent

1.4 The threshold voltage of a MOS transistor increases if
(a) gate oxide thickness increases
(b) gate oxide thickness decreases
(c) does not depend on gate oxide thickness

1.5 How is the transistor depletion width related to silicon substrate doping concentration? Is it
(a) proportional to the square root of substrate doping
(b) proportional to the square of substrate doping
(c) inversely proportional to the square root of substrate doping

1.6 For higher transistor current, width of the transistor should be
(a) bigger (b) smaller (c) does not matter

1.7 For higher transistor current, length of the transistor should be
(a) bigger (b) smaller (c) does not matter

Descriptive Problems

1.1 Explain Moore's law and its importance in the present VLSI scenario.

1.2 Explain the working principles of NMOS and PMOS.

1.3 What is reason that E_i is not in the centre of the forbidden gap of a silicon semiconductor?

1.4 Calculate the intrinsic carrier concentrations at the temperatures 0 K, 50 K 100 K, 150 K, 200 K, 250 K and 300 K, and plot the intrinsic carrier concentrations with respect to the above-mentioned temperatures.

1.5 When intrinsic silicon wafer is doped with 1×10^{16} boron atoms per cubic centimetre,

(i) Calculate the hole concentration, electron concentration and resistivity

(ii) Show the relative energy level in the energy-band diagram.

1.6 A silicon wafer is doped with 1×10^{15} cm^{-3} of boron, Q_f charge is 1.6×10^{-8} cm^2 and the oxide thickness is 800 Å. Calculate the threshold voltage at the gate of the MOS transistor.

1.7 Find the depletion width of an abrupt junction, doped with 10^{11}, 10^{13} and 10^{15} cm^{-3} of boron atoms in a silicon wafer uniformly doped with 10^{14} cm^{-3} phosphorous atoms at 300 K temperature.

1.8 The gate of a MOS transistor is connected to 5 volts, which has 500 Å gate oxide and 1-volt threshold voltage. Calculate the gate surface charge Q_S.

1.9 Draw the $E_F - E_i$ versus phosphorous doping at 10^{14}cm^{-3}, 10^{16}cm^{-3}, 10^{18}cm^{-3} and 10^{20}cm^{-3}.

CHAPTER

2

Silicon Wafer Preparation for MOS Transistor Fabrication

2.1 INTRODUCTION

In this chapter, the silicon crystal with different types of crystal defects in the context of MOS transistor fabrication is described. Furthermore, the fabrication of the silicon crystal is explained. Thereafter, silicon wafer sizing, grinding, polishing and identification masks are covered. The wafer specifications in view of MOS transistor fabrication and electrical performance are highlighted. The techniques to control crystal defects and the incorporation of impurities at the time of wafer fabrication are discussed. Thereafter, wafer contaminations and their preventions are described.

2.2 SILICON CRYSTAL STRUCTURE

Crystalline silicon is a face-centred cubic unit cell and belongs to the diamond crystal family. Crystalline silicon is made of unit cells. There are three basic forms of cubic unit cells, namely, simple cubic cell, body-centred cell and face-centred cell, as shown in Fig. 2.1. In a unit cell, a silicon crystal is surrounded by four equidistant atoms that are at a fixed distance called the **lattice constant**, and generally denoted by a; the position occupied by the silicon atom in the lattice is called the **lattice site**. The silicon structure is made of two face-centred cubic (FCC) unit cells, where the cells are interlocked at $a/4$ distance from

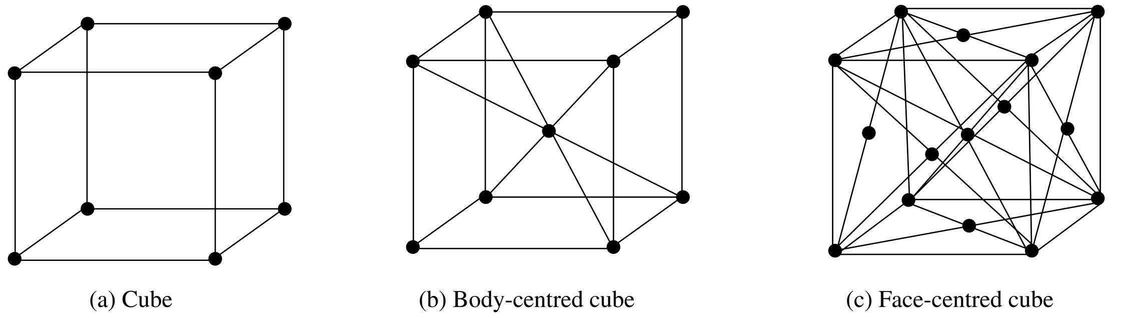


Fig. 2.1 Cubic cell, body-centred cell and face-centred cell

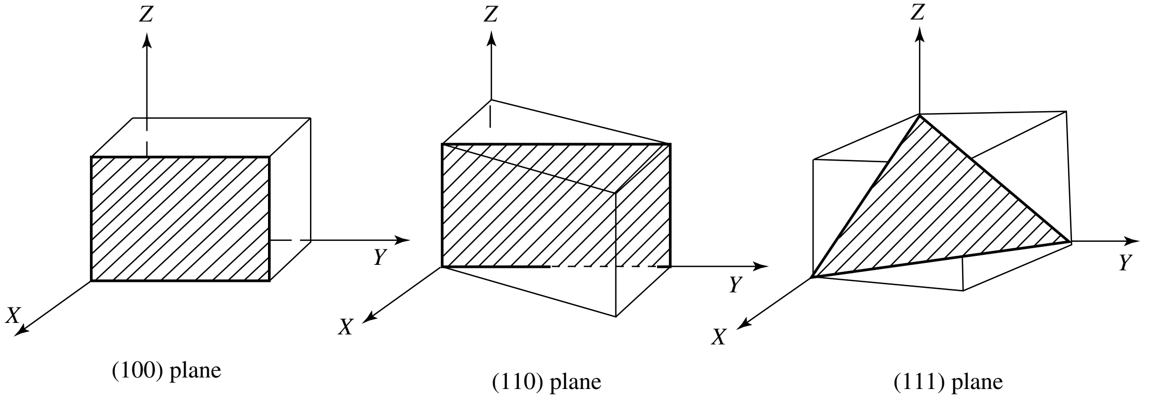


Fig. 2.2 Direction of crystal plane

each other in all the three directions. The face of the unit cells is called **crystal plane** and the growing direction of unit cells is called **crystal direction**. Crystal plane and crystal growth directions are defined by the Cartesian coordinate system, as shown in Fig. 2.2. Crystal orientation dictates its mechanical and electrical properties. Therefore, it is essential to know the crystal orientation and the crystal plane on which the MOS transistor is to be made. A convenient way to represent crystal planes is the Miller index presentation. If the crystal is growing parallel to the *X*-axis by a unit cell ($a = 1$) and does not meet (or meets at infinity) the *Y* and *Z* axis then the Miller index is represented by ($1/a$, $1/\infty$, $1/\infty$) or conveniently represented by (100). This means that the (100) plane intercepts the *X*, *Y* and *Z* axes at 1, ∞ and ∞ respectively. Furthermore, the direction of the crystal growing plane is represented by [100] or {100} as shown in Fig. 2.3. In the generic case, where the crystal is growing in all the three axes in the ratio of *a*, *b*, and *c*, the Miller index is represented in a simplified form as ($1/a$, $1/b$, $1/c$). For example, if the crystal is growing in

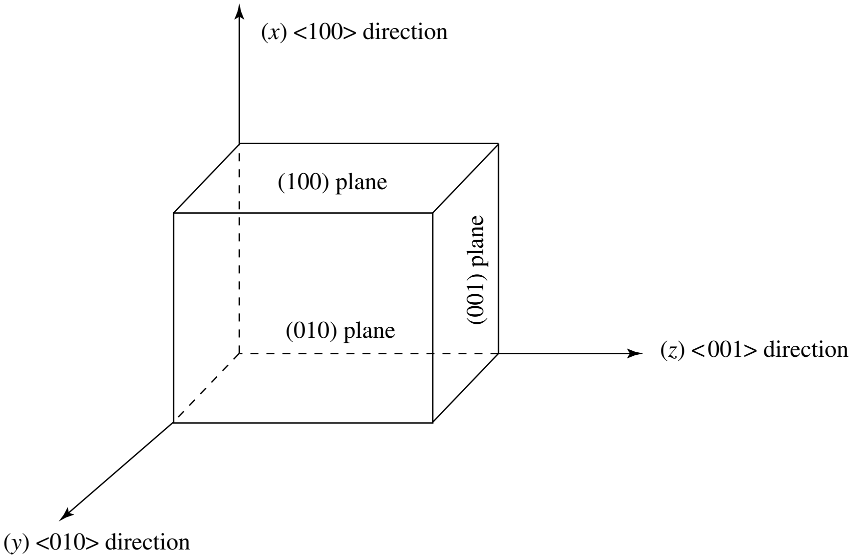


Fig. 2.3 Direction of crystal plane

the ratio of 1, 2, 3 in the *X*, *Y*, and *Z* axes, the Miller index can be represented by (1/1, 1/2, 1/3) or in whole numbers as (632). It is quite possible that the crystal grows in one of the negative coordinates, say on the *X*-axis; in this case, the Miller index is written as (6¯32).

If the cubic structure is symmetrical in all the three axes then the crystal has the same properties in all the three axes and is called the **isotropic crystal**. Different orientation planes contain different numbers of atoms; for example, the number of silicon atoms per centimetre are more in (111), (110) and (100). Thus, the fabrication of the MOS transistor and its electrical properties are highly dependent on crystal orientation. In view of MOS fabrication, the rate of oxidation is significantly dependent on crystal orientation. Similarly, the electrical parameters are significantly different for different crystal orientations; so are the silicon and silicon dioxide interface properties. Interface properties are a function of the number of atoms present in that crystal plane (see Chapters 4 and 11).

2.3 DEFECTS IN A SILICON CRYSTAL

The defects in a silicon crystal degrade the MOS transistor electrical properties like decreased mobility and minority carrier lifetime, lower output current, increase in leakage current, and poor reproducibility; and lead to the creation of generation and recombination centres. In reality, a silicon crystal cannot be made completely defect-free, but the defects can be minimised to the minimum possible values. There are four types of inherent defects present in a silicon crystal as highlighted by the circle in Fig. 2.4. These defects are

1. Point defects
2. Line defects
3. Area defects
4. Volume defects

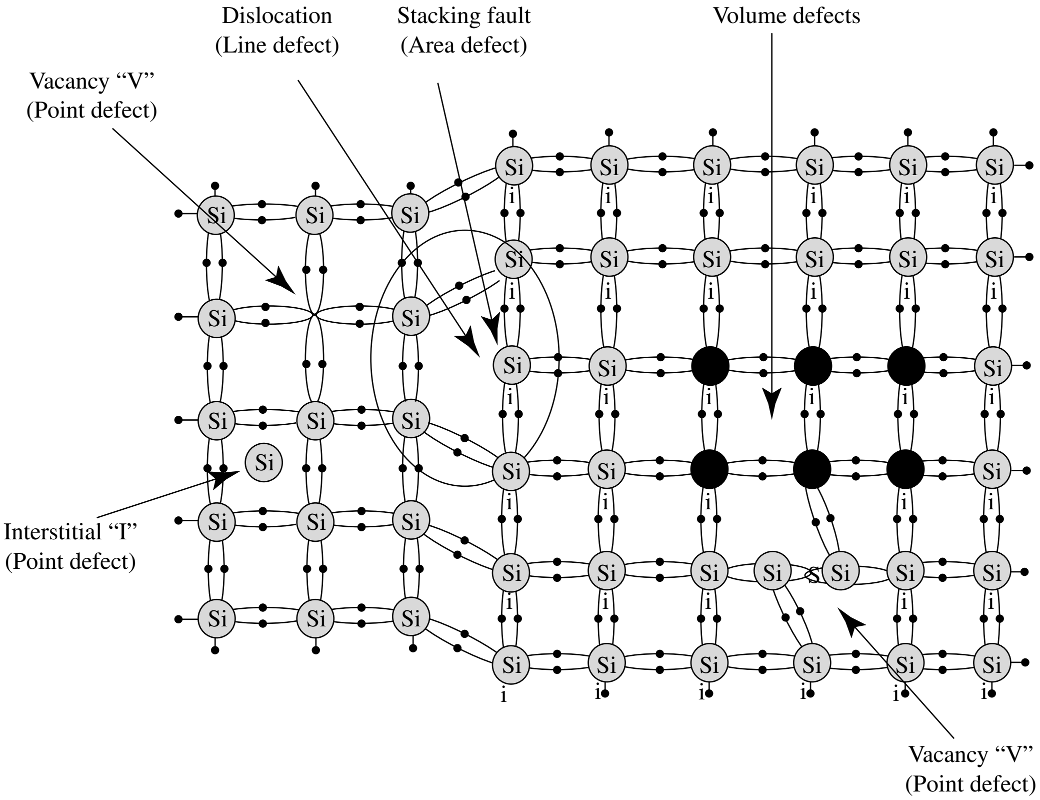


Fig. 2.4 Defects in crystalline silicon

2.3.1 Point Defects

Point defects are the atomic-level defects; therefore they are called "point defects". Sometimes these point defects are also called **native point defects** and are shown in Fig. 2.5.

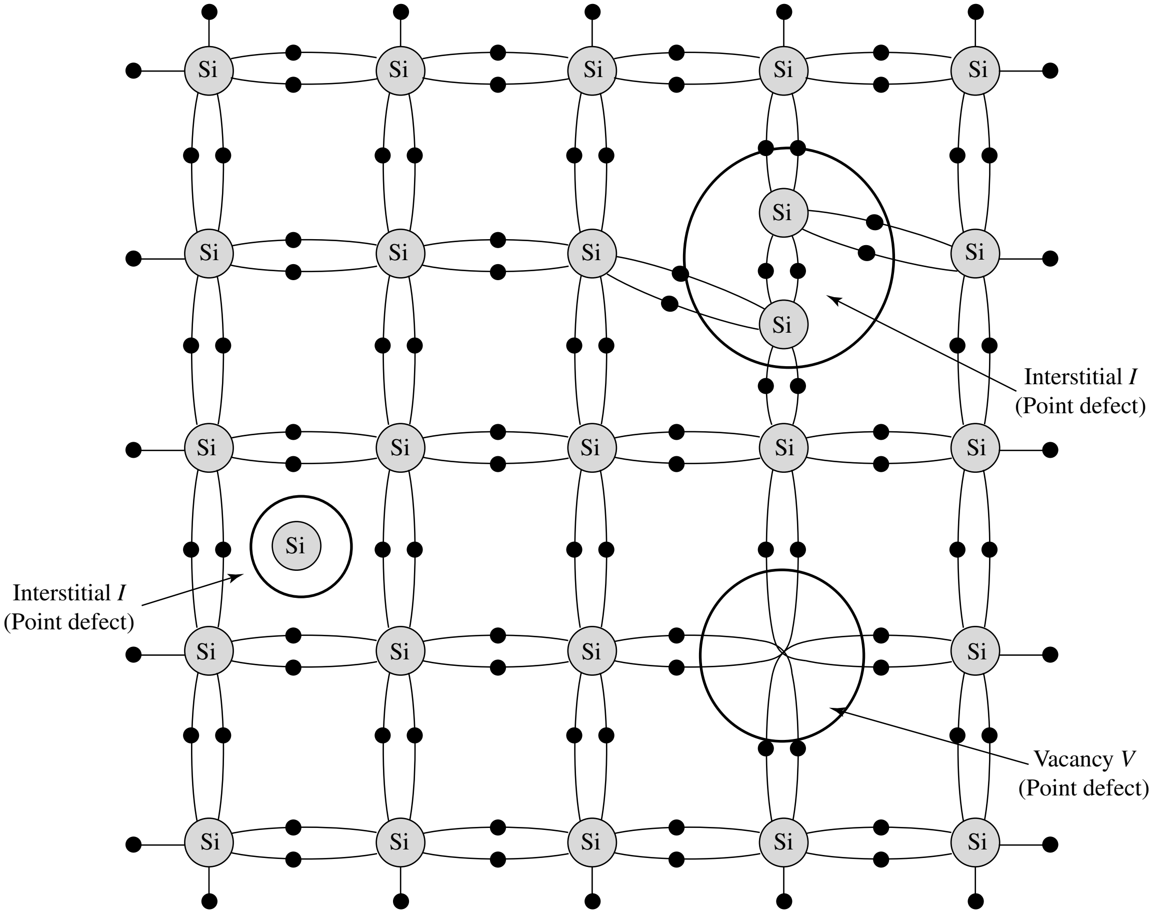


Fig. 2.5 Point defects in crystalline silicon

There are two types of point defects present in silicon. They are external point defect and internal point defect. **External point defects** are due to the presence of different types of elements (impurities) other than silicon atoms. These elements get introduced during the process of silicon-crystal growth. In addition, these elements are also introduced by the diffusion process (intentional doping during MOS fabrication). These atomic impurities either replace the silicon atoms from their lattice sites (substitutional), or they are present in between the lattice sites (interstitial) of the crystal. **Internal point defects** are created due to the absence (vacancy) of silicon atoms in the crystal lattice sites and are represented by *V*. Another type of internal point defect is created due to the presence of silicon atoms themselves in between the crystal lattice sites. These atoms may be bounded or may not be bounded (unbounded) with the neighbouring silicon atoms. These types of defects are called **interstitial defects** and are denoted by *I*, as shown in Fig. 2.5. These point defects (*V* and *I*) increase significantly with temperature and their concentration can go as high as 10^{12}/cm^3 to 10^{15}/cm^3 at high temperatures (see Chapter 8).

2.3.2 Line Defects

Line defects are formed in a line (one dimensional) in the crystalline direction. These defects lead to the growth of extra crystal planes along with the regular crystal planes. Hence, the extra planes are dislocated from the regular crystal planes. These extra line defects may run through from one end to the other end of the crystal. These types of defects are called **edge dislocations** and are shown in Fig. 2.6. If two or more dislocated lines join together and make loops then these dislocations are called **loop dislocations**. A sudden thermal change and stress in the crystal at the time of wafer fabrication and MOS fabrication at high-temperature processes are the major causes of these defects.

2.3.3 Area Defects

When a large number of close-loop dislocations are stacked one above the other, they form **stacking faults**. These stacking faults are in two dimensions; hence they are called “area defects”, as shown in Fig. 2.6. Stacking faults are more in (111) crystal orientation than in (100) orientation. At high temperatures, a large number of silicon atoms leave their lattice site (point defects) which leads to the creation of a large number of extra planes which manifest as stacking faults. This is one of the reasons that MOS transistor fabrication is preferred on the (100) crystal plane. Oxidation of the silicon wafer creates extra silicon atoms and that results in stacking faults. Generation of stacking faults by oxidation is called **Oxidation Induced Stacking Faults (OIFS)**.

2.3.4 Volume Defects

Volume (three dimensions) defects occur due to the agglomeration defects in the silicon. Therefore, these defects are called “volume defects”, as shown in Fig. 2.7. Volume defects are generated mostly at very high temperatures. At high temperatures, the impurities get dissolved in the silicon during crystal formation. When the silicon cools down, the impurities start precipitating and make clusters at preferred places in the silicon due to low solubility in the cooled solid silicon. For instance, the presence of a high concentration

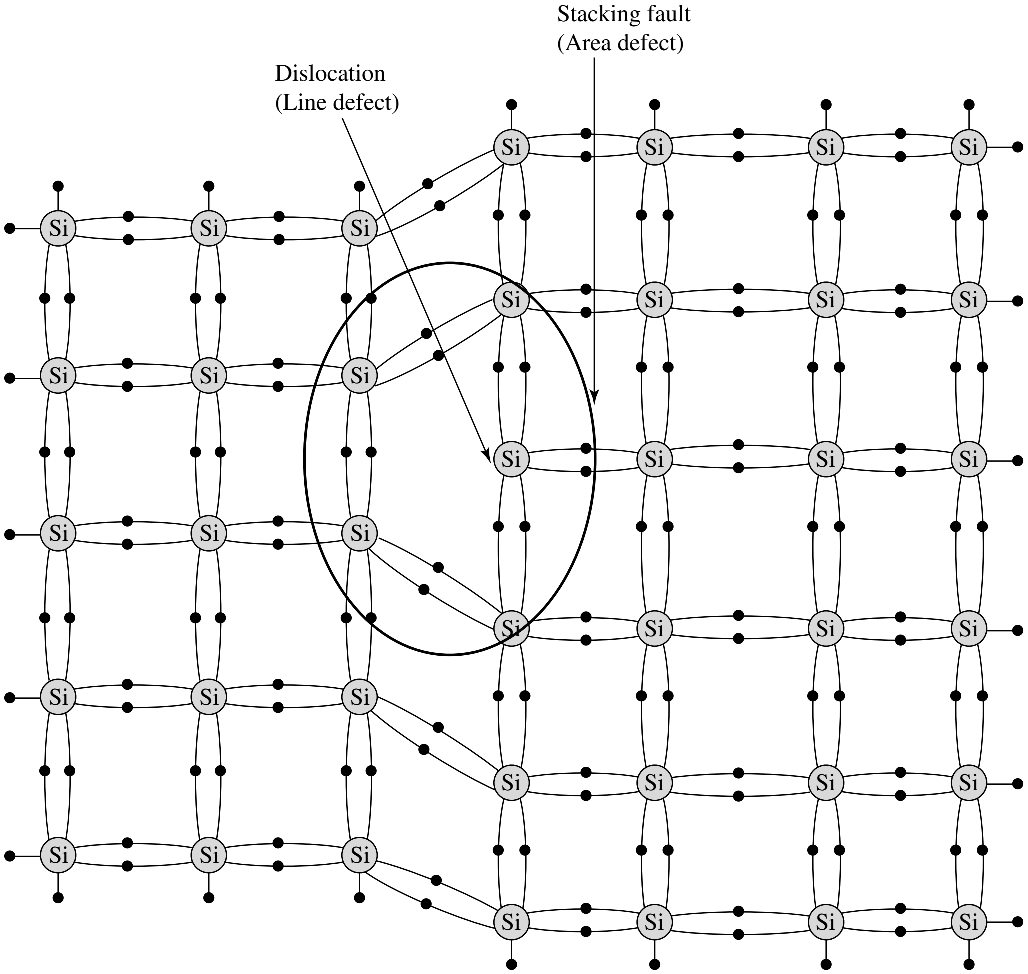


Fig. 2.6 Line and area defects in crystalline silicon

of oxygen atoms occurs in the silicon during its crystallization; especially, in a particular process called the **Czochralski (CZ) crystallization process**. These impurities precipitate and make clusters in the crystal silicon after solidification. Volume defects can be of crystalline and non-crystalline (amorphous) nature. Many of the unattached (unbounded) mobile impurities also move to volume defects and get attached to them. The impurities have serious impact on the threshold voltage of the transistor, when they are formed near the critical areas of the transistor such as gates or junctions. The arrest of mobile impurities in a localized area is essential to improve IC stability. The process of arresting mobile impurities is called **gettering**.

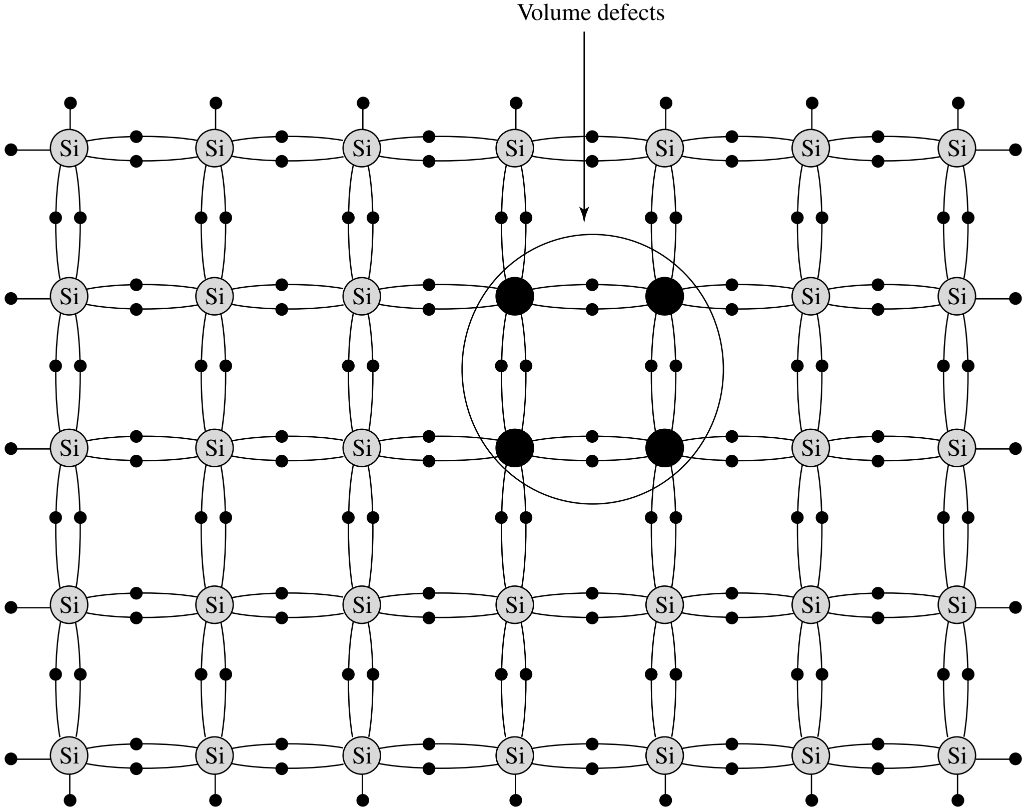


Fig. 2.7 Volume defect in crystalline silicon

2.4 SINGLE CRYSTALLINE SILICON-WAFER FABRICATION FOR MOS TRANSISTOR APPLICATIONS

Silicon crystals with minimum possible defects are made in three phases. In the first phase, the raw material is purified and this purified silicon is called the **Metallurgical Grade Silicon (MGS)**. Thereafter, in the second phase, the MSG is further purified and highly purified semi-crystalline silicon (polysilicon) is obtained. The silicon crystal which comes out after the second phase is called the **Electronics Grade Silicon (EGS)**. In the third phase, the highly purified EGS grade crystalline silicon is transformed into crystalline silicon. Once the crystalline silicon is obtained, it is cut (sliced) into thin pieces. This sliced silicon is called the **silicon wafer**. Sometimes these sliced silicon wafers are also called **starting silicon wafers** in the context of IC fabrication. The processes of making a crystalline silicon wafer using the raw materials are explained below in detail.

2.4.1 Metallurgical Grade Silicon (MGS)

The silicon wafer (starting silicon wafer) is prepared from a particular type of sand (silicon dioxide) called **quartzite** and it is found in abundance on the earth. Quartzite is mixed with carbon (coal or coke) and heated at around 2000°C using the arc-heating technique. At this high temperature, quartzite reacts with carbon and produces liquid silicon and carbon dioxide gas. The carbon dioxide gas is vented away and liquid silicon is transported to separate containers where it cools down and becomes solid. At this stage, silicon comes out to be ~98% pure and the remaining 2% are impurities that contain iron (Fe), aluminium (Al) and other elements. This stage of silicon is called Metallurgical Grade Silicon (MGS).

Solid silicon is then ground and mixed with catalysts and heated at 300ºC temperature in the presence of gaseous hydrochloride. Silicon reacts with the gaseous hydrochloride and converts into silane (SiH_4), chlorosilane (SiH_3Cl), dichlorosilane (SiH_2Cl_2), trichlorosilane ($SiHCl_3$), silicon tetrachloride ($SiCl_4$) and some other products. Out of these products, trichlorosilane ($SiHCl_3$) is chosen for further processing. Trichlorosilane has the property that it boils at around 32°C and stays in liquid form at room temperature. This property is exploited to separate $SiHCl_3$ from the other co-products and other impurities using fractional distillation technique. Then, $SiHCl_3$ is dissociated into silicon and chlorine by the **Chemical Vapour Deposition (CVD)** technique. In the CVD process, $SiHCl_3$ and H_2 are passed in a heated furnace called the **CVD reactor**. At high temperature, $SiHCl_3$ reacts with H_2 and produces Si vapour and HCl gas, as given by the chemical formula mentioned below.

$$2SiHCl_3 + 3H_2 \Rightarrow 2Si + 6HCl$$

Prior to the CVD process, a thin polysilicon diameter rod is fixed (lengthwise) at both ends of the reactor. The silicon vapour produced in the reactor deposits in the form of polysilicon on the thin polysilicon rod. When the required silicon-rod diameter is reached, the CVD process is stopped. This highly pure polysilicon is referred to as the **Electronics Grade Silicon (EGS)**; the impurity level in the EGS is in the range of 10^{13}–10^{14}/cm^3. In other words, one impurity atom is present in a billion silicon atoms and is referred to as parts per billion (ppb)! After obtaining the EGS rod, it is further processed for single-crystal formation in the third phase of the process.

There are two techniques widely used for making single crystals from the EGS polysilicon rod for MOS fabrication. These techniques are known as the Czochralski (CZ) technique and the float-zone technique. Both the techniques are described below with their merits and demerits.

2.4.2 Czochralski (CZ) Technique for Silicon Crystallization

In the C*Z* technique, the electronics grade silicon (EGS) rod is broken into pieces and put into a silicon dioxide crucible for heating. Along with the EGS pieces, a calculated amount of the dopant (boron or phosphorous) is added to make doped silicon wafer for MOS fabrication. The amount (concentration) of the dopant decides the resistivity (in other words conductivity) of that silicon wafer. Then, the crucible is heated at around 1500°C by passing current through wounded carbon (graphite) strips, as shown in Fig. 2.8. At this temperature, both the polysilicon and the dopant are melted. Thereafter, a piece of highly pure and single-crystal silicon rod is inserted from the top of the crucible till it touches the silicon melt. This pure crystal silicon rod is called **seed crystal**. To ensure uniform heating of the melt and thorough mixing of the dopant with the silicon, both the crucible and the seed rod are rotated in opposite directions. Then, the seed rod is slowly pulled upwards.

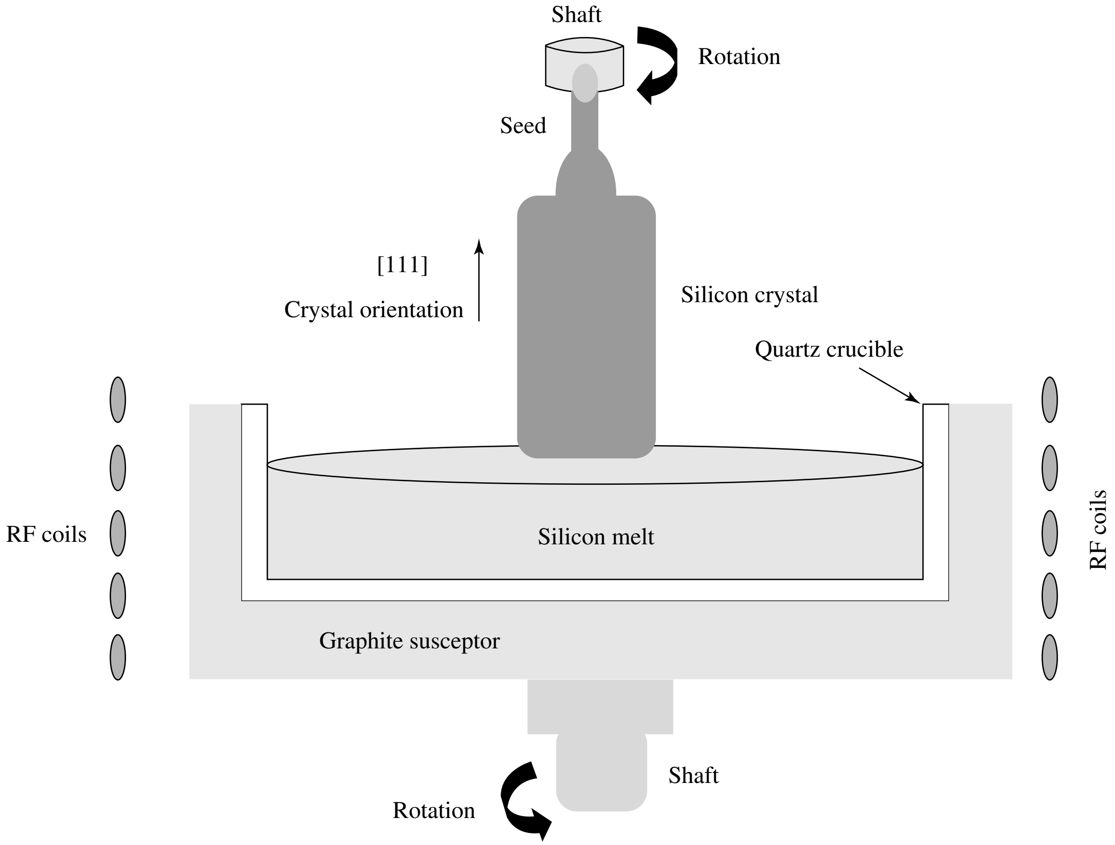


Fig. 2.8 Czochralski (CZ) technique for growth of silicon crystallization

The silicon atoms that get adhered to the bottom of the seed rod become solid when they come out from the silicon melt. The adhered silicon atoms orient themselves in the same fashion as the seed crystal. The crystalline silicon drawn out of the melt is called **boule**. The diameter of the boule is controlled by the pulling rate of the seed rod. The pulling rate, temperature stability, quality of the seed rod, melt mixing, crucible quality and a few other parameters decide the quality of crystallisation and the density of defects in the silicon.

To get the same concentration of the boron and the phosphorous in the silicon boule, different amounts of boron and phosphorous are introduced. This is because of their different segregation coefficients and atomic weights. In the solid silicon, boron incorporates more solidification; whereas, it is the reverse in the case of phosphorous. This is called the **segregation coefficient** and is defined as the ratio of the concentration of the dopant in the solid state with respect to the liquid state of silicon. The segregation coefficient is mathematically defined as

$$k \approx C_s/C_l \tag{1}$$

where, C_s and C_l are the equilibrium concentrations of the dopant in the solid and liquid states near the interface respectively.

Another important parameter of the CZ technique for crystallisation is the pull rate of the boule, as shown in Fig. 2.9. The pull rate has to be well optimised so that the boule

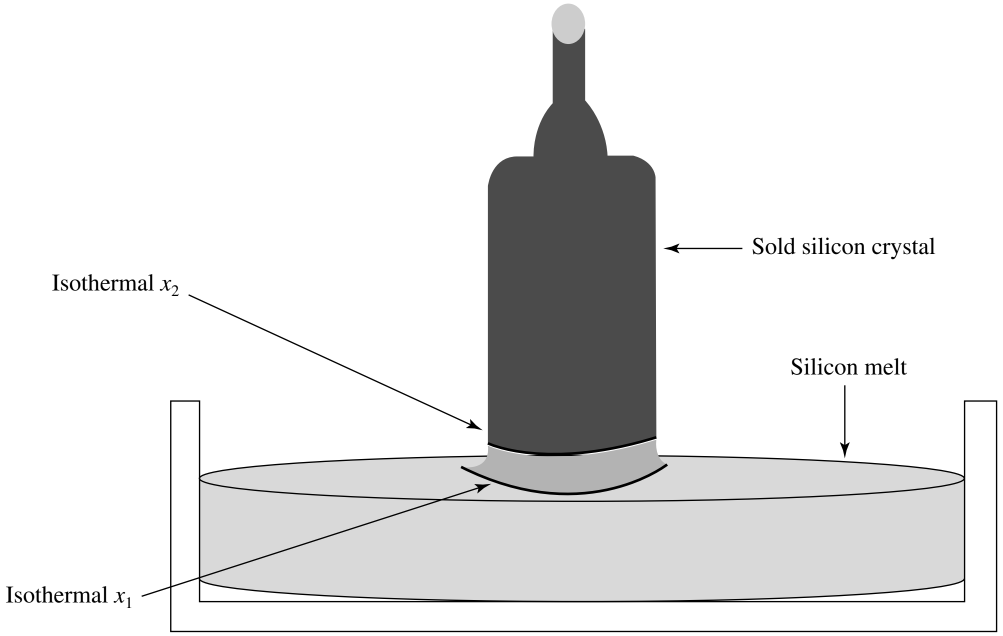


Fig. 2.9 Pulling rate of silicon crystallisation in Czochralski (CZ) technique

has minimum crystal defect and less stress. These defects and stress mainly occur due to the thermal gradient during the pulling out of boule from the silicon melt. The pulling rate is directly related to the thermal gradient and can be expressed as

$$L\frac{dm}{dt} + K_L\frac{dT}{dx_l}A_l = K_s\frac{dT}{dx_s}A_s \tag{2}$$

where L is the latent heat of fusion, dm/dt is the rate of freezing of silicon, K_L is the thermal conductivity of molten silicon, dT/dx_l is the temperature gradient at x_l, K_s is the thermal conductivity of the solid silicon, dT/dx_s is the temperature gradient at x_s, and A_l and A_s are the liquid and solid cross sections respectively.

The pulling rate of silicon can be expressed as

$$\frac{dm}{dt} = P_r AN \tag{3}$$

where P_r is the pulling rate of solid silicon, A is the area of the solidified bottom silicon and N is the density of silicon.

The maximum pull rate can be expressed as

$$P_{r\max} = \frac{K_s}{LN}\frac{dT}{dx_s} \tag{4}$$

if we neglect the term $K_L\frac{dT}{dx_l}A_l$, and take A_s as A.

The heat diffused in the solid is produced by the latent heat at the solid/melt interface; hence the term $K_L\frac{dT}{dx_l}A_l$ can be substituted by zero.

Example 2.1

100 kg of electronics grade silicon is loaded into a crucible to make crystalline silicon ingot by the Czochralski technique. To get 10^{15} atoms/cm^3 boron-doped silicon, how much grams of boron are required? The boron segregation coefficient is 0.8, the atomic weight is 10.8, Avogadro's number is 6.02×10^{23} atoms per mole, and the molten silicon density is 2.53 g/cm^3.

Answer

The segregation of boron in silicon = $10^{15}/0.8$ atoms/cm^3 = $1.25\ 10^{15}$ atoms/cm^3

The volume of boron can be neglected as compared to the volume of silicon.

The volume of silicon = $100 \times 10^3/\ 2.53 = 3.95 \times 10^4$ cm^3.

The number of boron atoms present in the molten silicon

$$= 1.25\ 10^{15}\ \text{atoms/cm}^3 \times 3.95 \times 10^4\ \text{cm}^3 = 4.94 \times 10^{19}$$

Hence, the required boron grams to get 10^{15} atom/cm^3

$= 4.94 \times 10^{19}$ atoms $\times$ 10.8 g/mole/ 6.02×10^{23} atoms/mole

$= 0.886$ mg

The silicon crystal grown by the Czochralski (CZ) technique is capable of producing different crystalline boule diameters with less crystallographic defects, but it suffers with the problem of high oxygen content and carbon impurities. Major sources of these impurities come from the crucible and carbon heating strip. The concentrations of oxygen and carbon are from 10^{17} to 10^{18} atoms cm^{-3} and from 10^{15} to 10^{16} atoms cm^{-3} respectively. These impurities create volume defects in the crystal. In addition, dislocations are also created due to the difference of the melt and seed rod temperatures.

2.4.3 Float Zone Technique for Silicon Crystallization

The Float Zone (FZ) technique is conceptually different from the CZ technique. In this technique, the EGS boule, which is in the form of a polysilicon rod, is directly crystallised by RF heating, as shown in Fig. 2.10. The boule is held vertically by jigs from both ends

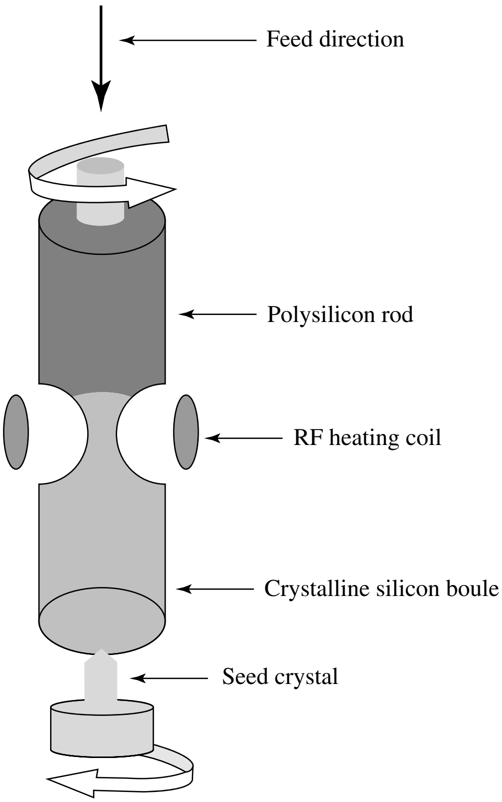


Fig. 2.10 Float-zone technique for single-silicon crystal growth

and a seed rod is placed in intimate contact at one end of the boule. High RF current is passed through a small segment (zone) of the boule across the diameter, from the seed along the length of the boule. The heated segment melts (softens) and becomes viscous. When this viscous segment turns into solid silicon, the silicon atoms rearrange themselves in the manner of seed orientation. Silicon wafer doping is done by introducing the dopant gas that diffuses at the time when the boule is becoming viscous.

The FZ technique is better than the CZ technique in view of the contamination of impurities. This is because the silicon boule never comes in contact with any part of the equipment. Furthermore, FZ is found to be a better technique for high-resistivity silicon fabrication. In contrast, the FZ technique is not suitable for large-size boule-diameter crystallisation. This is mainly due to the limitation of mechanical stability and zone heating. Furthermore, FZ has microscopic resistivity variation due to doping variation. In addition, it is not possible to make silicon wafers (starting wafers) with a wide range of resistivity. Therefore, the FZ technique is hardly preferred for IC applications.

2.5 FABRICATION OF SILICON WAFER FROM THE BOULE

The MOS transistors are made on circular silicon wafers for the sake of convenience in the fabrication process. In addition, MOS fabrication processing needs wafers polished on one side. To make the required wafer thickness and one side polish, the crystalline silicon boule is cut across its diameter so that thin pieces of circular wafers are obtained. The diameter of the wafers is made from a slightly bigger boule diameter using a lathe type of machine. Thereafter, two different sizes of cuts are made along the boule length, as shown in Fig. 2.11. The bigger cut is called the **primary cut** which indicates crystal orientation, and the second cut which is smaller in size than the primary cut, is used for dopant identification (type of wafer). These cut configurations are shown in Fig. 2.12. Usually, the primary cut is referred to as the wafers cut and is used for wafer alignment with respect to the first mask in the lithography process. Detail of mask making is covered in Chapter 5. The lithography process is described in detail in Chapter 6; where, the wafer-to-mask alignment procedure is described through illustrations. Scribe track patterns are the boundaries of the integrated circuits called **dies** and when dies are separated by cutting after IC fabrication, it is called the **IC chip** or simply the **chip**. Usually, dies are of the same size, so the scribe track patterns happen to be a straight line. In addition, the scribe track marks also guide to separate (dice) one die from the other for bonding and packaging for end use.

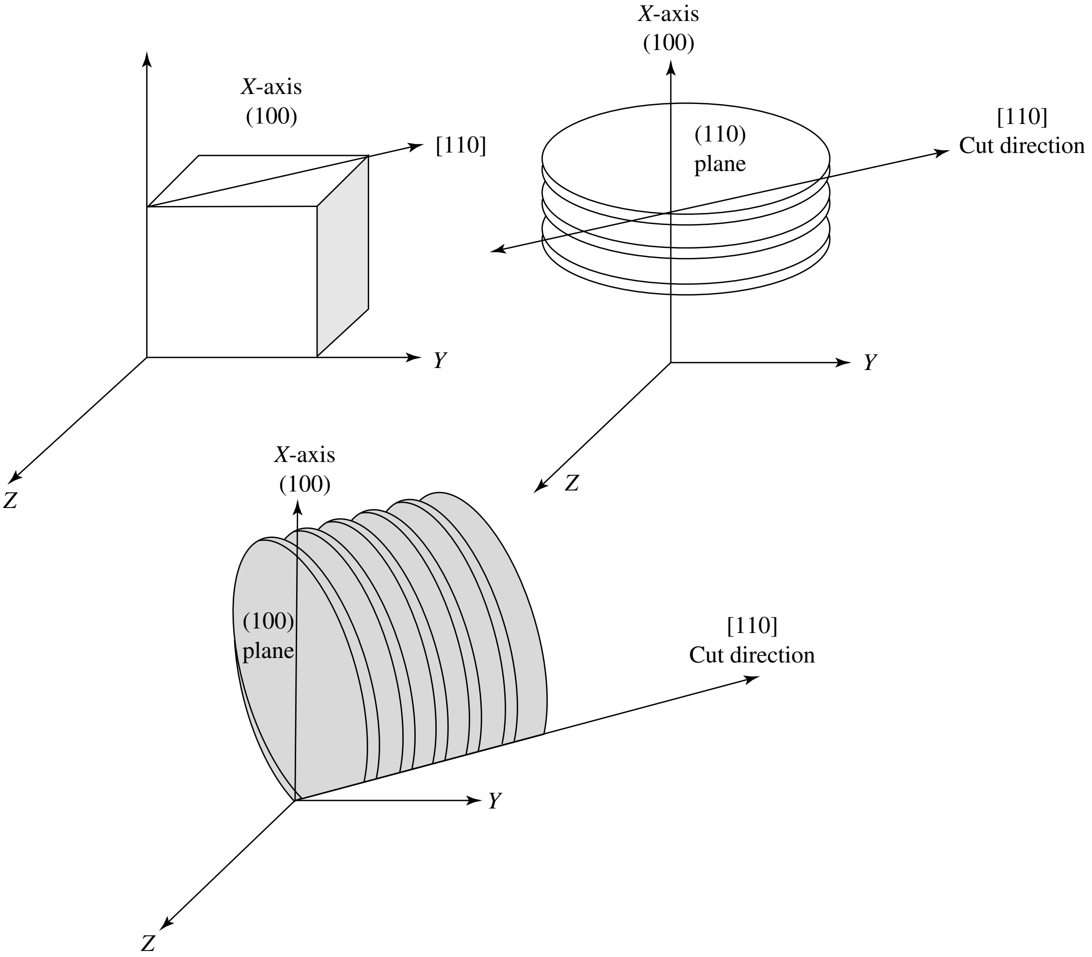


Fig. 2.11 Wafer cut for lithography and identification of wafer

Then the boule is cut along the crystal orientation (110) which is perpendicular to the crystal orientation (100), as shown in Fig. 2.13. The wafer thickness is kept marginally more than the required wafer thickness; the desired wafer thickness is finally obtained by the grinding and polishing processes. These slashed wafers are then mounted onto a special type of flat cast iron jig for grinding and polishing, as shown in Fig. 2.14. After that, the wafers are ground by rubbing with an abrasive, onto a highly polished flat cast iron tool, as shown in Fig. 2.15. Grinding starts from a coarse abrasive and leads to a fine abrasive in steps. Generally, silicon carbide or diamond powder is used as the abrasive. After the grinding process, the wafers are demounted and again mounted with the opposite side onto the jig, and thereafter, the wafers are ground in the same way as explained before.

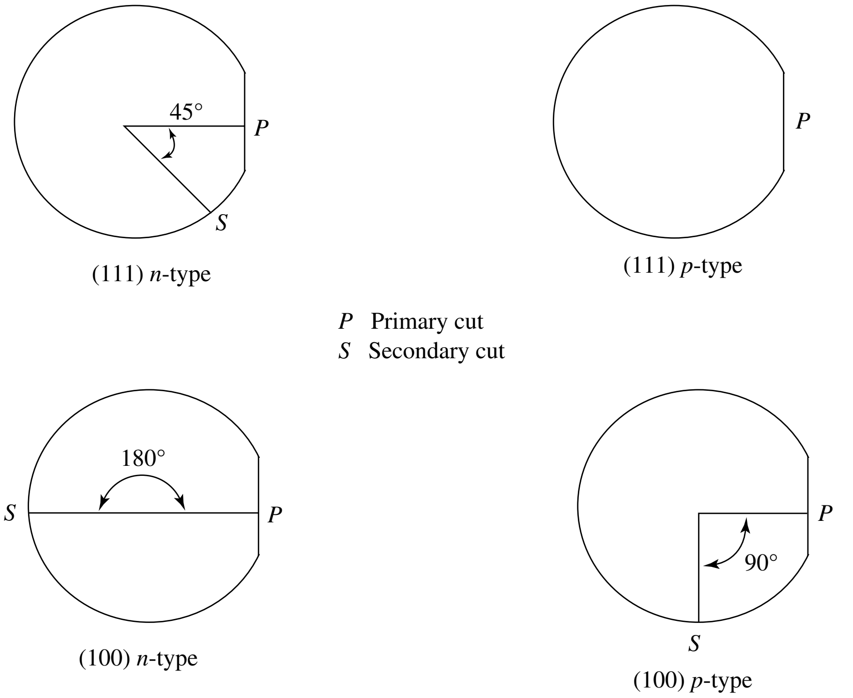


Fig. 2.12 Identification of type of wafer through wafer cuts

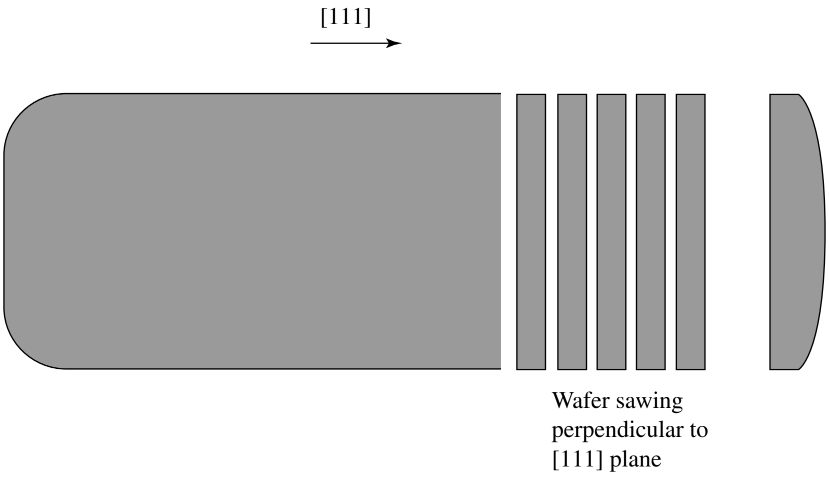


Fig. 2.13 Boule cutting perpendicular to (100) crystal orientation

Next, the fine grinding is carried out with fine alumina abrasives. Thereafter, the wafer is polished with a fine grade ferric oxide or cerium oxide using optically flat polishing tools. In place of ferric oxide or cerium oxide, sometimes, a diamond abrasive is also used.

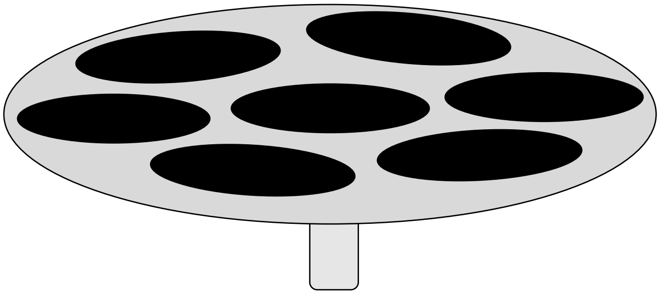


Fig. 2.14 Wafer mounting on optically flat plate for grinding and polishing machine

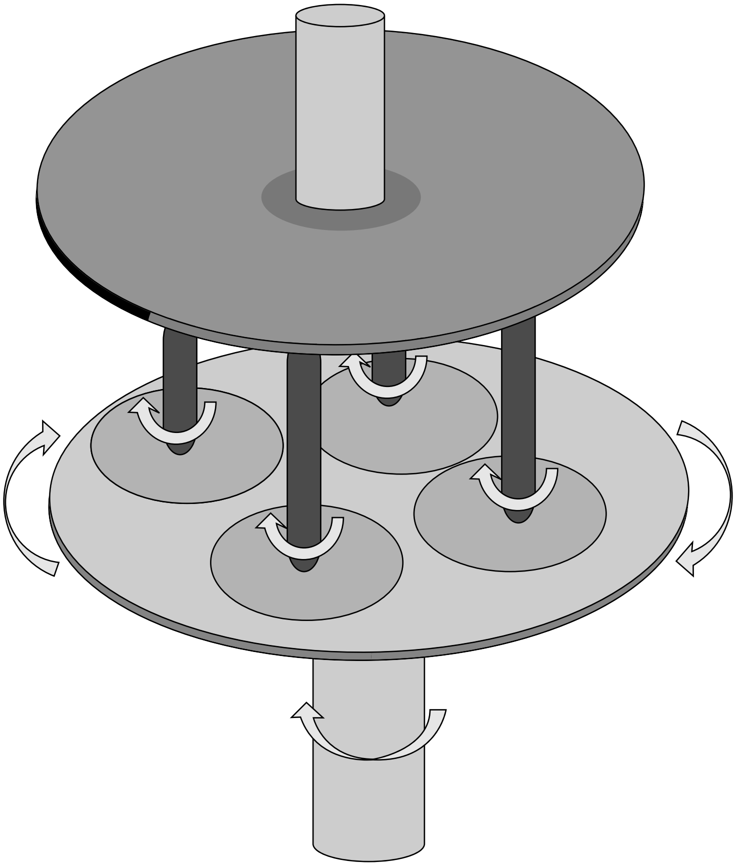


Fig. 2.15 Wafer grinding and polishing machine

The final polishing of the wafer is carried out by the Chemical Mechanical Polishing (CMP) technique. CMP polishing is done with fine silicon dioxide (or diamond) particles in an alkaline solution using optically flat polishing tools. CMP polishing results in a highly polished (mirror polished) wafer surface. After that, the wafers are demounted from the jig and finally, the wafers are ultra cleaned and then packaged.

2.6 SPECIFICATIONS OF THE SILICON WAFER

The electrical characteristics of a MOS transistor are governed by the silicon wafer specifications. Therefore, the selection of a proper silicon wafer is extremely important for a particular integrated circuit. For instance, if the defects density is higher, it will lead to higher carrier traps, leakage current, lower breakdown and change in the threshold voltage (V_T). The main specifications of a wafer are its thickness, diameter, wedge, flatness, crystallography, defects, impurities, the type of dopant, and resistivity. The diameter of the wafer is selected according to the IC fabrication (process) line. For example, if a manufacturer has a fabrication process capability (equipment) of 12-inch diameter silicon wafer, it is called the **12-inch fabrication process line** or in short, the **12-inch process line**. The wafer wedge (taper) is another important specification of wafers. Integrated circuit process equipments have certain tolerance to the adjustment of the tapered wafers. In the lithography process, the distance between the wafer and the mask will vary from one end of the wafer to the other end. This will affect the fidelity of the mask patterns translated on the wafer due to light diffraction. In addition, if the wafer has wedge then the mechanical pressure on the wafer will be uneven and that may lead to breaking of the wafer during the lithography process. The surface of the wafer must be optically flat and well polished. This ensures the fidelity of pattern transfer by lithography process on the wafer. The electrical characteristics of a transistor are found to be the best if the transistors are made on the (100) crystal plane. Hence, an accurate wafer cut of the crystallography axis is very important, as the cut depicts the orientation of the silicon crystal. It has been mentioned previously that the CZ grown crystal contains significant levels of oxygen and carbon elements in the wafer. In addition, other traces of impurities such as iron, gold and other inorganic elements are also present in parts per million (ppm) in the wafer. These impurities create stacking faults and other crystallographic defects in the wafer. Hence, these impurities should be within a specific level. The type and concentration of the dopant in the starting silicon wafer is very important. For the CMOS process, the wafer should be lightly doped, so that subsequent doping, threshold voltage adjustment and channel stop can be achieved. The details of these processes are described under MOS process flow, Chapter 3. Generally, wafer-dopant concentration is specified in terms of the wafer's resistivity.

2.7 DEFECTS AND IMPURITIES IN THE SILICON WAFER

Impurities and defects in silicon wafers cannot be eliminated completely. On top of that, these impurities and defects increase during IC fabrication. To control and minimise these defects and impurities, additional process steps are needed during IC fabrication. These processes are the **annealing process** and the **gettering process**. Annealing minimises the defects in the wafers and gettering arrests the mobile impurities.

2.7.1 Annealing Process

The defects in the silicon introduces allowed electronic levels in the forbidden gap; hence, the characteristic of MOS transistor changes. To curb the defects, the annealing process is carried out. In the annealing process, wafers are kept at high temperatures in nitrogen or a mixer of $N_2 + H_2$ gases for a specific period of time. The annealing process repairs the initial defects as well as the defects introduced during the high-temperature IC fabrication process. Annealing also helps to place the dopant into substitutional sites. It is found that 90% of the dopant should occupy the substitutional sites of the silicon lattice site to get good electrical characteristics of MOS transistor. Generally, annealing is immediately followed by the high temperature process steps.

2.7.2 Gettering Process

Apart from crystal defects, impurities also introduce energy levels in the forbidden gap of the crystal. Generally, these impurities are inorganic and mobile in nature. Their movement increases with an increase in temperature. When impurities move to the critical areas of a transistor, say below the gate, the threshold voltage of that MOS transistor changes. As the movements of the impurities are erratic in nature, the transistor's electrical performance also becomes erratic with time and that leads to unstable electrical characteristics of the MOS transistor. Therefore, it is essential to arrest these impurities and confine them far away from the critical areas of the transistors. It is mentioned under process flow in Chapter 3 that either Phosphorous Silicate Glass (PSG) or silicon nitride film is deposited on top of the wafer for IC passivation. The PSG glass has the property to getter alkaline ions such as Na, K, etc. Unfortunately, in the presence of moisture, the phosphorous dopant which is more than 4%, reacts and produces phosphoric acid that corrodes metal lines used for interconnections. Therefore, the silicon nitride film over the IC is preferred in place of PSG. Silicon nitride prevents alkaline ions to diffuse from the environment into the wafer.

Metal impurities like Fe, Au, etc., are getters that use different mechanisms. These mechanisms are: intrinsic gettering and extrinsic gettering. In the mechanism of intrinsic gettering, impurities that are already present in the silicon wafer are exploited for gettering purpose. For example, under appropriate processing conditions, the oxygen atoms react with the silicon atoms and form silicon dioxide. Due to a mismatch of the silicon and the oxygen atom sizes, they create volume defects. These volume defects have enough space to accommodate (getter) bigger metals impurities. The defects are intentionally created at the back of the wafer by ion-implantation, phosphorous diffusion and polysilicon deposition, and this is called **extrinsic gettering**. These defects produce enough space to accommodate wandering impurities. In the extrinsic gettering process, defects present in the wafer such as dislocations, stresses, grain boundaries also getter the impurities.

2.8 WAFER CONTAMINATIONS

In the previous sections, types of impurities and their sources have been mentioned. These impurities are inherently present in the starting wafers. These wafers get further contaminated with impurities during the IC fabrication stages. Generally, environment conditions, chemicals and gases are the key external sources of these contaminations.

2.8.1 Environmental Contaminations and Precautions

Particulate contamination has a great impact on the IC yield and its electrical behaviour. A large number of different sizes of the particulates are suspended in the atmosphere. In addition, these particulates are generated by the operators, the equipments and the surroundings. Hence, it is essential to check and remove these particulates from the IC processing areas. Unfortunately, particulates cannot be completely eliminated. The air-filtering technique is used to remove particulates from the air. Effective air filtering is done by close loop (recirculation in the processing area) filtering system. For this process, a high-quality air filter is used. These air filters are called **High-Efficiency Particulate Air (HEPA)** filters. Air is passed through the HEPA filters which are fitted in the **Air Handling Unit (AHU)** of the air conditioning system. The HEPA filter is made of perforated fibre sheets. Air is pushed through the HEPA filter in the processing area from the top of the processing room and sucked from the floor of the processing area. Now the sucked air from the floor is filtered through the HEPA filter and sent back to the processing room again. This process continues and the air particulates are reduced significantly. To increase the particulate filtering efficiency, the HEPA filters are folded in a corrugated fashion. This corrugated filter also makes laminated (sheet) air flow in the process areas for better air filtering. This system of air filtering is called the **laminar flow** system. Nearly 99.9% of the particulates are filtered by the laminar flow system. To make laminar flow system for the whole fabrication area is a costly affair.

The particulates filtered IC processing areas are called **clean rooms** and they are classified on the basis of particle filtering size. In the low budget research laboratories, small-sized localised laminar systems are used. Generally, these air filters are configured as bench type, thus it is called the **laminar bench flow system** or in short, the **laminar bench**.

2.8.2 Clean-Room Classification

The presence and size of particulates in the clean room is classified by **Class *X***, where *X* is denoted by the total number of particulates per cubic foot. The nomenclature of Class *X* means that the particulates of 0.5 μm size should not be more than *X* in number. For example, Class 1000 means that the particulates of 0.5 μm size should not be more than 1000 in number per cubic foot of air. Similarly, Class 100 means that the particulates of 0.5 μm size should not be more than 100 in number per cubic foot. In other words, clean-room classification is done on the basis of the number of particulates of 0.5 μm size present per cubic foot. Clean-room classifications are shown in Fig. 2.16.

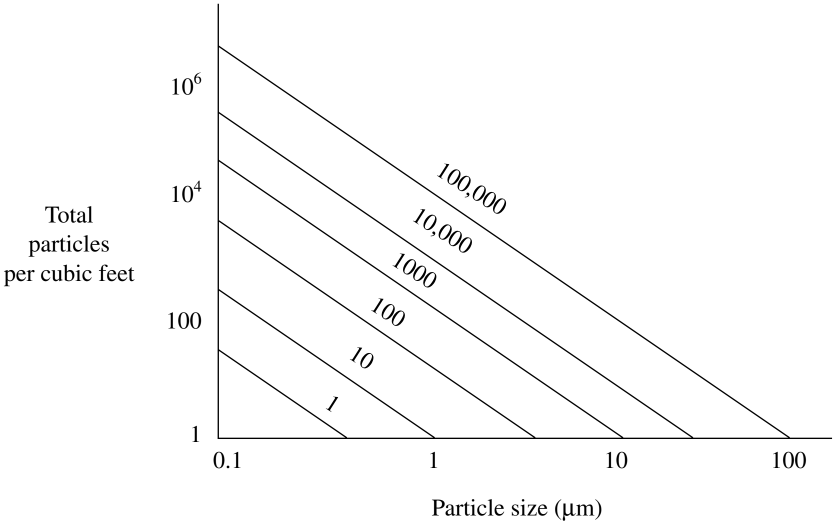


Fig. 2.16 Classification of laminar flow systems

The particulates are also emitted from the operators, the processing equipments, walls of the room, etc. To prevent particulates emanating from the bodies of the operators, the operators are covered by special type of suits and masks. In addition, human beings are substituted by robots to reduce human-emitted particulate to the maximum extent. Furthermore, the moving parts of the processing equipment are located (if possible) away from the IC processing areas. Furthermore, the processing rooms and the surrounding areas are made free of particulate emanation.

2.8.3 Gas Contamination and its Preventions

Gases are another source of particulates. Gases are extensively used in the MOS transistor fabrication process; such as for drying, cleaning, and control of fabrication processing equipments, etc. Highly pure and particulate free gases are used for these purposes. In spite of this, particulates come from gas and its containers, gas pipelines and delivery systems. To further check for particulates that are coming from the gases, a localised gas filter is used.

Apart from the particulates, gases may contain moisture. Moisture sources are the same as those of the particulates. Moisture has severe effects on IC fabrication, especially in the lithography process. To overcome this problem, gases are filtered for moisture in the wafer-processing area and in the wafer-processing equipment. To reduce the moisture of the room, dehumidifiers are also attached with the air-conditioning systems.

2.8.4 Chemical Contamination and its Prevention

Chemicals are another source of wafer contamination. The chemicals used for IC fabrication can have foreign particulates and the particulates of the chemical itself. In addition, chemical contamination comes from the IC fabrication process in the form of chemical residuals. Chemical contamination has adverse effects on the MOS transistor geometries and electrical characteristics. Therefore, chemical particulates should be filtered from the chemicals; in addition, chemical residuals should also be removed completely from the wafers. Apart from chemical particulates and contaminations, wafer cleaning has great impact on the transistor's electrical characteristics. In earlier days, it was noticed that the threshold voltage of the MOS transistor used to shift in spite of the transistor being made in identical process conditions. It was later found that the cleaning procedure was responsible for the threshold voltage shift of the transistor. In 1970, Kern and Puotinen of RCA, USA, came up with a better cleaning procedure with the chemistry of silicon cleaning understanding. Later, Kern and Puotinen's cleaning procedure came to be known as the **RCA wafer cleaning** procedure. With time, many derivatives of RCA cleaning have been developed, but the connotation remains the same. The RCA cleaning procedure is done in two process steps. These steps are based on the pH values of the cleaning solutions. In the first step, the wafers are cleaned in a high-pH (SC-1) solution. SC-1 is made of hydrogen peroxide (H_2O_2), ammonium hydroxide (NH_4OH) and water (H_2O) in the ratio of 1:1:5. Typically, the wafers are kept in the SC-1 solution for 10 minutes at 70–80°C. A high-pH (SC-1) solution reacts with the silicon surface which contains Cu, Ag, Au, Zn, Hg, Co, Ni, and Cd metals; it oxidises these impurities and then these impurities get dissolved in the same SC-1 solution. This oxidising and dissolving process in the SC-1 solution not only removes the impurities, but also removes the particulates from the wafer surface. In the second step, wafers are dipped into a low-pH (SC-2) solution. The low-pH solution (SC-2) is made of hydrogen peroxide (H_2O_2), hydrochloric acid and water (H_2O) in the ratio of 1:1:6. Typically, wafers are kept for 10 minutes at 70–80ºC. In this process, the alkaline ions and cat ions like Al^{+3}, Fe^{+3}, and Mg^{+2} are dissolved in the SC-2 solution. These alkaline ions and cat ions are formed in the SC-1 solution.

Sometimes, the particulates firmly stick onto the wafer and cannot be removed easily by the chemical cleaning procedures. Hence, to remove them from the wafer, a rigorous and brute physical force is required. The best way to remove these particulates is by ultrasonic cleaning. Wafers are placed inside a water container (or in a chemical) and that is placed inside the water filled ultrasonic machine. The ultrasonic machine generates ultrasonic waves excited by a piezoelectric crystal that is fitted at the bottom of the container. The ultrasonic waves are transmitted to the water through the container and finally to the wafer. The transmitted ultrasonic waves create bubbles in the water (or chemical); these bubbles increase in volume with time and finally they burst and produce tremendous shock waves in the water (or chemical). These enormous mechanical shocks knock out the sticking particulates from the wafer surface.

The wafers procured from the wafer-manufacturing industries are highly cleaned; hence, in the first step (usually the oxidation step) of IC fabrication, wafers are not cleaned and they can be directly processed for MOS fabrication. If wafers are exposed to the environment, which usually happens in the laboratories, then they must be removed prior to oxidation.

Summary

In this chapter, the silicon crystal, silicon crystallisation, silicon wafer sizing, grinding, polishing and wafer identification marks, in reference to the MOS transistor fabrication, have been described. The techniques to control crystal defects, and the incorporation of impurities at the time of wafer fabrication have been discussed. Thereafter, preventions of wafer contamination are described.

Crystalline silicon is a face-centred cubic unit cell and belongs to the diamond crystal family. This silicon unit cell is surrounded by four equidistant atoms at a fixed distance and is called *lattice constant*, and the position occupied by the silicon atom in the lattice is called *lattice site*. If a crystal is growing parallel to the *X*-axis then it is represented by (100) and the direction of the crystal growing plane is represented by [100] or {100}. Similarly, the silicon growing in the *Y* and *Z* planes are represented by (010) and (001) and their crystal-growing directions are represented by [010] and [001] respectively. As the (100) plane has less number of silicon atoms, MOS transistor is made on this plane, so that the oxide and silicon oxide/silicon charges are less manifested on the plane. Planes of different orientation contain different number of atoms; hence, the silicon crystal possesses different electrical and mechanical properties.

Crystalline silicon has four defects, namely, point defects, line defects, area defects and volume defects. These defects degrade the MOS transistor electrical properties like mobility, minority carrier lifetime, lower output current, increase in leakage current and poor reproducibility; and lead to the creation of generation and recombination centres. These defects are further increased due to the incorporation of impurities from environment conditions and chemical gases during IC processing.

To control the dust particulate, in the IC processing areas, air is filtered using HEPA filters. The number of air particulates are classified by *Class X*, where *X* denotes the total number of particulates per cubic foot. The nomenclature of *Class X* means that the particulates of 0.5 μm size should not be more than the number *X*. The contamination of the silicon surface is removed by the RCA wafer cleaning process developed by Kern and Puotinen of RCA, USA, and it is widely used in the IC manufacturing companies.

References

- J D Plummer, M Deal and P B Griffin; *Silicon Fundamental Technology: Fundamentals, Practice and Modelling*, Prentice Hall, 2000
- S M Sze; *VLSI Technology*, Second Edition, McGraw-Hill, 1988
- S K Gandhi; *VLSI Fabrication Principles*, Second Edition, Wiley, 1994
- D Nagchoudhari; *Principles of Microelectronic Technology*, Wheeler, 1998
- S A Campbell; *The Science and Engineering of Microelectronic Fabrication*, Oxford University Press, 1996
- S M Sze; *Semiconductor Devices Physics and Technology*, John Wiley and Sons, 1985

Multiple-Choice Questions

2.1 Does the point defect increase with substrate doping?
(a) Yes (b) No

2.2 Does the silicon defect increase when the silicon is stressed?
(a) Yes (b) No

2.3 Volume defect is mostly found in
(a) float-zone technique (b) Czochralski technique

2.4 A silicon wafer produced by the float-zone technique has more defects than the Czochralski technique.
(a) Yes (b) No (c) Same

2.5 Generally, the purity of the metallurgical grade silicon is
(a) 98% (b) 99% (c) 99.9%

2.6 A silicon wafer produced by the float-zone technique is more resistive than the Czochralski technique.
(a) Possible (b) Not possible

2.7 The technique of wafer crystallisation that is more popular in the IC industry is
(a) Float-zone (b) Czochralski

2.8 Which silicon crystalline technique wafer contains more carbon and oxygen atoms?
(a) Float-zone technique (b) Czochralski technique

2.9 In clean-room classification, the reference particle size is taken as
(a) 0.5 micron (b) 5 microns (c) 10 microns

2.10 In RCA cleaning, the SC-1 solution is
(a) alkaline (b) acidic (c) neutral

Descriptive Problems

2.1 In IC fabrication, why is the Czochralski technique preferred over the float-zone technique?

2.2 Calculate how much phosphorous in grams is required to get 10^{15} atoms/cm^3 of boron doped silicon, if 100 kg of electronics grade silicon is loaded in the crucible to make crystalline silicon ingot by the Czochralski technique. The phosphorous segregation coefficient is 0.35, atomic weight is 30.97, Avogadro's number is 6.02×10^{23} atoms per mole and the molten silicon density is 2.53 g/cm^3.

2.3 A silicon wafer of 6" (125 mm) contains 0.886 mg (refer Example 2.1) of boron uniformly. Calculate the concentration in atoms per cm^3.

2.4 Elaborate why the term $K_L \dfrac{dT}{dx_l} A_l$ is approximated and neglected in the CZ recrystallisation technique?

2.5 Write RCA wafer cleaning steps with reasoning.

CHAPTER

3

MOS Transistor Process Flow

3.1 INTRODUCTION

The first bipolar transistor was realised in polycrystalline germanium by John Bardeen, Walter Brattain and William Shockley in 1947 at the BELL laboratories, USA. In the year 1935, Liandrat had revealed that the surface conductivity of a semiconductor changes when an electric field is applied perpendicular to it. His finding led to the birth of the unipolar Field Effect Transistor (FET). Thereafter, one of the FET family members called the **Metal-Oxide-Semiconductor Field Effect Transistor (MOSFET or MOS)** led to the creation of the integrated circuit (IC) in 1960 by Jack Kilby of Texas Instrumentation and Robert Noyce of Fairchild Semiconductor, USA. In this chapter, MOS and CMOS fabrication procedures have been focussed with the introduction of major technological evolution of transistor fabrication.

3.1.1 Transistor Fabrication by the Grown Junction Technique

In the growth technique, the *NPN* transistor is fabricated during the crystal growing stage (Chapter 2). The *N*-type silicon-crystal layer is drawn from the phosphorous added silicon melt and then boron is added to the silicon melt for a very short time to get the *P*-type silicon layer. Thereafter, phosphorous as dopant is added into the silicon melt and an *N*-type crystalline silicon is obtained. The process is optimised such that a thin layer of the *P*-type silicon is sandwiched between the *N*-type silicon. These three layers of the *N*, *P* and *N* silicon constitute the collector, base and emitter of a bipolar transistor. To make a large number of *NPN* transistors, the silicon crystal is cut into small pieces from the grown *NPN* transistor, as shown in Fig. 3.1. Similarly, the *PNP* transistor is made by this technique by adding the dopant in the reverse order.

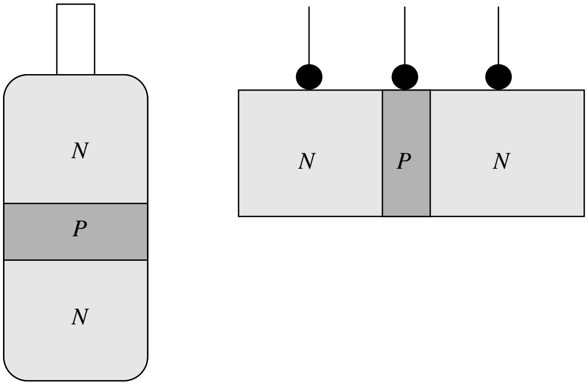


Fig. 3.1 Transistor fabrication by growth-junction technique

3.1.2 Transistor Fabrication by the Alloy Technique

In the same era as the junction-grown transistors, the alloy technique was also developed. In this technique, a small piece of the *P*-type indium is placed on each side of the *N*-type germanium semiconductor slice (wafer) and then heated for a short period of time. The indium pieces melt and diffuse inside the germanium wafer and the *NPN* transistor is formed. The process steps (usually called **process flow**) of the *NPN* transistor fabrication by alloy technique are shown in Fig. 3.2.

3.1.3 Transistor Fabrication by Double Diffusion Mesa Technology

In the late 1950s, transistor fabrication by double diffusion mesa technology was developed. The process flow (sequence) of this technique of transistor fabrication is shown in Fig. 3.3. The *N*-type silicon wafer is exposed to the boron gas environment at a high temperature. The boron atoms diffuse in the *N*-type silicon and then a thin layer of the *P*-type is formed on the silicon. Once the boron layer is formed, the wafer is subjected to high temperature in the gaseous environment of phosphorous. This converts to an upper layer of boron the phosphorous-rich layer. As the structure of the *PNP* transistor is in stack (mesa) form,

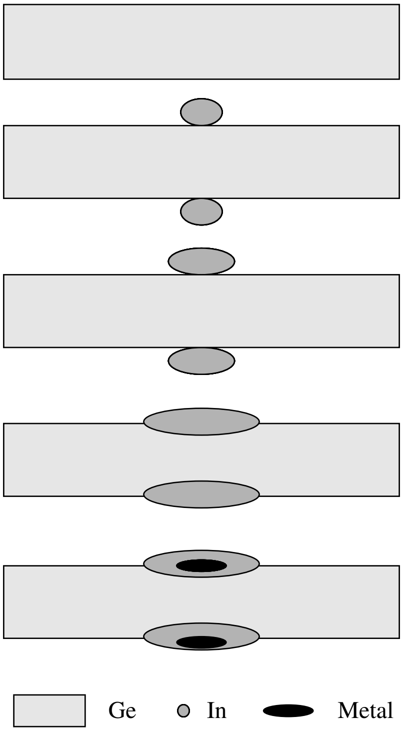


Fig. 3.2 Transistor fabrication by alloy-junction technique

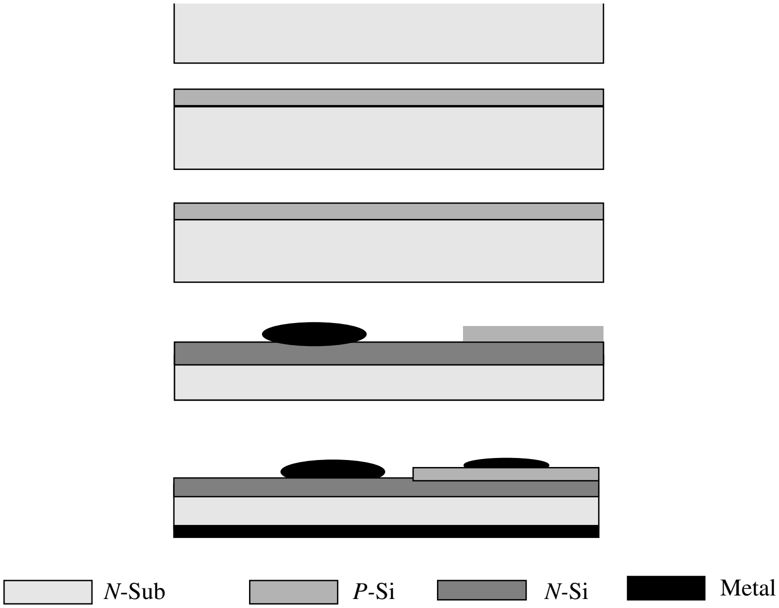


Fig. 3.3 Transistor fabrication by double-diffused mesa transistor technology

the technique is called the mesa technology. Once the *PNP* structure is made, the metal contacts are taken from two layers of doped silicon and one from the silicon wafer, as shown in Fig. 3.3. A large number of devices can be obtained from a single silicon wafer by this double diffusion technique.

3.1.4 Transistor Fabrication by Diffusion Technique (Planar Technology)

The fabrication of planar MOS transistor by the diffusion technique was developed in the 1950s by Jean Hoerni. The MOS transistor is fabricated by the diffusion process and it is described as follows: An *N*-type silicon wafer is heated in the presence of boron vapours in a furnace; at high temperature, the boron atoms diffuse inside the *N*-type silicon and convert the top layer of the *N*-type silicon into the *P*-type. Then, a thin silicon dioxide film (layer) is grown on the *P*-type silicon. Thereafter, the grown silicon dioxide film is removed from the localised areas and the wafer is heated in the presence of phosphorous vapours. The phosphorus atoms diffuse into the *P*-type silicon where no oxide is present, and convert the top *P*-type layer into *N*-type, resulting in an *NPN* transistor. Thereafter, metal connections are taken. The transistor fabrication process by diffusion technique is shown in Fig. 3.4.

By the diffusion technique, a large numbers of transistors are made on one side of the silicon wafer in layers (plane); therefore, this process is also called the **planar process**, as shown in Fig. 3.4. The present-day planar process is used to fabricate the MOS transistor.

3.2 MOS TRANSISTOR FABRICATION

MOS transistor physics is described through the energy-band diagram given in Chapter 1. In this section, the fabrication of a MOS transistor is described. Depending on the transistor's gate material, the MOS is classified as **metal-gate MOS** or **polysilicon gate MOS** or **polygate MOS**. In the metal-gate MOS transistor, a metal film is deposited on the gate oxide, and its process sequence (process flow) is shown in Fig. 3.5; whereas in the polygate MOS, a conducting polysilicon film is deposited, and over that, a metal film is deposited on the MOS gate.

3.2.1 Metal-Gate MOS Transistor Fabrication Process

In metal-gate MOS fabrication, the silicon wafer is first oxidised to a thickness of around 1 micrometre, and thereafter, silicon oxide is removed from the source and the drain areas,

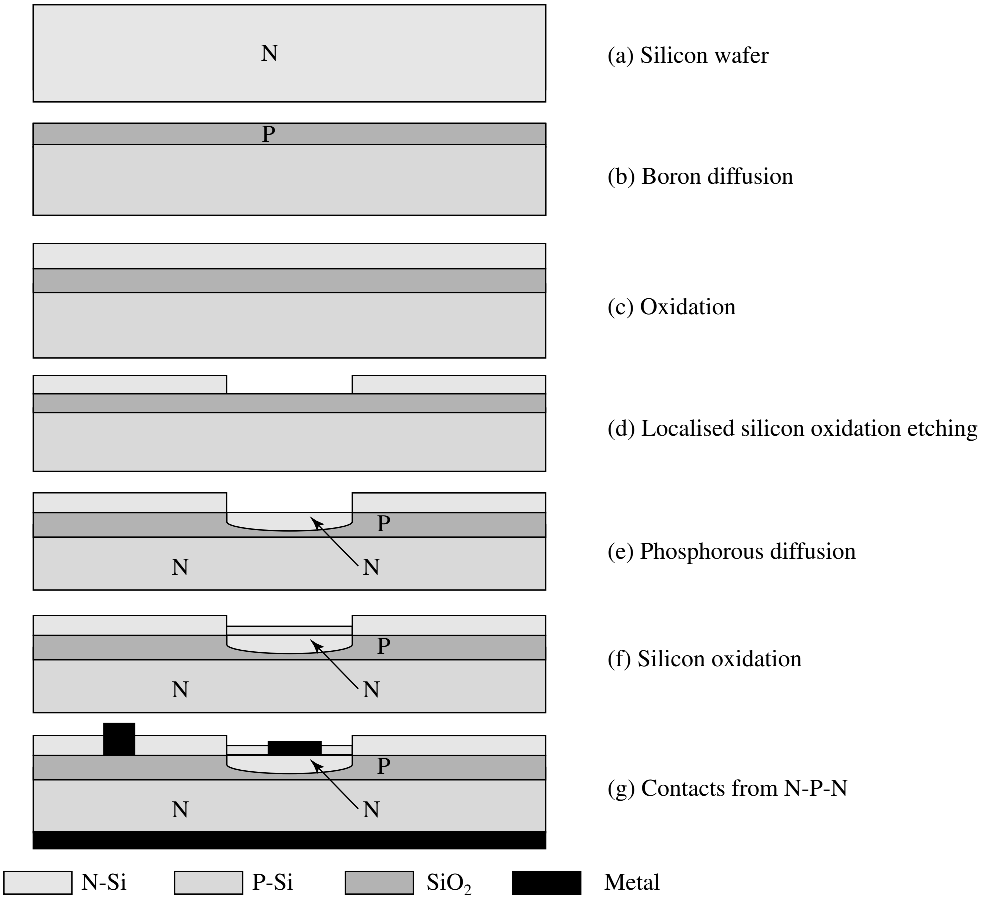


Fig. 3.4 Transistor fabrication by diffusion technique

as depicted in Figs. 3.5(a) and 3.5(b). Then, the source and the drain areas are diffused with phosphorous dopant and simultaneously, the wafer is oxidised as shown in Fig. 3.5(c). Thereafter, the thick oxide is etched between the source and the drain leaving behind the sides of the source and the drain for gate fabrication, as shown in Fig. 3.5(d). Then a high-quality oxide, called the **gate oxide,** is grown over the entire wafer, as shown in Fig. 3.5(e). Thereafter, the oxide is etched smaller than the source and drain areas, as shown in Fig. 3.5(f). These etched areas are called **contact areas** or **contact windows**. Once the oxide is removed from the contact areas, metal is deposited over the entire wafer and patterned, as shown in Fig. 3.5(g). It is to be noted that the metal gate is overlapping the source and the drain areas. This gate overlapping introduces gate-source capacitance at the source end and similarly, gate-drain capacitance at the drain end. The overlapping is required to take care of mask misalignments during lithography, etching and human error during MOS fabrication. As there are millions of transistors in an IC, a large capacitance appears in the

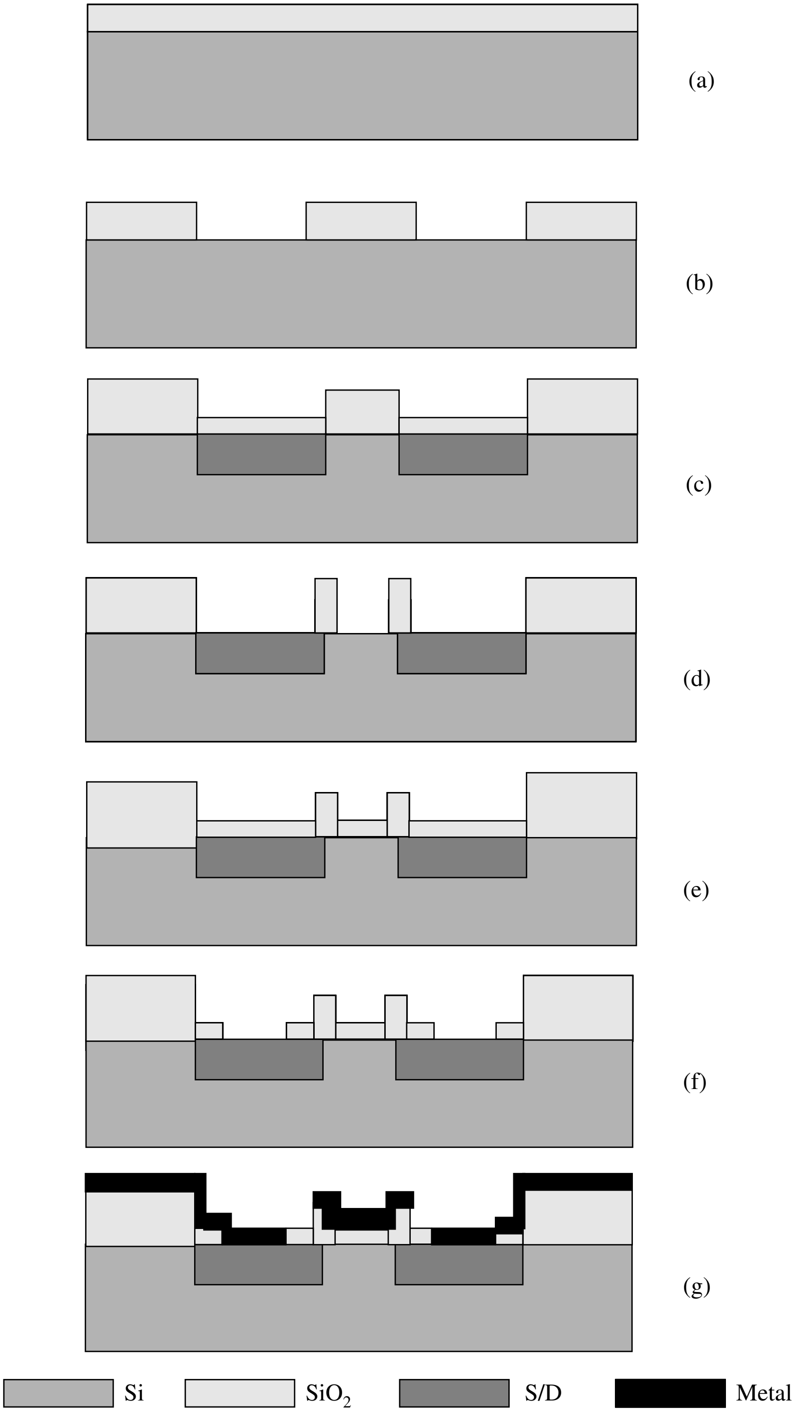


Fig. 3.5 Metal-gate PMOS transistor process

circuit and that reduces the speed (operating frequency) of the IC drastically. Furthermore, the metal-gate MOS transistor requires more silicon area due to the gate overlap. Hence, metal-gate MOS is not used for VLSI and ULSI applications. To avoid gate overlapping significantly, polysilicon gate is used in VLSI and ULSI applications.

3.2.2 Polysilicon-Gate MOS Transistors Fabrication Process

In polysilicon-gate MOS transistor fabrication, the polysilicon gate is made prior to the source and the drain areas. Thereafter, the source and the drain are made, as the source and the drain dopant diffuses almost vertically, and also diffuses marginally in the horizontal direction and leads to less overlap with the gate. As the dopant diffuses along the polysilicon gate and no extra alignment mask is needed, therefore, this technique of MOS fabrication is called the **self-align process**. In reality, at high temperature, the dopant diffuses laterally; this results in overlapping between the gate and the source and the drain areas and leads to the formation of an overlapping capacitor. The gate overlapping with the source and the drain is significantly less than that in the case of metal-gate MOS, thus resulting in the saving of silicon area. Other than these advantages, the wafer can be subjected to a high temperature as the polysilicon can withstand very high temperature for further processing; this is not possible with metal. The polygate thick oxide isolation is described in Section 3.3.1. The difference between metal and polygate transistors with regard to overlapping of the gate over the source and the drain is illustrated in Fig. 3.6.

An IC is made up of a large number of MOS transistors very closely spaced; hence, each MOS must be electrically isolated from all sides. Therefore, it is essential to describe the electrical isolations of MOS transistors.

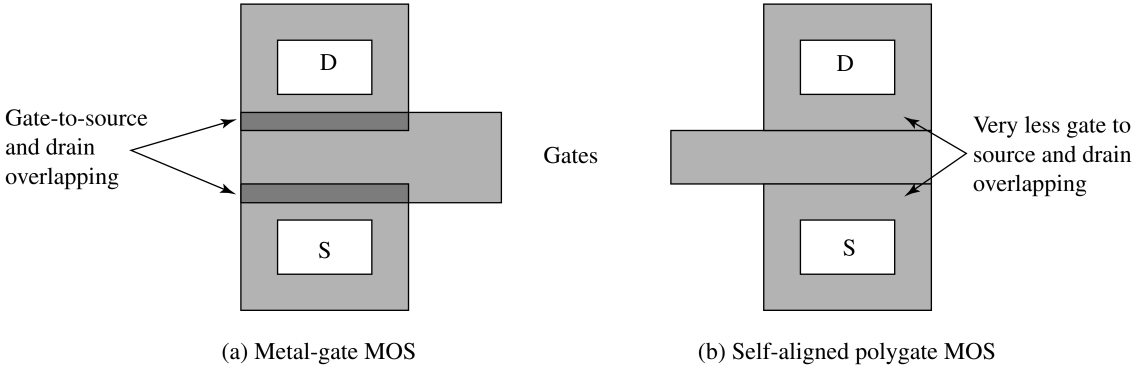


Fig. 3.6 Comparison of gate-to-source and drain overlapping in metal and polygates

3.3 DEVICE ISOLATION

To avoid electrical shorts between the closely placed MOS transistors, the electrical isolation of the MOS from all sides is a must. Usually, silicon dioxide dielectric is used for this purpose. Silicon dioxide has the advantage that it can be easily etched, besides being process compatible; and it can be grown or deposited easily as compared to other dielectric materials. There are many oxide-based isolation schemes, but the three most popular oxide isolation schemes, namely, thick oxide isolation, Local Oxidation of Silicon (LOCOS) isolation and shallow trench isolation are discussed below. To describe the process of device isolation, polysilicon-gate PMOS process is taken as the vehicle.

3.3.1 Thick-Oxide Isolated Silicon Gate PMOS Transistor Process

Thick-oxide isolated silicon-gate PMOS transistor process is shown in Fig. 3.7. The *N*-type silicon wafer is oxidised to a thickness of around 1 micrometre for MOS isolation and the active areas in the silicon are defined using active mask, as shown in Figs. 3.7(a) and 3.7(b) respectively. Then, the gate oxide is grown and the polysilicon film is deposited on the entire wafer as depicted in Fig. 3.7(c). Thereafter, the polysilicon is doped with phosphorous dopant to make it conducting. Then, the upper layer of the polysilicon is oxidised and thereafter, the polyoxide and polysilicon are removed excepting the MOS gate region using a gate mask as shown in Fig. 3.7(d). The source and the drain region are then doped with boron at high temperature and simultaneously, the oxide is grown as shown in Fig. 3.7(e). Then, the contact window is opened by removing the silicon oxide from the source and the drain areas and the polyoxides from the gate for electrical connection using contact mask as shown in Fig. 3.7(f). Thereafter, the entire wafer is metallised and the metal is patterned using a metal mask (electrical wiring) as shown in Fig. 3.7(g).

3.3.2 Local Oxidation of Silicon (LOCOS) Isolated Silicon Gate PMOS Transistor Process

In this section, LOCOS isolation polysilicon-gate PMOS process is described through Fig. 3.8. A thin layer of oxide, around 500 Å in thickness, is grown onto N-type wafer, as shown in Fig. 3.8(a) and then a nitride film is deposited over the grown oxide. Generally, the thickness of the nitride and oxide films is in the ratio of 3:1. The process of deposition of these oxide and nitride films on the wafer is shown in Fig. 3.8(b), and Fig. 3.8(c), respectively. Then, the nitride is removed from the wafer excepting the active areas, as shown in Fig. 3.8(d).

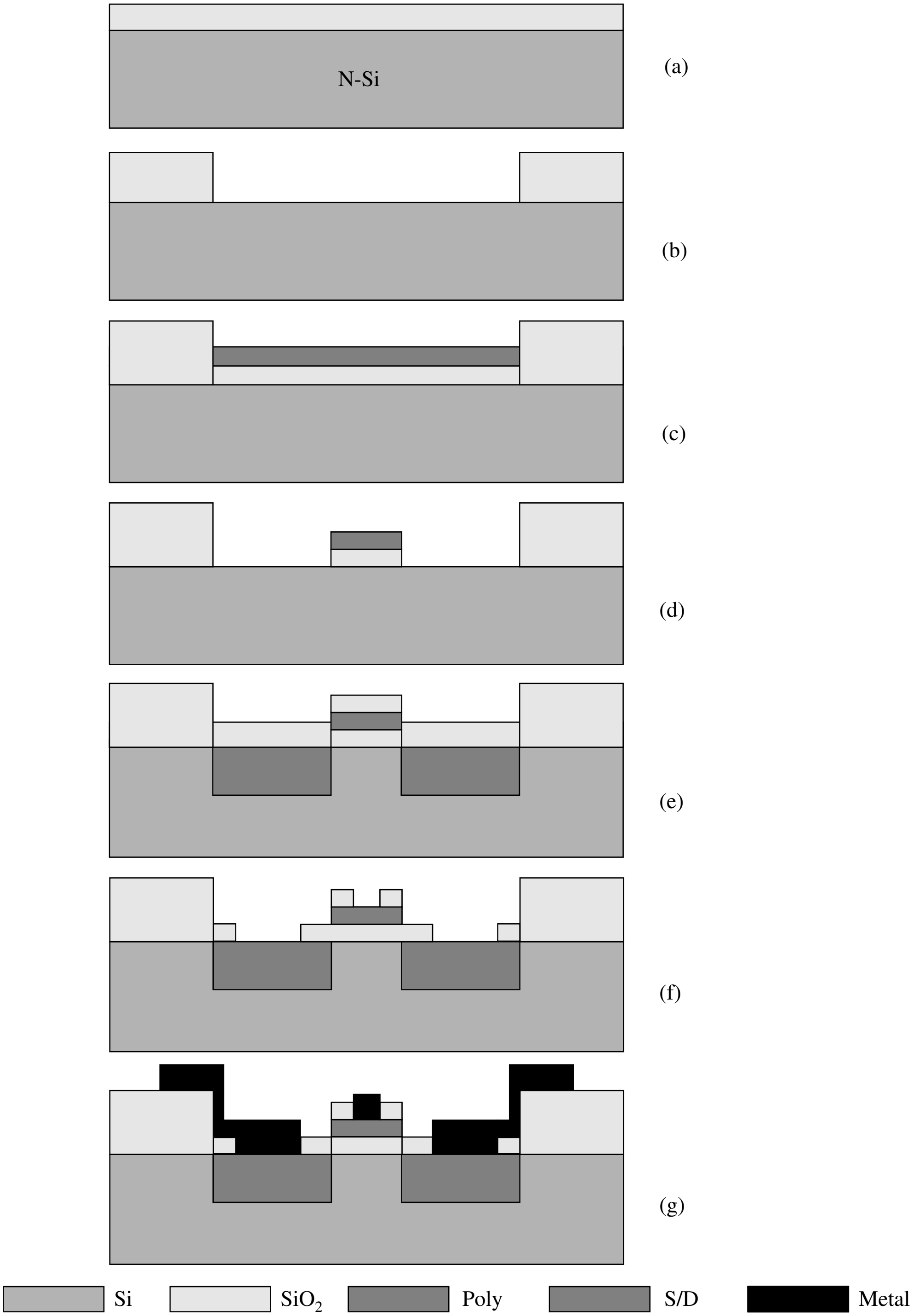
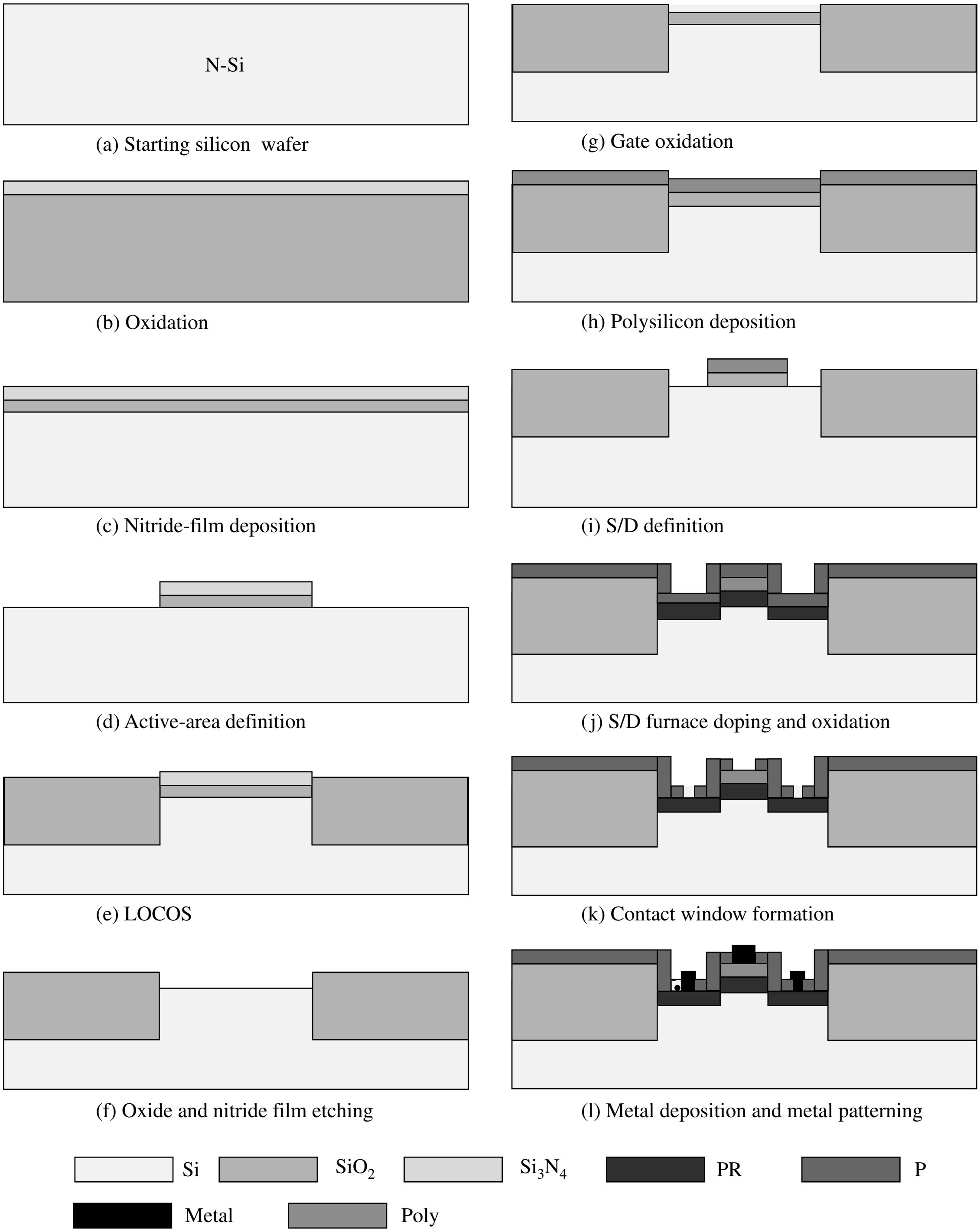


Fig. 3.7 Polygate PMOS transistor process



(a) Starting silicon wafer

(b) Oxidation

(c) Nitride-film deposition

(d) Active-area definition

(e) LOCOS

(f) Oxide and nitride film etching

(g) Gate oxidation

(h) Polysilicon deposition

(i) S/D definition

(j) S/D furnace doping and oxidation

(k) Contact window formation

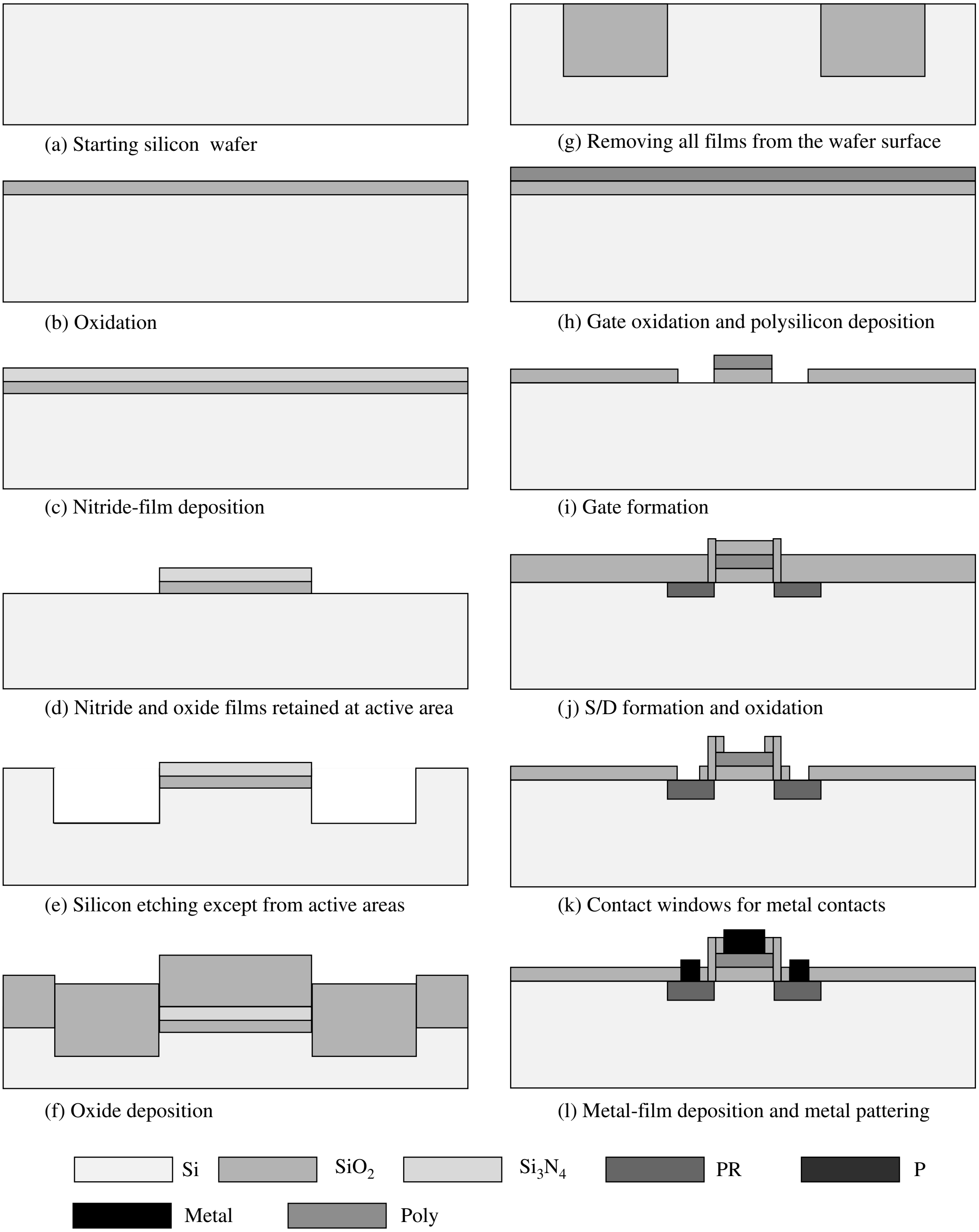
(l) Metal deposition and metal patterning

Fig. 3.8 LOCOS isolated silicon-gate PMOS transistor process

Thereafter, a thick oxide layer is grown by the oxidation process as shown in Fig. 3.8(e). Generally, the thick oxide is called **Field Oxide (FOX)**. Here, these FOX areas are used for MOS isolation. Oxygen atoms cannot diffuse through the nitride film; thus, no oxide is formed in the active areas. This eventually leads to the local oxidation of the silicon wafer. For this reason, this process is called **local oxidation of silicon (LOCOS)**. It is essential to mention that the FOX is not sharp vertical, but slanted inside the active area (not shown in figure). Once the local oxidation is completed, the oxide and the nitride films are etched out from the entire wafer as shown in Fig. 3.8(f). Then, a gate oxide is grown on the entire wafer as illustrated in Fig. 3.8(g). Then, the polysilicon film is deposited and doped with phosphorous, as illustrated in Fig. 3.8(h). The polysilicon is then etched from everywhere, except the MOS gate area, as shown in Fig. 3.8(i). Thereafter, boron is doped for the source and the drain and simultaneously oxidised as illustrated in Fig. 3.8(j). Then, the contact windows are opened in the source, the drain and the gate areas. Thereafter, a metal film is deposited on the entire wafer and as depicted in Fig. 3.8(k) then metal is patterned for electrical connections, as shown in Fig. 3.8(l).

3.3.3 Shallow-Trench Isolated PMOS Process

The shallow-trench device isolation scheme is described through Fig. 3.9. The shallow-trench MOS isolation scheme has been found to be the best option for VLSI and ULSI applications till today. The wafer is first oxidised and then, a nitride film is deposited on the oxide film, as shown in Fig. 3.9(a) to Fig. 3.9(c). Then, the oxide and the nitride films are removed from these boundaries of the active areas and silicon is etched from these boundaries, as shown in Fig. 3.9(d) and Fig. 3.9(e), respectively. Generally, shallow silicon etching is done, but it should be deep enough to isolate the MOS transistor. As the silicon etching is shallow and vertical, this isolation technique is called **shallow trench**. Once the trenches are formed, a thick oxide is deposited on the wafer till the trenches are covered as shown in Fig. 3.9(f). Thereafter, the oxide over the wafer surface is removed by the **Chemical and Mechanical Polishing Process (CMP)**, as shown in Fig. 3.9(g). Once the oxide above the wafer surface is removed, the active areas are enclosed inside the trench. Then, the gate oxide is grown over the wafer and thereafter, the polysilicon and the deposited and doped with phosphoros on the gate oxide as shown in Fig. 3.9(h). Thereafter, the oxide is removed from the active area except the gate region as shown in Fig. 3.9(i). Once the gate is defined, the source and the drain diffusion, and oxidation are carried out; as shown in Fig. 3.9(j). Thereafter, the contact windows are opened in the source, the drain and the gate regions as shown in Fig. 3.9(k). Then, the global metal film is deposited on the wafer and metal patterns are made as shown in Fig. 3.9(l).



(a) Starting silicon wafer

(b) Oxidation

(c) Nitride-film deposition

(d) Nitride and oxide films retained at active area

(e) Silicon etching except from active areas

(f) Oxide deposition

(g) Removing all films from the wafer surface

(h) Gate oxidation and polysilicon deposition

(i) Gate formation

(j) S/D formation and oxidation

(k) Contact windows for metal contacts

(l) Metal-film deposition and metal pattering

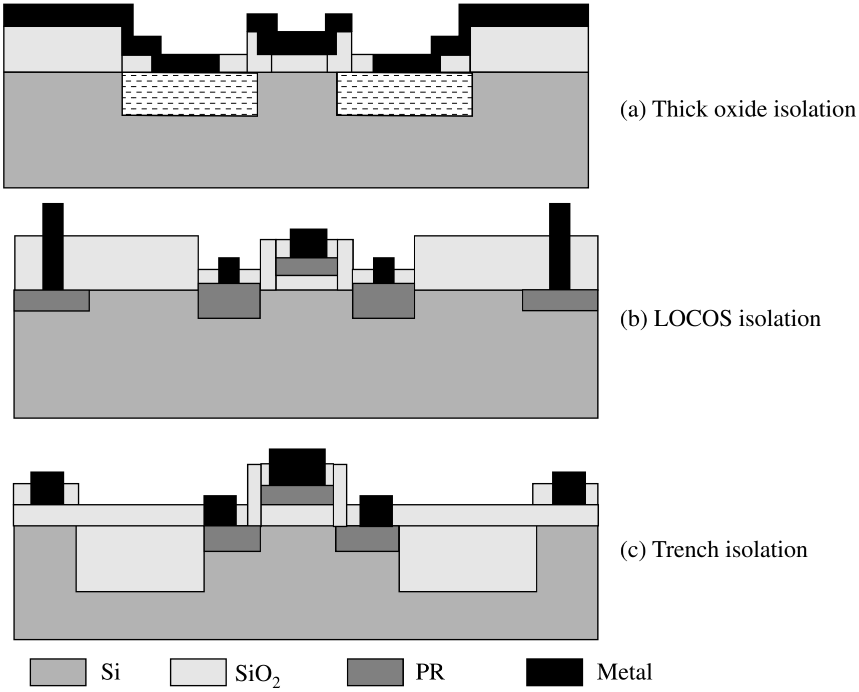
Fig. 3.9 Trench isolated silicon-gate PMOS transistor process

3.3.4 Comments on the Thick Oxide, LOCOS and Shallow-Trench Isolation Schemes

The above schemes of transistor isolation are essential to circumvent the electrical shorts between the MOS transistors. Unfortunately, all of these schemes suffer from some disadvantages. Shortcomings and strengths of these isolation schemes are discussed below with illustrations.

Thick Oxide Isolation

The thick oxide isolation scheme is much simpler than the LOCOS and the shallow-trench isolation schemes, but it suffers from the high vertical oxide step at the edge of the FOX and the active areas. This sharp vertical oxide step creates problems for subsequent MOS fabrication steps, especially the film deposition and lithography processes. In addition, the silicon area is more consumed than the other two schemes. Furthermore, in the thick oxide isolation scheme, the transistors are isolated from the top of the wafer surface, but not from the bottom of the silicon wafer. For these reasons, the thick oxide isolation scheme is not opted for VLSI and ULSI applications. The structure oxide isolation is shown in Fig. 3.10(a).


(a) Thick oxide isolation

(b) LOCOS isolation

(c) Trench isolation

Fig. 3.10 Structure of thick oxide isolation, LOCOS isolation, Trench isolation

Local Oxidation of Silicon (LOCOS)

In the LOCOS isolation scheme, the MOS transistors are isolated (separated) from the top as well as the bottom of the wafer to some extent. The LOCOS oxide has a gentle slope between the FOX and the active areas with less oxide height that minimises the problems of subsequent film deposition and lithography processes. Unfortunately, the LOCOS scheme suffers from the loss of silicon due to a "bird beak" formation. The bird beak is formed due to the lateral diffusion of oxygen atoms inside the active regions, resulting in the oxidation of the silicon not shown in Fig. 3.10(b). In addition, the LOCOS isolation process needs extra process steps compared to the thick oxide isolation process.

Shallow-Trench Isolation

Shallow-trench isolation is the best isolation scheme till today. Hence, this shallow-trench isolation scheme is presently opted for VLSI and ULSI applications. The trench isolation scheme isolates the MOS transistor from deep inside the wafer. Furthermore, the trench isolation is vertical; hence, there is least loss of silicon. Trench isolation needs extra processing steps as compared to thick film and LOCOS isolation schemes. In addition, highly sophisticated machines are needed to carry out the trench isolation process. The structure of shallow trench isolation is depicted in Fig. 3.10(c).

3.4 CMOS FABRICATION

3.4.1 Introduction

The circuit made of NMOS transistors or PMOS transistors in separate chips is called IC, i.e. **integrated circuit**, but the circuit made of NMOS transistors and its complementary PMOS transistors in a single chip is called the **Complementary Metal-Oxide-Semiconductor (CMOS) integrated circuit** and is abbreviated as **CMOS IC**; and the process by which these two transistors are made in a single chip is called the **CMOS process**. The two-dimensional structures of CMOS, NMOS and PMOS transistors are shown in Fig. 1.4. The fabrication of CMOS is complex and needs around 20% extra processing than required for the individual NMOS or PMOS. The CMOS IC has many advantages over the individual NMOS and PMOS integrated circuits out of which some of the advantages worth mentioning are low power consumption, wide power supply range, higher speed, and wide output swing voltages. Nowadays, almost all of the ICs are made by the CMOS process; hence, the CMOS fabrication technology and its related technological issues are discussed here in detail. Generally, a CMOS IC is also referred to as an IC.

3.4.2 CMOS Process

In the previous section, the basic MOS structure and its fabrication sequence is described. To make NMOS and PMOS in a single wafer, more precisely in one chip, both the *P* and *N* types of substrates are needed. This can be achieved by the boron and phosphorous dopings at different areas in the chip on which these MOS transistors are to be made. The area which is doped by boron is called the ***P*-well** and the area which is doped by phosphorous is called the ***N*-well**. In many literatures, these wells are also named as **tubs**. In this section, LOCOS isolated CMOS fabrication process is described in detail, with illustrations; and for the sake of better visualisation, the dimensions of the transistor layers (structures) are not made to scale in the illustrated figures. Also, the typical fabrication processing data, such as temperature, time and other process parameters are not the actual values pertaining to the generation of IC. In fact, these process parameters are called the **recipe**, and these recipes are governed by the electrical parameters and the MOS geometry. The individual fabrication processes, for example oxidation, diffusion are called the **unit processes**. When these unit processes are integrated (sequentially processed) to realise the CMOS IC (or IC), it is called **process integration**. The process integration recipe may be slightly different than the individually optimised unit process recipe. This can be explained through an example: say the source and the drain have 1-micrometre junction depth. The junction depth is optimised by the diffusion unit process, but during MOS fabrication, the wafer passes through many high-temperature processes, for instance, oxidation. The dopants go deep into the silicon that increases the junction depth more than 1 micrometre. To start with the MOS process, wafer selection (specifications) is important. These wafer specifications are mentioned below.

Wafer Selection

The electrical characteristics of an IC significantly depend on the silicon-wafer specifications. These specifications are mainly, the type of wafer (*N* or *P*), resistivity (doping level), crystal orientation (mostly 100), wafer size (on wafer process line), wafer flatness, defect density, impurity level, etc. To demonstrate the CMOS process flow, a wafer is taken with specifications such as *P*-type, surface orientation (100), resistivity 5 to 50 Ωcm or conductivity from 10^{15} to 10^{16} $(\Omega\text{cm})^{-1}$ and depicted in Fig. 3.11.

Fig. 3.11 Silicon wafer

1st step: Oxidation To fabricate the CMOS (for that matter any type of MOS), the silicon wafer is first oxidised in the oxidation furnace. In the oxidation furnace, the N_2 gas is introduced at around 1 L/min, till the oxidation temperature reaches say around 1100°C. The N_2 gas flushes out the impurities (if any) from of the furnace, and also keeps the furnace in a positive pressure to prevent impurities from entering. Then, the wafers are loaded in a quartz jig called the "boat" and kept at the mouth of the furnace for 5 minutes for warming. Thereafter, the boat is pushed inside the central zone of the furnace for 10 minutes; where, the central zone is maintained at around 1100°C ± 1°C temperature. The slow push of the boat in the high-temperature central zone avoids the generation of dislocations and cracks in the wafer due to the sudden change of temperature. This entire process is called **wafer loading**. After the wafer is loaded inside the furnace, the N_2 gas is closed and O_2 gas is introduced at around 1 L/min. At a high temperature, the oxygen atoms react with the silicon atoms and silicon dioxide is formed. After 10 minutes of oxidation, O_2 gas is stopped and N_2 gas is passed at the rate of 1 L/min; the wafer is kept there for about 10 minutes. With this dry oxidation process, the oxide thickness comes out around 500 Å. The process of heating of the wafers in nitrogen gas is called the **annealing process**. The annealing process reduces the oxide and oxide/silicon interface charges. Thereafter, the boat is withdrawn from the central zone of the furnace to the mouth of the furnace in around 10 minutes. The boat and the wafer are left for 5 minutes in the furnace mouth for cooling and thereafter, the wafers are unloaded from the boat to the wafer-carrying container. The process of withdrawing the jig from the centre of the furnace till the unloading of wafer from the boat is called **wafer unloading**. The oxidised silicon wafer is shown in Fig. 3.12. The details of silicon oxidation are covered under oxidation in Chapter 4.

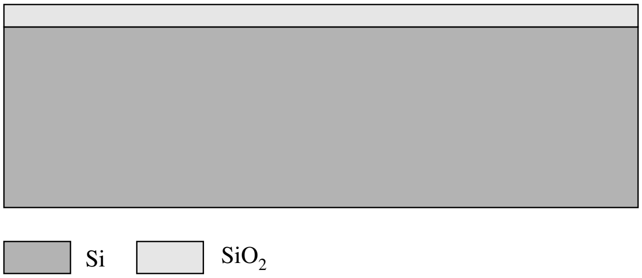


Fig. 3.12 Oxidation

2nd step: Nitride Deposition In the second step, a thin nitride film is deposited over the grown silicon dioxide. Generally, in order to reduce stress, the thickness of the nitride film is kept around three times that of the previously grown silicon dioxide. Generally, the nitride film is deposited using the **Low Pressure Chemical Vapour Deposition (LPCVD)** technique. In the LPCVD technique, a nitride film is deposited in the furnace with around 5 mTorr pressure at about 800°C temperature in the presence of silane and ammonia gases. The furnace of LPCVD is called the **reactor**. The wafer is loaded in the quartz boat and then the boat is pushed into the centre of the furnace, as explained in the oxidation process, Step 1. Thereafter, the reactor is evacuated and the silane (SiH_4) and the ammonia (NH_3) gases are introduced. These gases react with themselves at high temperature and produce silicon nitride vapours that get deposited on the wafer. The chemical reaction is mentioned below.

$$3SiH_4 + 4NH_3 \rightarrow Si_3N_4 + 12H_2.$$

After the silicon nitride film deposition, the silane and ammonia gases are closed and nitrogen gas is introduced to bring back the furnace at atmospheric pressure and anneal the wafer. Thereafter, the wafer is unloaded in a similar way as explained in the oxidation process, Step 1. The details of LPCVD are covered in Chapter 9. Figure 3.13 shows the nitride-deposited film on the wafer.

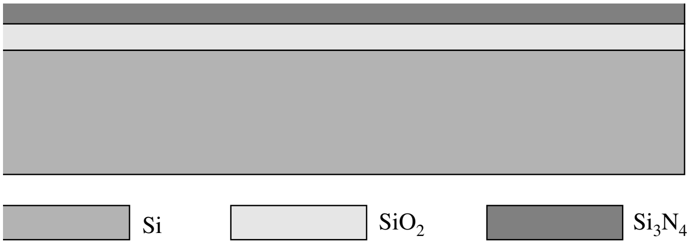


Fig. 3.13 Nitride deposition

3rd step: Active Regions Definition After the nitride deposition, the *N* and *P* well areas are created in the wafer. These wells are defined on the wafer by replicating the active mask using the processes of lithography and etching. The wafer is held on a spinner machine, sucked by vacuum and then the PR (say positive PR), is spread all over the wafer. After that, the wafer is rotated at around 6000 rpm for 1 minute. At this high speed, around 1 micrometre thick PR is left over the wafer, and the excess PR is thrown out off the wafer due to centrifugal force. For better PR adhesion, the wafer is coated with liquid Hexa-Methyl-Di-Silane (HMDS) prior to coating the PR, as shown in Fig. 3.14. The process of HMDS coating is similar to PR coating. Thereafter, the wafer is placed inside the oven at 90°C for 45 minutes for PR hardening. This process step is called the **PR pre-bake** or simply **pre-bake**.

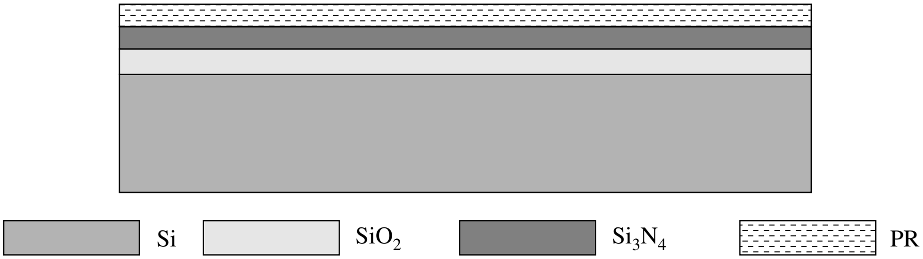


Fig. 3.14 PR coating

Then, the active mask and the wafer are placed in the mask alignment equipment and brought closer around 2 μm. Then, looking through the microscope, one of the scribe tracks of the mask is aligned with the primary cut of the wafer. Once the alignment is done, the wafer and the mask are brought in close contact (one of the modes of lithography), and the PR is exposed through the mask using UV light for a few seconds, as shown in Fig. 3.15. Thereafter, the exposed wafer is taken out from the mask alignment machine and developed in the PR developer solution for 1 minute. The exposed PR (particularly positive PR) dissolves in the PR developer, exposing the nitride film, as shown in Fig. 3.16. Thereafter, the wafer is rinsed thoroughly in de-ionised (DI) water and then heated in the oven at 120°C for 45 minutes for further PR hardening and adhesion. This process step is called the **PR post-bake** or simply **post-bake**. The details of lithography are described in Chapter 6.

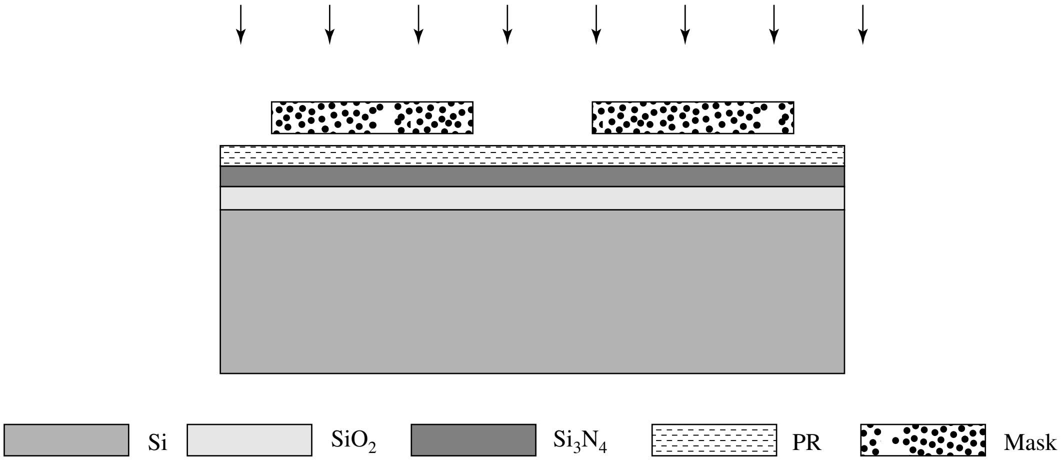


Fig. 3.15 Active areas lithography

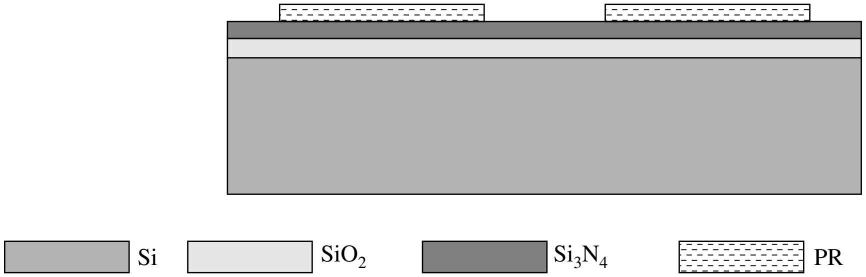


Fig. 3.16 After PR development

4th step: Nitride Etching In the next step of CMOS process, the nitride and oxide films are etched out from the exposed (FOX) areas from where PR has been removed, as shown in Fig. 3.17. Generally, the nitride film is etched by the plasma etching technique. In the

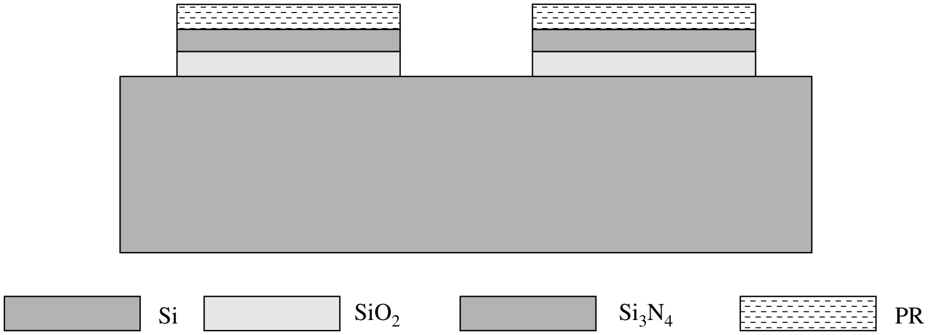


Fig. 3.17 Nitride etching

plasma etching technique, the wafer is placed on one of the electrodes of a partially evacuated discharged chamber. Usually, a high RF voltage at 13.56 MHz frequency is used to create the electrical discharge (plasma) between the electrodes. In the discharge chamber, a particular gas (etchant gas) is introduced. The etchant gas dissociates and gets converted into highly reactive atoms (or molecules) in the electric discharge (plasma) condition. For example, the NF_3 (or CF_4) gas dissociates in the electric discharge (plasma) condition and gets converted into highly reactive fluorine atoms. The fluorine atoms react with the exposed silicon nitride film and produce volatile products. On the other hand, the fluorine atoms do not etch the nitride film from other places (active areas) as it is protected by the PR film. The dissociation of NF_3 in plasma condition can be written as

$$4\ NF_3 \rightarrow 12\ F + 2\ N_2$$

$$Si_3N_4 + 12\ F \rightarrow 3\ SiF_4 \text{ (volatile product)}$$

The plasma etching process is always carried out by a gas or a mixture of gases in the plasma (electric discharge); therefore, this process is called the **plasma etching** or the **dry etching** process. Apart from plasma etching, the silicon nitride film can also be etched by a liquid chemical, but it is not preferred due to the high etching temperature required. Etching of the film by a chemical is called **wet etching** and the etching chemical (or mixture of chemicals) is called the **etchant**. Details of nitride etching are covered in Chapter 7.

Film etching has two important criteria, namely selectivity and anisotropicity. The ratio of the etch rates of different materials is defined as **selectivity**. As the IC is made of different film materials, the selective etching of film is very essential. In ideal film etching, one film should be etched without getting the other films or materials etched. Unfortunately, dry etching has poor selectivity than wet etching. Apart from film selectivity, another important criteria is that the film should be etched anisotropically (vertically). The anisotropic etching of the film helps to maintain the fidelity of the dimensions of the mask patterns on the wafer, as also the MOS transistor dimensions. As the MOS dimensions are reducing day by day, the anisotropic (vertical) etching is becoming indispensable. A good anisotropic etching is obtained by the dry etching process as compared to the wet etching process. This is one of the reasons why dry etching is always preferred over wet etching for VLSI and ULSI fabrication. Furthermore, PR as a mask can be used for the dry etching process for almost all types of film etching, but it is not true in the case of wet etching. In case of wet etching, at high temperature, the PR either comes out (lift off) from the wafer or reacts with the etching chemical, especially in the nitride etching process. Wet etching process is preferred when the nitride film is globally removed from the wafer. Once the nitride film is etched from the FOX areas, the remaining PR is removed using acetone, as shown in Fig. 3.18. The PR can also be removed using sulphuric acid (or plasma in O_2 ambient).

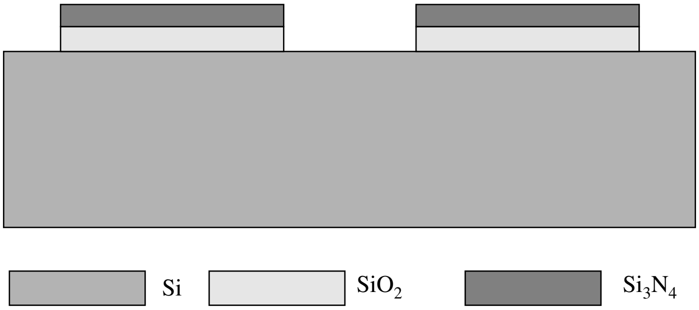


Fig. 3.18 After PR removal

5th step: Field Implant Once the nitride film is removed from the field areas, the wafer is processed for field implantation, as shown in Fig. 3.19. Field implantation reduces the chances of forming a parasitic transistor under the field oxidation (FOX) area. For this reason, this process step is called **field implantation**. In the field ion-implantation process, the boron dopant ions are produced in the plasma and accelerated to a great speed by the electrostatic field, and then they are made to lodge in the wafer. The details of ion implantation are covered in Chapter 9. Generally, the field implantation is carried out at low implantation energy (~30 keV) with low boron dose ($\sim 1 \times 10^{13}$ atoms per unit area) for marginal increase of the dopant concentration just below the wafer surface. In practice, the ion implantation is done through a thin oxide film. It has the advantage of wafer protection from environment contamination as well as it minimises the channelling effect. The majority of the ions stop near the surface, but a few ions go deep inside the wafer due to the **channelling effect**. The channelling effect occurs because the implanted ions enter in the silicon lattice gaps and travel to a long distance before they stop. The presence of

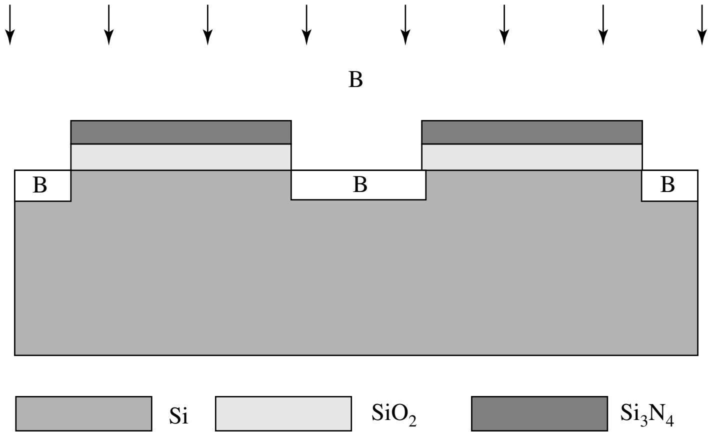


Fig. 3.19 Field implant

oxide on the wafer drastically reduces this channelling phenomenon, as the dopant ions collide first with the amorphous oxide atoms, and get scattered in all possible directions. This reduces the probability of the ions entering into the lattice gap of the silicon. The details of the channelling effect are covered under ion implantation in Chapter 9. Once field implantation is completed, the PR is removed (stripped) completely from the wafer.

6th step: Local Oxidation (LOCOS) After field implantation, the wafer is processed for local oxidation (LOCOS). To get a fast and thicker oxide, the dry-wet-dry oxidation recipe is used. The wafer is loaded in the oxidation furnace and then dry oxidation is carried out at around 1100°C for 10 minutes to get a good-quality oxide film; this also helps in a better PR adhesion on the oxide film. Thereafter, the oxidation process is carried out in the presence of water vapour for fast oxidation. This particular process step is called **wet oxidation**. In the wet oxidation process, the wafer is heated in the presence of water vapour for around 230 minutes. Once the wet oxidation is completed, dry oxidation is again carried out for 10 minutes to get a good oxide film and better oxide/silicon interface. This combination of the oxidation process is called the **dry-wet-dry oxidation** process. With this oxidation recipe, the oxide film comes out to be around 1 µm thick. After LOCOS is done, the nitride film is removed from the wafer, as shown in Fig. 3.20.

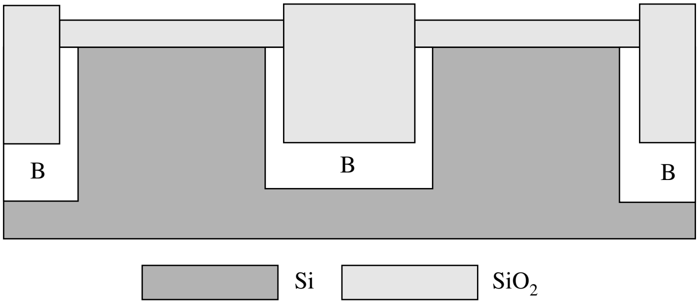


Fig. 3.20 Nitride film is removed

7th step: *N*- and *P*-wells In this step, the nitride is removed from the active areas for well formation. Depending on the circuit layout, the *N*-well and the *P*-well are fabricated in different locations of the die (chip). These *N*- and *P*-wells are fabricated in two steps. In the first step, the *N*-well is made using the ***N*-well masks** in conjunction with the lithography and oxide etching processes, and thereafter, the implantation of phosphorous ions is done. In the second step, the *P*-well is made in the rest of the active areas using the ***P*-well masks** in conjunction with the lithography and oxide etching processes, and thereafter, the implantation of boron ions is done. The order of well formation can be altered. Process flows of these well formations are shown in Figs. 3.21 and 3.22.

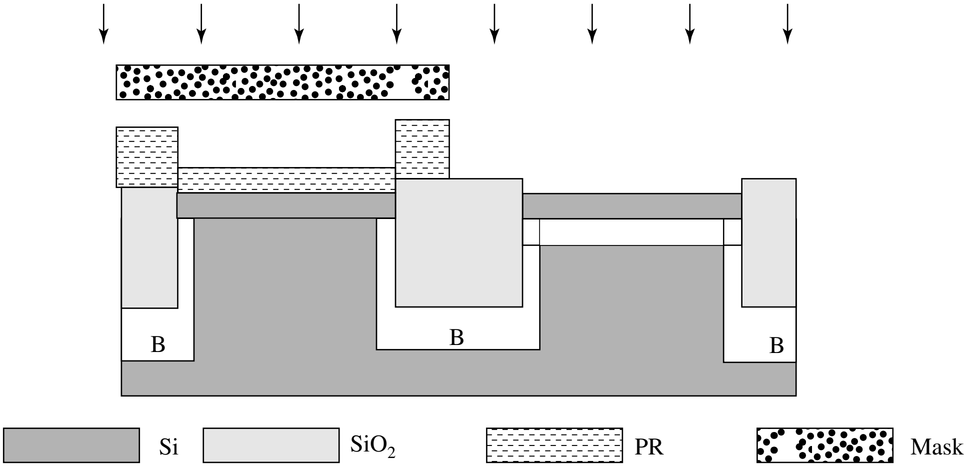


Fig. 3.21 *P*-well implant

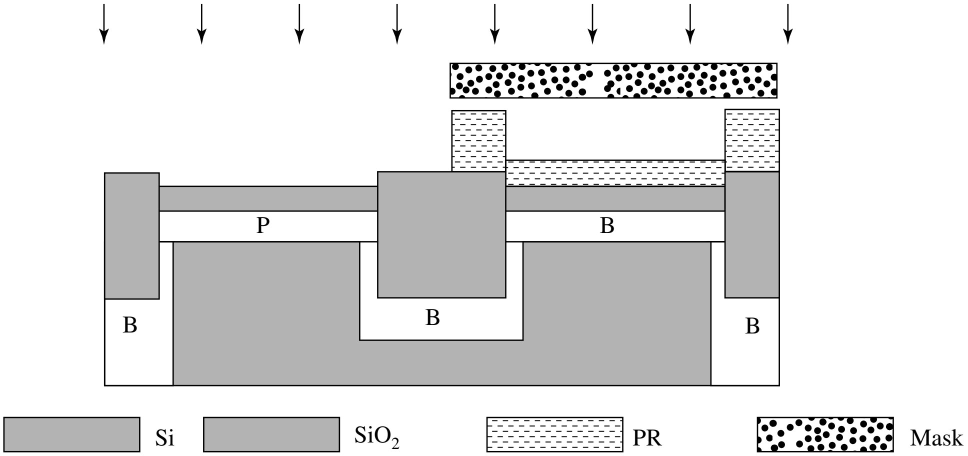


Fig. 3.22 *N*-well implant

Generally, the implantation doses for the wells are chosen in the range of 5 × 10^{16} to 1 × 10^{17} atoms cm^{-2}. These well dopings serve as substrates for the MOS transistors; hence, the implantation doses are chosen with great care. These well doping concentrations should be somewhat less than the source and drain dopant concentrations; on the other hand, it should be more than the starting silicon wafer doping concentration.

After implantation of the wells, the wafer is put in a high-temperature furnace having temperature around 1150°C, in the presence of nitrogen gas for about 6–8 hours, to push (diffuse/drive) the dopants 5–6 micrometre deep inside the wafer, if the source and drain junction are around 1 micrometre. This process is called the **drive-in process**. It is essential to mention that the well concentration decreases when the drive-in is done. The detail of the diffusion and the drive-in process is covered in Chapter 8. It is essential to mention that the diffusion constant (diffusivity) of boron and phosphorous is almost equal; hence, the junction depths of the *N*- and *P*-wells are also nearly the same, as shown in Fig. 3.23.

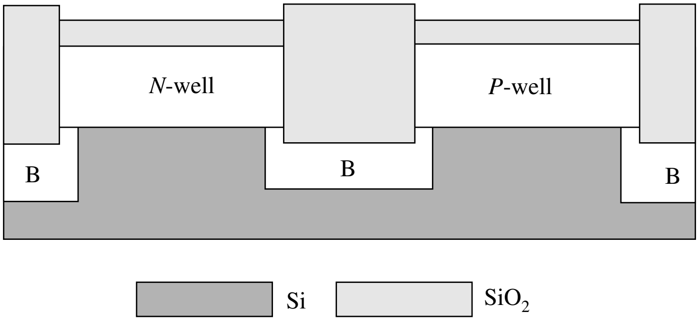


Fig. 3.23 Well drive-in

8th step: Transistor Threshold (V_{th}) Adjustment It is essential that the threshold voltages of both the transistors are the same or within a tolerable limit. The V_{th} of the MOS transistor is a function of the well (substrate) doping concentration. Unfortunately, an appropriate well doping may not be achieved during well formation; hence, well-doping concentrations are tailored according to the desired threshold voltage of the MOS transistors. The process of tailoring of well concentration for the required threshold voltage is called **threshold voltage adjustment**.

The threshold voltage of a MOS transistor is expressed as

$$V_{\text{th}} = V_{FB} + 2\phi_f + \sqrt{\frac{(2\varepsilon_s q N_A 2\phi_f)}{C_{\text{OX}}}} + q\frac{Q_{\text{imp}}}{C_{\text{OX}}} \tag{1}$$

where V_{FB} is the flat band voltage due to the work function, $\phi_f$is the position of the Fermi level in the bulk with respect to the intrinsic level, ε_s is the permittivity of silicon, N_{A} is the substrate dopant concentration, i.e. the dopant concentration of the well, Q_{imp} is the required implanted dose to adjust the V_{th}, C_{ox} is the oxide thickness and q is the electronic charge.

For the MOS threshold voltage for a particular MOS, the implantation dose QI_{imp} is calculated. Threshold voltage adjustments may be done using the *N*-well and *P*-well masks; hence, no extra masks are needed. The threshold voltage implantation process steps for the NMOS and PMOS transistors are shown in Figs. 3.24 and 3.25 respectively. Generally, the threshold voltage implantation dose and the energy for the NMOS and PMOS transistors are in the range of 1–5 × 10^{12} atoms cm^{-2} and 50–75 keV respectively.

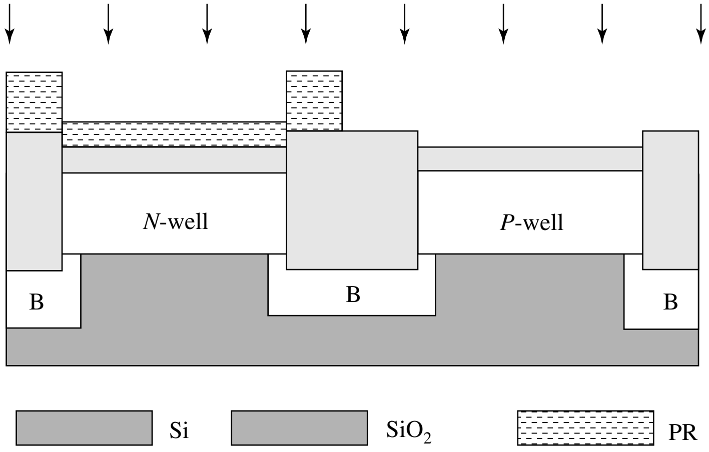


Fig. 3.24 NMOS threshold adjustment

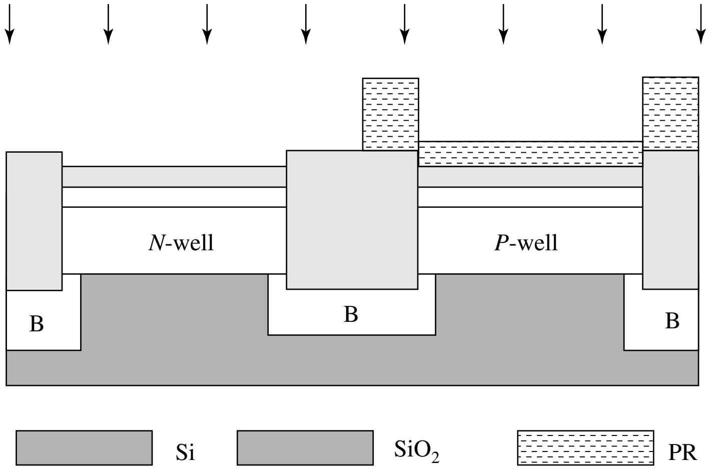


Fig. 3.25 PMOS threshold adjustment

9th step: Gate Oxide Growth Once the threshold voltage is adjusted, the wafer is processed for gate oxidation. The gate oxidation process is one of the most critical steps in MOS transistor fabrication. The gate oxide should have minimum oxide charges as well as oxide/silicon interface charges. Prior to gate oxidation, the PR and the oxide over the well are completely removed from the wafer and then the wafer is thoroughly cleaned. Then, the wafer is loaded into the gate oxidation furnace and dry oxidation is carried out in the presence of oxygen and chlorinated gas. The chlorine gas reacts with the impurities such as sodium, potassium, etc. and reduces the oxide charges. The gate oxide grown in this way is shown in Fig. 3.26. The details of oxidation are described in Chapters 4 and 11.

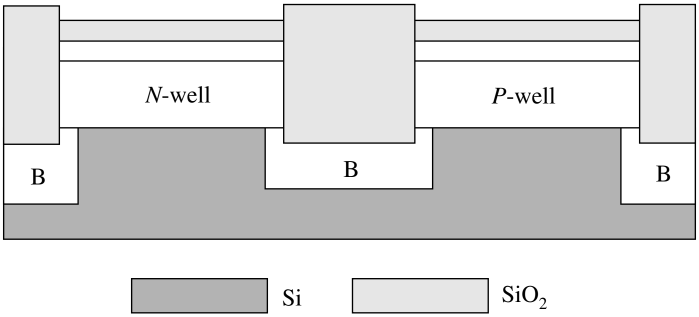


Fig. 3.26 Gate oxide

10th step: Polysilicon Deposition and Gate Electrode Definition To make the polysilicon gate transistor, a polysilicon film of about 1 µm thickness is deposited on the gate oxide by the LPCVD technique. Polysilicon deposition is done in a furnace (reactor) that is similar to the nitride reactor. Generally, silane gas is used as a precursor to polysilicon deposition. Polysilicon is deposited at a low pressure (~5 mTorr) at a temperature of about 650°C temperature. At this temperature, the silane gas decomposes into vapour of silicon atoms and hydrogen gas, as mentioned below:

$$SiH_4 \rightarrow Si + 2H_2$$

The silicon vapours deposit on the wafer in a polycrystalline (partially crystalline) form. The partially evacuated reactor allows the silicon atoms to move all around the reactor, resulting in uniform film deposition on the wafer. The polysilicon deposited wafer is shown in Fig. 3.27. In this figure and onwards the threshold voltage implantation is not shown. As the deposited polysilicon has very high resistivity, it is heavily doped with phosphorus to make the polygate conductive. Due to this, the gate electric field becomes uniform across the gate after voltage is applied to the gate. Usually, two techniques are used for poly doping, namely, **diffusion** and **in-situ doping**. In the diffusion technique,

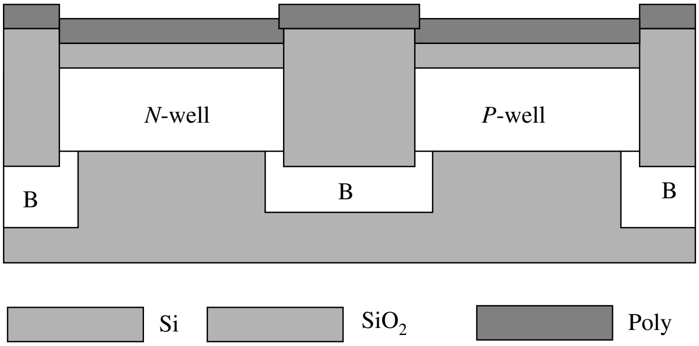


Fig. 3.27 Poly deposition and poly doping

poly is doped with $POCl_3$ vapour at 950°C for 15 minutes after polydeposition; whereas, in the in-situ doping technique, phosphorous is doped during the polydeposition. In spite of heavy phosphorous doping, the sheet resistivity of polysilicon is much higher (~100 times) than that of an aluminium metal film. For this reason, the doped poly is used only for MOS transistor gates and short-distance (local) electrical connections. Doped polysilicon has two major advantages over aluminium material. Firstly, poly can be used for a self-aligned process that reduces the parasitic capacitance of the source and the drain, and that in turn, increases the circuit speed; secondly, it can withstand subsequent high-temperature processes.

Polysilicon can also be doped by an implantation process. Generally, the dose of phosphorous is kept around 5×10^{15} atoms cm^{-2}. It is essential to mention that phosphorous ions should not penetrate through the polysilicon film and enter into the gate oxide; otherwise, the quality of the gate oxide will deteriorate.

Once the polysilicon doping is completed, the wafer is processed for dry oxidation. Thereafter, the polysilicon gate and the short connections (not shown in figure) are made using the **gate mask** as shown in Fig. 3.28, and the gate patterns are made on PR as shown in Fig. 3.29. Thereafter, the polysilicon is etched out from all places except the gate areas and local connections; subsequently, the oxide is etched out and then the PR is removed as shown in Figs. 3.30 and 3.31, respectively. Generally, the poly is etched by plasma (dry) process, where chlorine or bromine gas is usually used as an etchant gas. Details of poly etching are covered in Chapter 7 (etching) and Chapter 11.

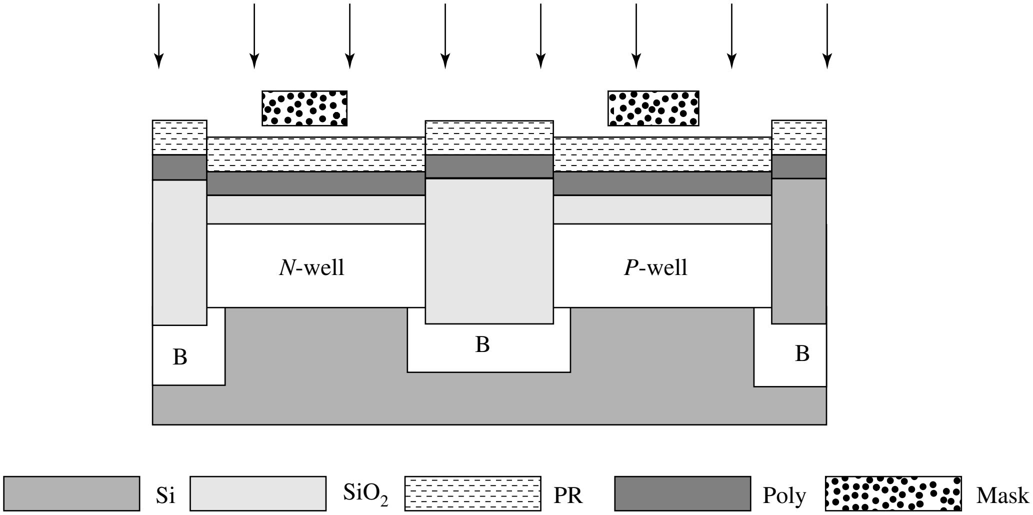


Fig. 3.28 Gate definition

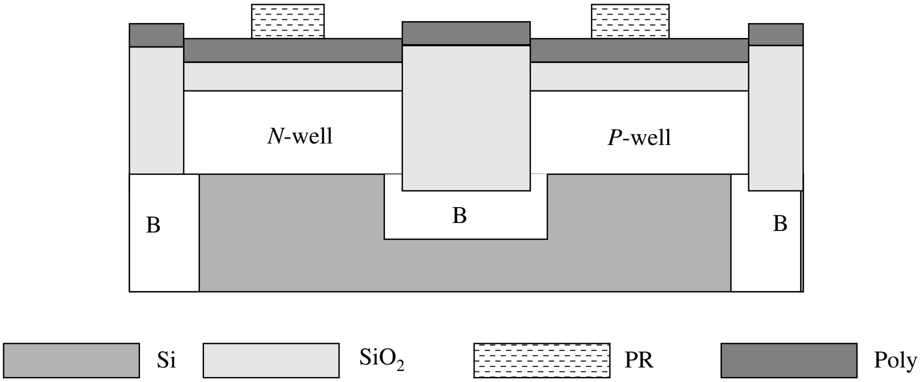


Fig. 3.29 Gate definition

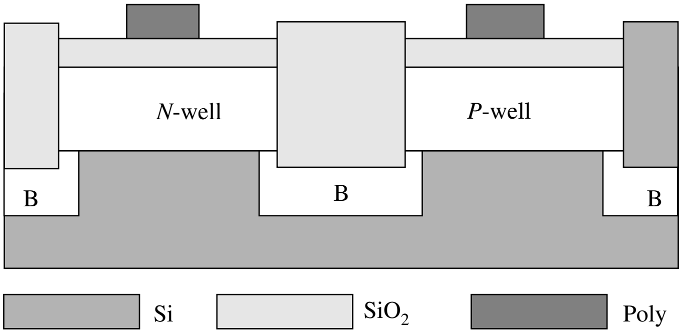


Fig. 3.30 Gate definition

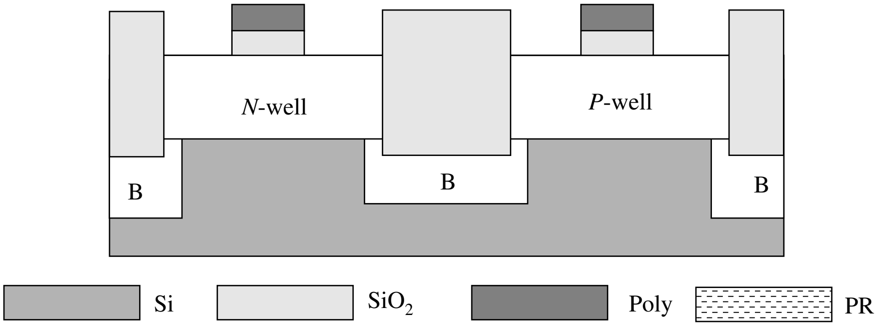


Fig. 3.31 Gate definition

11th step: Tip (or Extension) of Lightly Doped Drain (LDD) for PMOS Transistor

If the MOS device dimensions are reduced and if the corresponding voltage is not reduced, then electric field will be very high across the source and drain. If the electric field exceeds 10^5 $V\,\text{cm}^{-1}$, **hot electrons** (or holes) will be generated; these electrons (holes) will further generate electrons or holes due to the cascading effect resulting from the breaking of the silicon–silicon bond. Consequently, a high current will flow from the source to the drain. Hence, to withstand the electric field, the drain and the source are lightly doped before source and drain formation and this is called a **Lightly Doped Drain**. It reduces the peak value of the electric field near the drain due to P^+P^-N (drain-tip-well) structure in the PMOS. In this structure, the drain voltage drops because of the presence of a P^- tip as compared to an abrupt P^+N junction.

To process the LDD for a PMOS transistor, the lithography step is shown in Figs. 3.32 and 3.33. Once the PR is removed from the source and the drain areas of the PMOS transistor, shallow doping is done by ion implantation using a low dose of phosphorous as shown in Fig. 3.34.

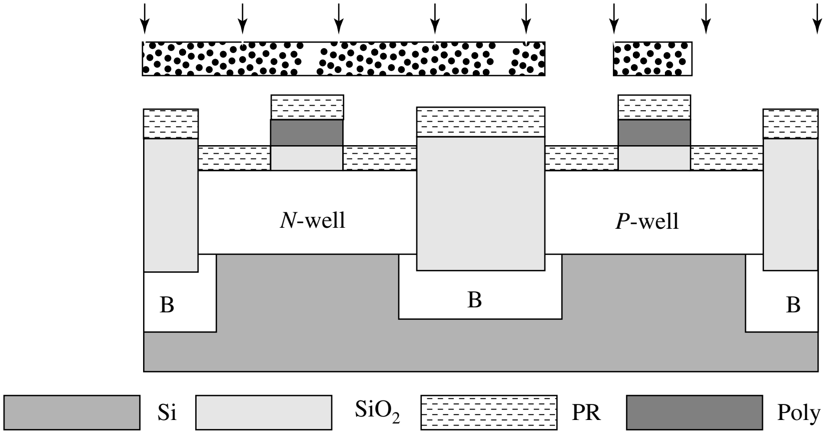


Fig. 3.32 PMOS LDD formation, PR deposition

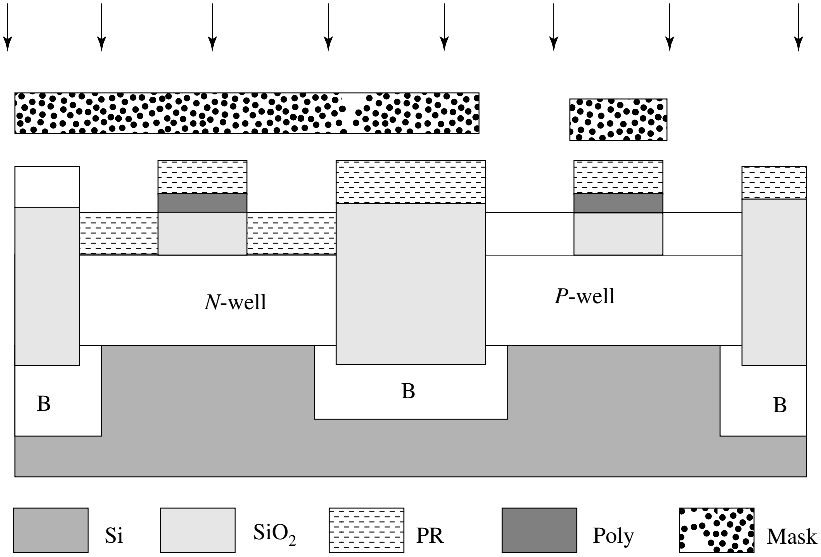


Fig. 3.33 PMOS LDD extension lithography

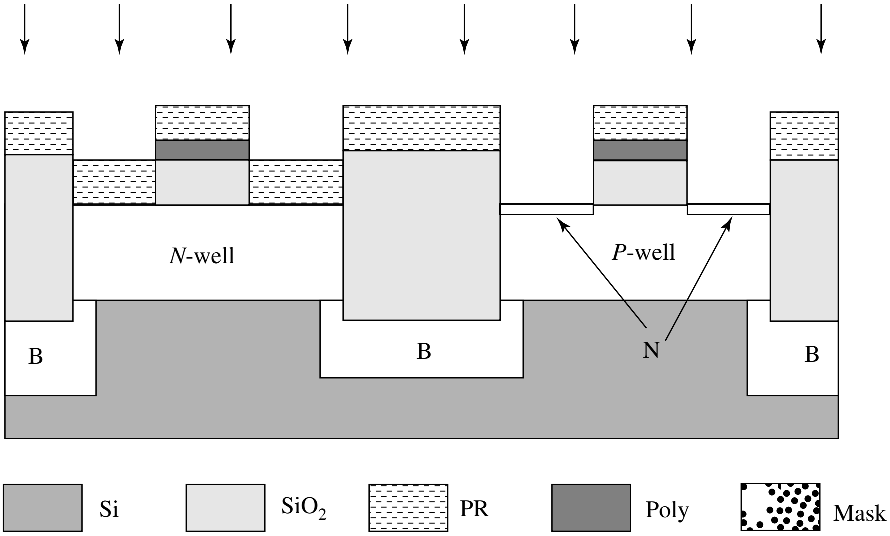


Fig. 3.34 PMOS LDD extension implant formation

12th step: Tip (or Extension) of Lightly Doped Drain (LDD) for NMOS Transistor Similarly, for the LDD of an NMOS transistor, the PR is coated on the wafer as shown in Fig. 3.35. Once the PR is removed from the source and the drain areas of the NMOS transistor, shallow doping is done by ion implantation using a low dose of boron. This reduces the peak value of the electric field near the drain due to $N^{+}N^{-}P$ (drain-tip-well) structure, as shown in Fig. 3.36.

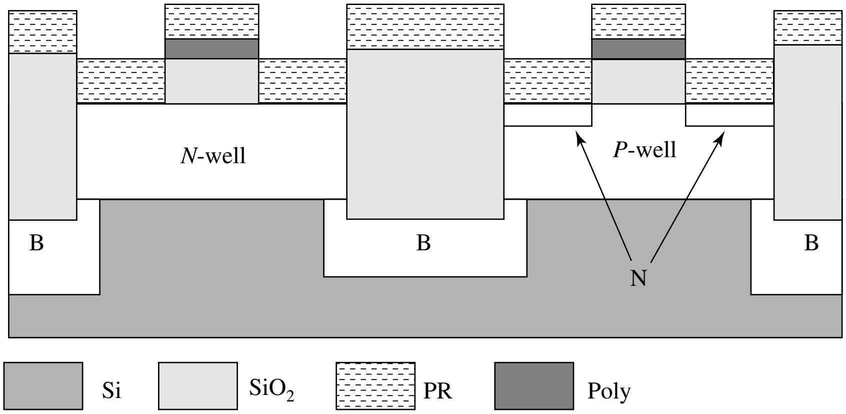


Fig. 3.35 NMOS LDD extension, PR deposition

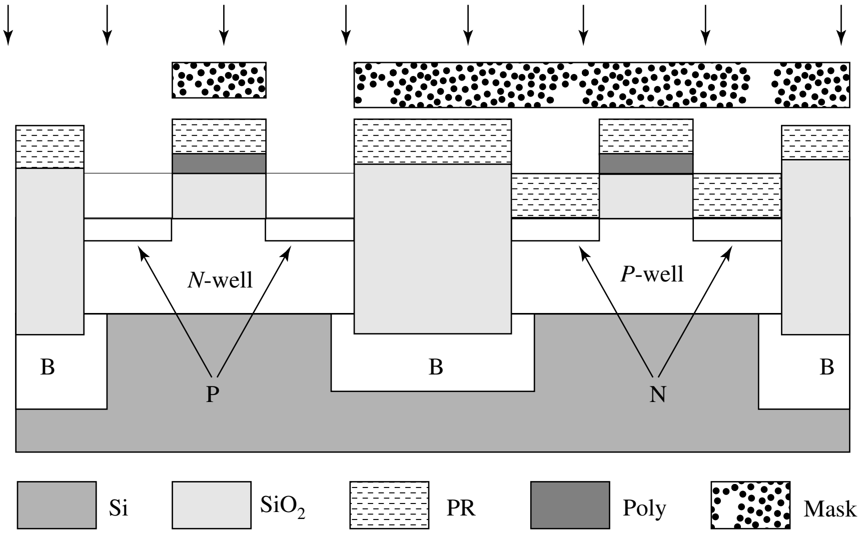


Fig. 3.36 PMOS LDD extension, boron implantation

13th step: Oxide Deposition Once both the NMOS and PMOS tip formations are completed, the oxide is deposited on the entire wafer. Generally, silicon oxide is deposited by the LPCVD process. In the LPCVD process, the oxide deposit is uniformly done with conformal steps inclusive of the source and the drain areas, as shown in Fig. 3.37. Normally, the thickness of the oxide is around 0.05 micrometre. To deposit the silicon dioxide, silane with oxygen or dichlorosilane with nitrous oxide are used.

$$SiH_4 + O_2 \rightarrow SiO_2 + 2H_2 \quad \text{at } 400°C$$

$$SiH_2Cl_2 + 2N_2O \rightarrow SiO_2 + 2N_2 + 2HCl \quad \text{at } 900°C$$

Details of the LPCVD process are described in Step 2.

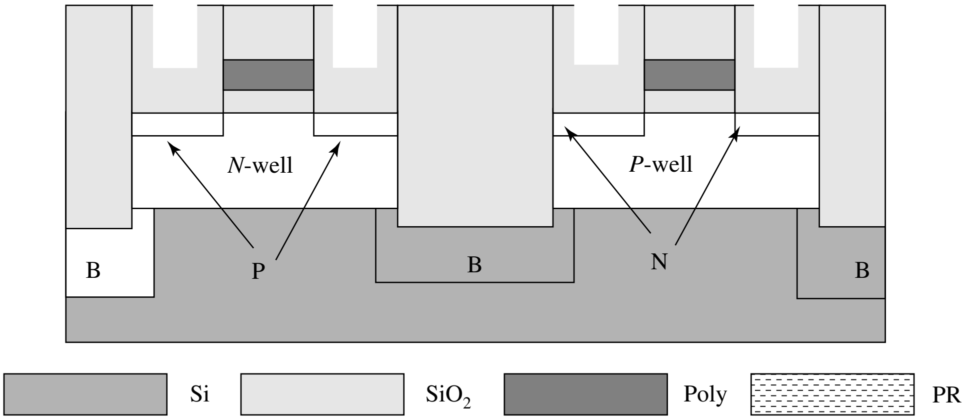


Fig. 3.37 Oxide deposition after LDD

14th step: Oxide Etching Once silicon dioxide is deposited, the wafer is coated with the PR and then lithography is carried out. The PR is removed from the source and the drain of the NMOS transistor except from the collars of the gates, as shown in Fig. 3.38. The oxide is removed by the dry (plasma) etching process so that anisotropic etching takes place; where, fluorine gas is used as the etchant gas.

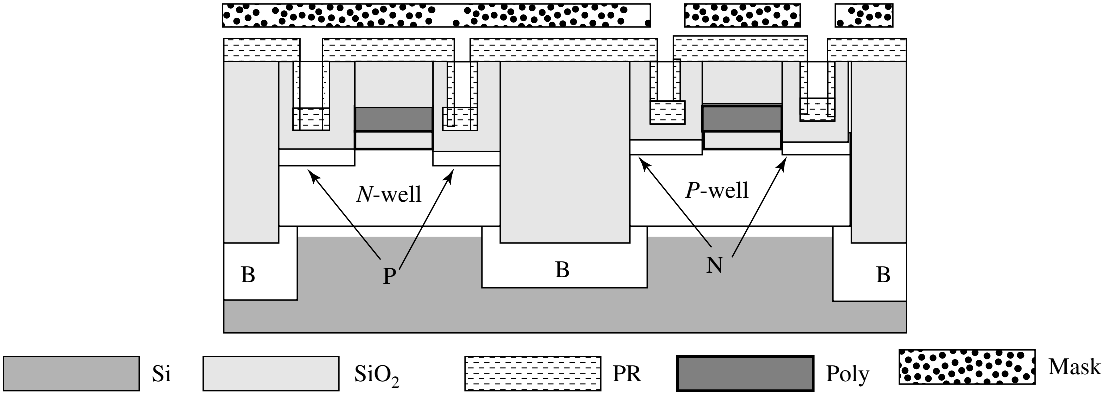


Fig. 3.38 Lithography for NMOS S/D formation

15th step: Self-align Source and Drain for NMOS Fabrication Once the oxide is etched from the source and the drain regions, leaving the collar at the gate, the wafer is implanted by the phosphorous dopant around 2–4 × 10^{15} atoms cm^{-2} with the energy of 75 keV for the formation of the source and drain of the NMOS transistor. The lithography, implantation and PR removal are shown in Figs 3.38, 3.39, and 3.40 respectively.

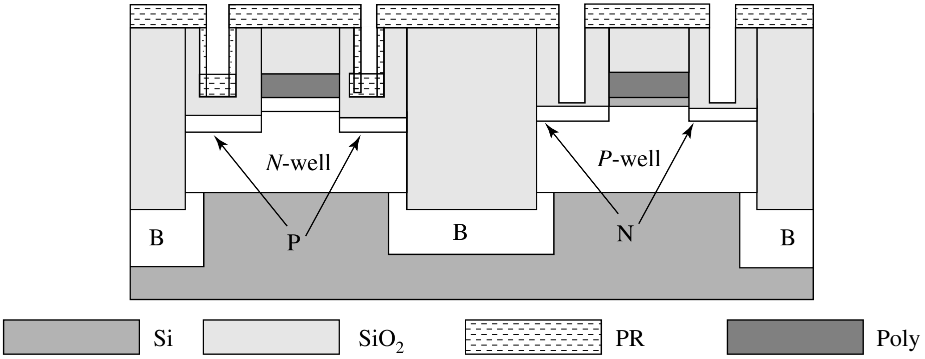


Fig. 3.39 Oxide etching for NMOS S/D formation

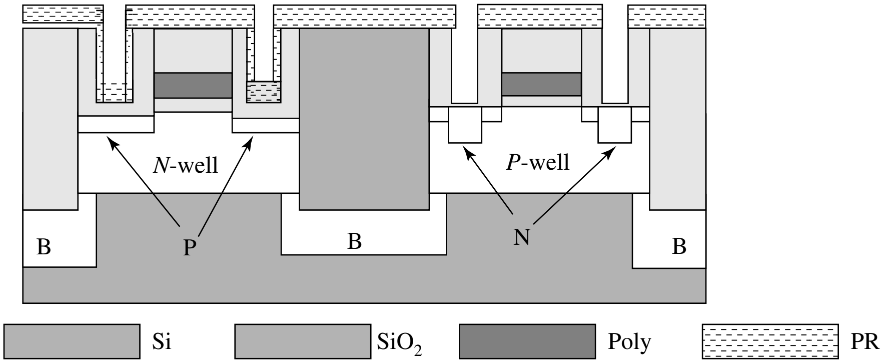


Fig. 3.40 NMOS S/D implant

Similarly, the PMOS source and drain are made by replicating the **PMOS source and drain mask** in conjunction with the lithography process and etching of the oxide in the *N*-well. Thereafter, boron is implanted by around 2 × 10^{15} atoms cm^{-2} dose with the implantation energy of 60 keV. The sequence is depicted by the Figs. 3.41, 3.42 and 3.43.

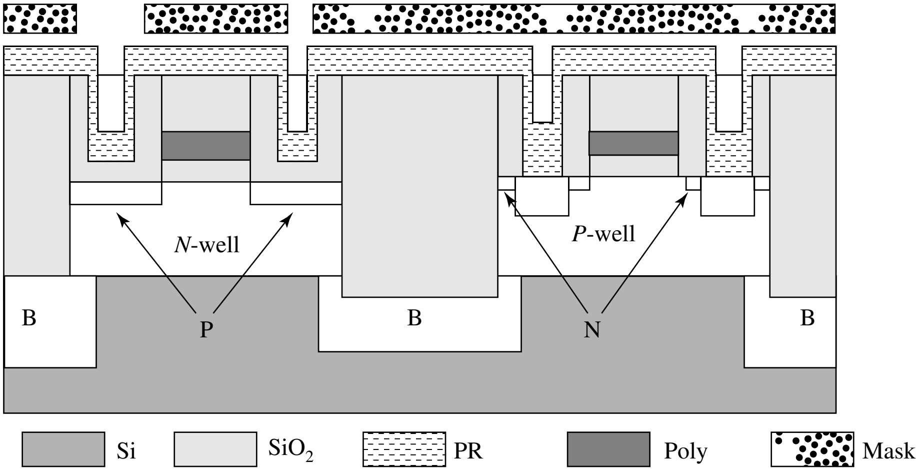


Fig. 3.41 Lithography for PMOS S/D formation

Apart from implantation, the source and the drain of the transistors can also be made by thermal diffusion in the furnace, if the transistors are well separated with little variation in the process flow. Phosphorous diffusion is generally carried out at around 950°C temperature, for about 30 minutes in the presence of the $POCl_3$ vapour. To carry the $POCl_3$ vapour into the furnace, nitrogen gas is used as the carrier gas. It is advisable that at the

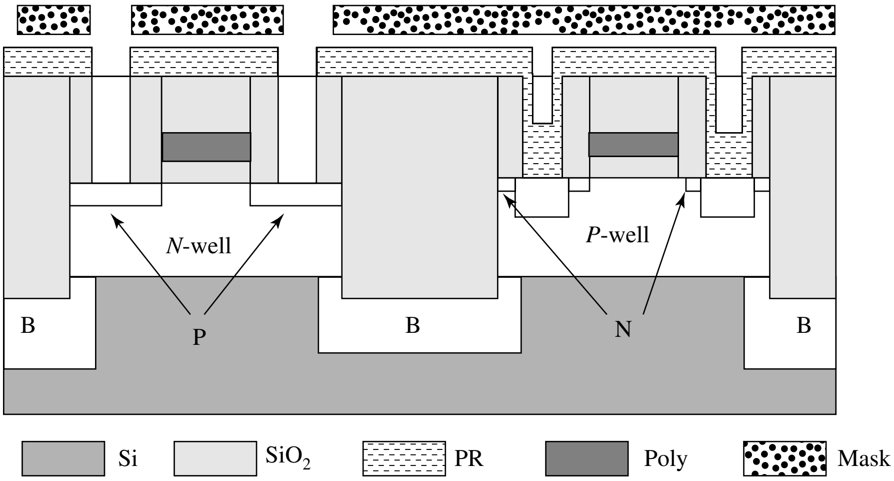


Fig. 3.42 Lithography for PMOS S/D formation

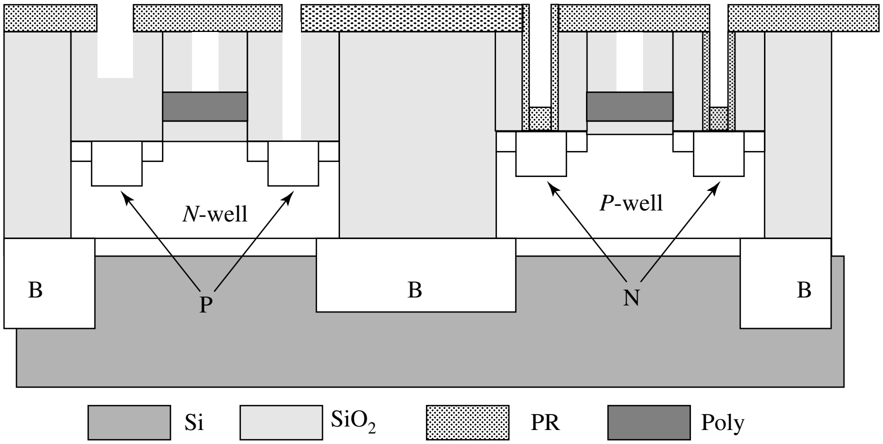


Fig. 3.43 PMOS S/D formation

time of phosphorus diffusion, the wafer must be oxidised to dilute the phosphorous dopant concentration in the oxide film. Similarly, the boron concentration has to be diluted by growing the oxide. Generally, an unintentionally thin layer of oxide is formed due to the trace of oxygen during dopant diffusion. This high-dopant-rich thin oxide is very difficult to etch by the wet etching process. For this reason, the dopant concentration is diluted in thick oxide film. This diluted dopant oxide can be easily removed by wet etching. It is essential that during the diffusion process, the conductivity of the polygate should not change significantly when boron is doped. Generally, the ion-implantation technique is used for the source and drain doping, due to its less lateral diffusion, accurate ion dose, room temperature process and many other advantages. The detail descriptions of the diffusion and ion-implantation techniques are covered in Chapters 8 and 9, respectively.

After the source and drain formation of both the transistors, the wafer is cleaned thoroughly (Fig. 3.44), and heated at around 1100°C temperature in the oxidation furnace for about 1 hour in the presence of oxygen. By this process, the required junction depths of the source and drain, and a thick oxide are obtained. This process is called the **oxidation and drive-in process**, as shown in Fig. 3.45. In the ion-implantation process, the silicon atoms are dislodged from their lattice site and this creates damages in the wafer due to the dopant ions strike. The oxidation and drive-in process repairs these damages and also puts about 90% of the dopant atoms in the silicon lattice site, which is required for good transistor operation and eliminates the need for extra annealing of the wafer.

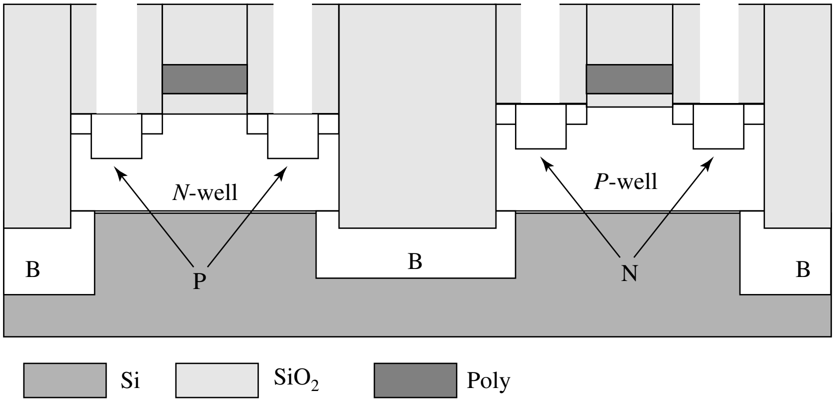


Fig. 3.44 PR removed

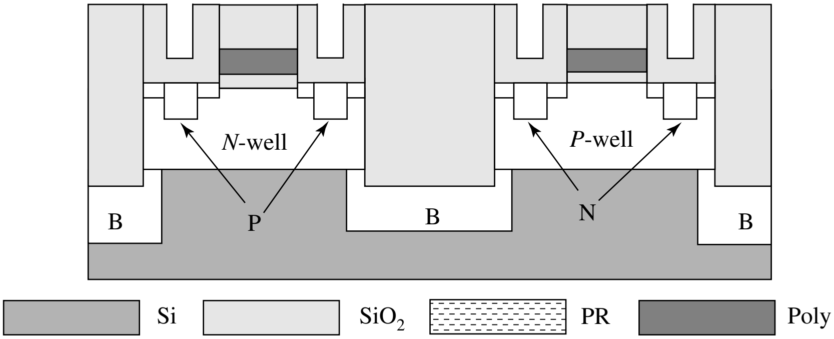


Fig. 3.45 S/D drive-in

16th step: Contact Window Once the source and drain fabrication of the NMOS and PMOS transistors is completed, the wafer is processed for electrical connections. It has been mentioned previously that local (shot) connections are made by doped polysilicon, but it cannot be used for long-distance electrical connections due to high resistivity. Therefore, for longer distance, electrical wiring is done using metal. To do the metal wiring, the contact windows are opened at the source, the drain and the gate regions. This is done by using contact mask in conjunction with the lithography and the oxide etching processes, as shown in Fig. 3.46; thereafter, the PR is removed, as shown in Fig. 3.47.

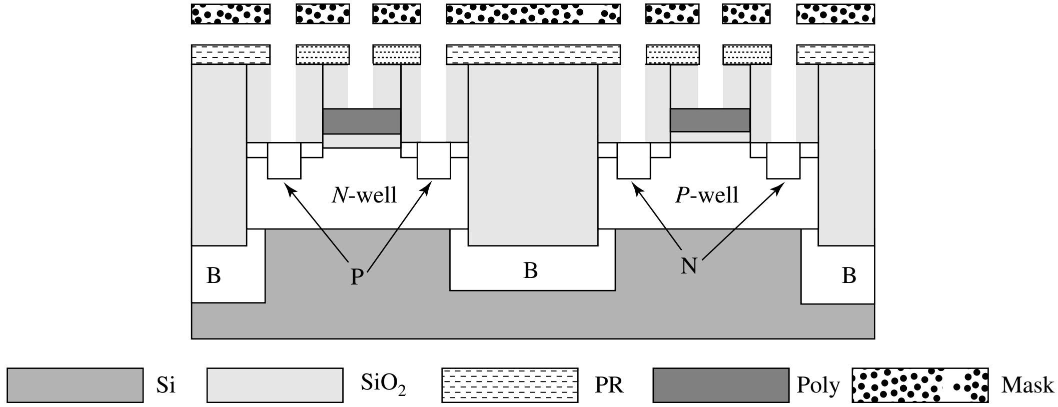


Fig. 3.46 Lithography for contact windows for S/D

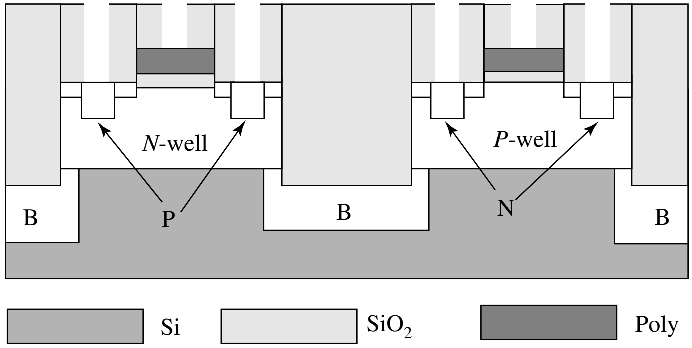


Fig. 3.47 Contact windows for S/D

17th step: Multi-level Wiring Process Steps High-density IC chip electrical connections are made using multi-level electrical connections. This is because of insufficient space available on the wafer surface (or single layer) for electrical connections. For local connections, a tungsten film is deposited using the sputtering or the CVD process, and thereafter, the tungsten is patterned. The tungsten deposition and patterning are shown in Figs. 3.48 and 3.49 respectively. The details of tungsten deposition are described in Chapters 10 and 11. Then, a dielectric film is deposited on the wafer and thereafter, holes are made at required places in the dielectric, as shown in Figs. 3.50 and 3.51. After making contact holes in the dielectric, tungsten metal is deposited on the wafer and it is etched except from the holes.

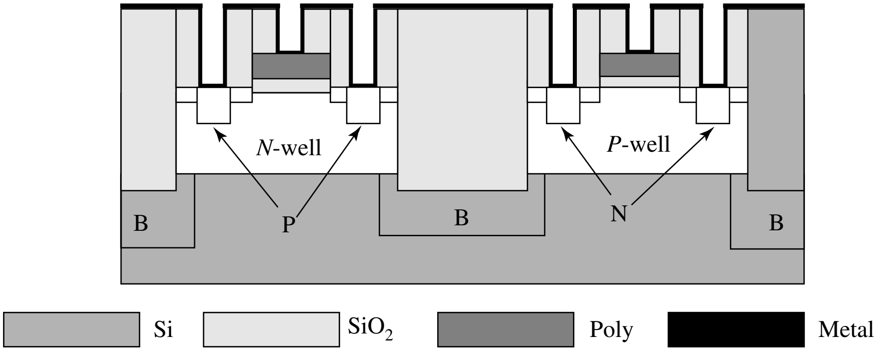


Fig. 3.48 First-level metallisation

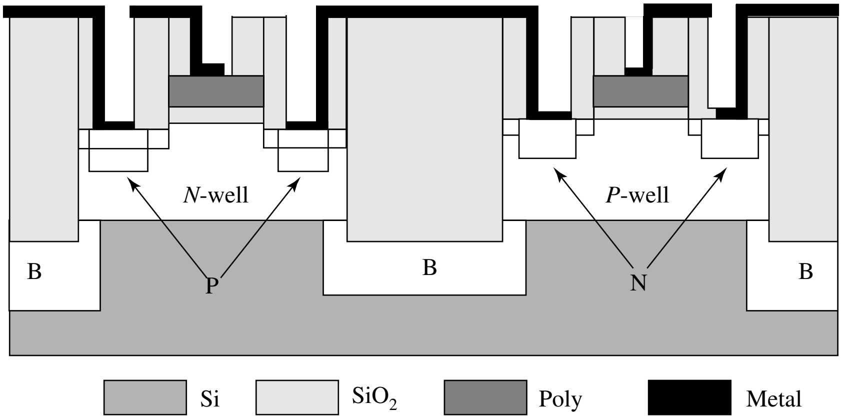


Fig. 3.49 First-level metal patterning

18th step: Aluminium Metal Deposition Once holes are filled with tungsten the wafer is deposited by an aluminium metal film. Generally, the aluminium metal film is deposited by the Physical Vapour Deposition (PVD) process, either by the **thermal evaporation** technique or by the **sputtering technique**. In both the techniques, vapours of aluminium atoms are produced and they deposit on the wafer in an evacuated chamber. The deposited aluminium film on the wafer is shown in Fig. 3.52. The equipment which is used for the evaporation

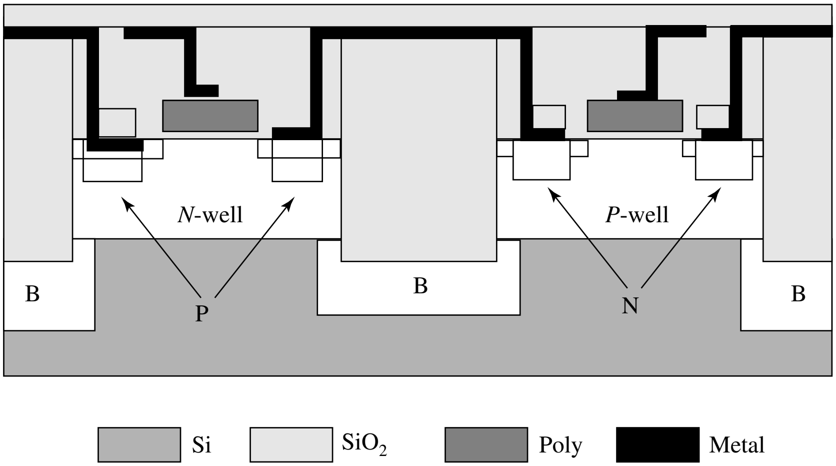


Fig. 3.50 First-level metal patterning

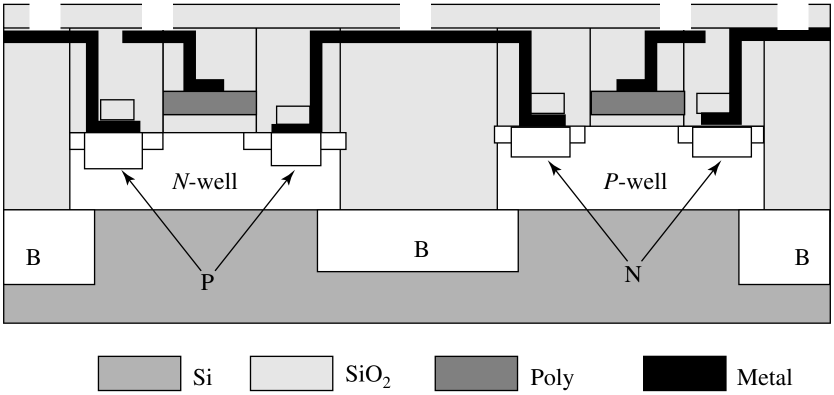


Fig. 3.51 Via for second metal

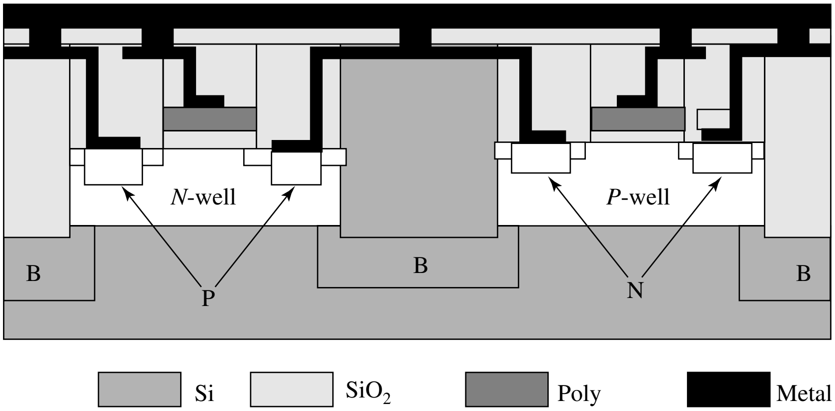


Fig. 3.52 Second-metal deposition

of aluminium (or other metal) is called the **evaporation metallization system (unit)**; and the equipment which is used for sputtering the metal to produce metal vapour is called the **sputtering system**. In the thermal evaporation deposition process, pieces of aluminium wire are hanged on the tungsten filament and then the tungsten filament is heated with the help of an electric current. The melted aluminium produces aluminium vapours and they get deposited on the wafer. In the sputtering process, vapours of aluminium ions are produced by knocking out aluminium atoms from the aluminium plate (called the **target**) by striking high energetic argon ions. In VLSI and ULSI metallization, sputtering technique is preferred because of its uniform metal deposition. The details of PVD and sputtering are described in Chapter 10.

It is important to mention that when bare silicon is exposed (especially after opening the contact windows) to the atmosphere, the silicon oxidises to a thickness of around 20 Å immediately and increases oxide thickness with time. This grown oxide is sandwiched in between the metal and the silicon which may lead to contact resistance. Hence, it is advisable that the wafer should not be exposed for a long time.

19th step: Aluminium Film Patterning (Wiring) Process Steps Once aluminium metal is deposited (coated) on the wafer, the PR coating is done on it. The aluminium is patterned from the metal mask by lithography. Once PR patterning is done, the wafer is dipped into orthophosphoric acid to etch the aluminium metal from the uncovered PR areas. Thereafter, PR is completely removed from the wafer in acetone and the wafer is cleaned thoroughly. Now the wafer is annealed in the furnace at around 450°C for 30 minutes in nitrogen gas or a mixture of H_2 and N_2 gases. This annealing process promotes good contact between the metal and silicon, and eliminates the radiation damage generated at the time of metal deposition. Aluminium wiring patterns are shown in Fig. 3.53; except the contact windows, all electrical wirings are run over the thick oxide (FOX). FOX being a thick oxide, it prevents electrical shorts and formation of parasitic transistors.

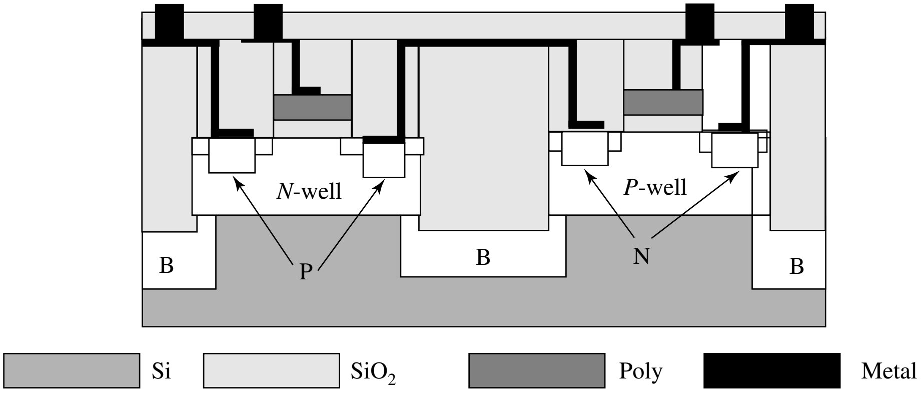


Fig. 3.53 Second-metal deposition and patterning

Summary

The process flow of the IC has to be frozen prior to fabrication. The architecture, design rules and process flow are interlinked and these criteria have to match with the manufacturer's process capabilities. Hence, it is extremely essential to know the IC process flow. Furthermore, the process flow also involves the physics of the semiconductor and the fabrication techniques to be used. In addition, it gives the guidelines for process sequences and process recipes. Presently, polysilicon gate and planar technology are used to realise ICs. Polysilicon gate has the advantages of self-align process, local electrical connections and low gate-to-source and drain capacitances; it also allows subsequent high-temperature processes.

To avoid the electrical shorts between the closely placed MOS transistors, usually shallow-trench electrical isolation is used. It has advantages of zero oxide step height, lower silicon consumption and good isolation of the source and the drain of one MOS transistor from the other MOS transistors inside the silicon. It seems that shallow electrical isolation will continue to be used in the near future.

Prior to IC fabrication, the silicon wafer specifications must be frozen as they govern the characteristics of the MOS transistor. These specifications mainly are type of the wafer, resistivity (doping level), crystal orientation (mostly 100), wafer size (wafer process line), wafer flatness, defect density, and impurity levels.

One of the essential parts of MOS fabrication is the threshold voltage (V_{th}) of transistors. The doping concentrations of the wells are tailored according to the desired threshold voltage of the MOS transistors. This process is done by ion implantation and is described in Chapter 9.

References

- J D Plummer, M Deal and P B Griffin; *Silicon Fundamental Technology: Fundamentals, Practice and Modelling*, Prentice Hall, 2000
- S M Sze, *VLSI Technology*, Second Edition, McGraw-Hill, 1988

Multiple-Choice Questions

3.1 Generally, what is the depletion width of a channel?
(a) 20 Å (b) 50 Å (c) 100 Å

3.2 The minimum geometry in a transistor is generally
(a) gate width (b) source and drain width (c) contact area

3.3 For MOS fabrication, which orientation of silicon is preferred?
(a) (1,0,0) (b) (1,1,1) (c) (1,1,0)

3.4 For device isolation, which technique is the best
(a) thick oxide (b) LOCOS (c) shallow trench

3.5 Which isolation techniques is a more complicated process?
(a) LOCOS (b) Shallow trench

3.6 Why is chlorinated gas used?
(a) To reduce mobile charges (b) To reduce the dislocation
(c) To reduce the oxide stress

3.7 Why is field implantation essential?
(a) To reduce parasitic transistor (b) To reduce the capacitance
(c) To reduce the resistance

3.8 Annealing helps the dopant to occupy the substitution site of silicon. Is the statement true or false?
(a) True (b) False

3.9 In planar technology, contacts are taken on
(a) top of the wafer (b) bottom of the wafer

3.10 Generally, the gate polysilicon is doped by
(a) boron (b) phosphorus

Descriptive Problems

3.1 Explain the diode process steps using planar technology.

3.2 Explain PMOS process steps through a sketch.

3.3 Write the relationship between the depletion width to wafer doping concentration, and explain the physics behind it.

3.4 Why is polygate generally doped by phosphorous dopant?

3.5 What would be the effect of PMOS characteristics if dopant penetrates through the poly film and enters into the gate oxide?

3.6 If the metal gate and the polysilicon gate are overlapping with the source and drain of different MOS transistors by 0.25 μm and 0.1 μm respectively then compare the capacitances for the same gate oxide of 500 Å.

3.7 What is the reason why the well depth is kept around 3 microns when the source and drain junction depth is made of 1 micron.

CHAPTER

4 *Oxidation*

4.1 INTRODUCTION

Silicon dioxide is used as a gate electrode besides being used for electrical isolation, surface planarisation and passivation in a MOS transistor; and also used as a mask barrier for impurities and a pad oxide for nitride deposition during MOS transistor fabrication. Therefore, the silicon dioxide must have qualities such as lower stress, high refractive index, higher density, higher breakdown, good composition, less pinholes, lower defect density, and amorphousness. These qualities of the oxide can be easily achieved when silicon is heated in a furnace at a high temperature in the presence of oxygen gas. This oxidation process is carried out in a furnace which is called an **oxidation furnace**. There are two types of oxidation processes used in MOS transistor fabrication, namely dry oxidation and wet oxidation. In dry oxidation, dry and pure oxygen gas is used; whereas, in the wet oxidation process, water vapour is used. The oxidation rate of wet oxidation is significantly higher than that of dry oxidation; hence, whenever a thicker oxide is needed, the wet oxidation process is used. In contrast to wet oxidation, the rate of dry oxidation is significantly low, but the oxide quality is far superior as compared to that of wet oxidation. There are two techniques used to produce the water vapour for wet oxidation. In the first technique, the water vapour is produced by boiling water around 95ºC in a special type of container called **bubbler**; then, the water vapour is transported into the high-temperature oxidation furnace. In the second (pyrogenic) technique, the water vapour is produced by burning pure hydrogen and oxygen gases and then, it is transported to the oxidation furnace. The silicon dioxide structure, the oxidation procedures, the modelling (kinetics) of oxidation, the quality of the oxide and other relevant issues are discussed below.

4.2 STRUCTURE OF SILICON DIOXIDE

The silicon dioxide is made of a large number of polyhedron (tetrahedral or triangular) networks as shown in Fig. 4.1. In a triangular polyhedron network, four oxygen atoms are placed at the corners and the silicon atoms are at the centre. Each of the oxygen atoms is covalently bonded with the silicon atoms. In the tetrahedral network, the internuclear distance between the silicon and the oxygen atoms and two oxygen atoms are 1.62 Å and 2.27 Å respectively. The different possible silicon dioxide structures are shown in Fig. 4.2. When the oxygen atoms of a polyhedron are bonded with two silicon atoms of neighbouring polyhedrons, the silicon dioxide is called **bridged oxide**. If the oxygen atoms are bonded with one silicon atom of the neighbouring polyhedron, the silicon dioxide network is called **nonbridged oxide**. It is found that the bridged oxide is more cohesive and less damage-prone as compared to the nonbridged oxide. In case of dry oxidation, a large fraction of the bridged network is present in the silicon oxide as compared to the nonbridged oxide. On the other hand, less bridged oxide is formed and a large fraction of the nonbridged network oxide is formed by the wet oxidation process. Furthermore, in dry oxidation, oxygen is directly bonded with the silicon atoms; whereas, in the case of wet oxidation, some of the hydroxyl (OH) molecules are directly bonded with the silicon atoms in the oxide network. These OH—Si bonds are weak and can be broken easily; and, elements (impurities) can pass through the oxide. The silicon dioxide may be in the crystalline, polycrystalline and amorphous forms. The crystalline silicon dioxide is also called a **quartz crystal**, where the oxide network structures are arranged in a particular orientation in a definite periodic manner. This crystalline silicon is never used in IC fabri-

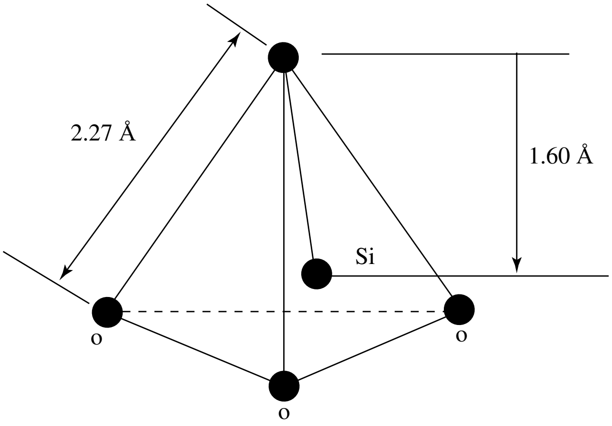
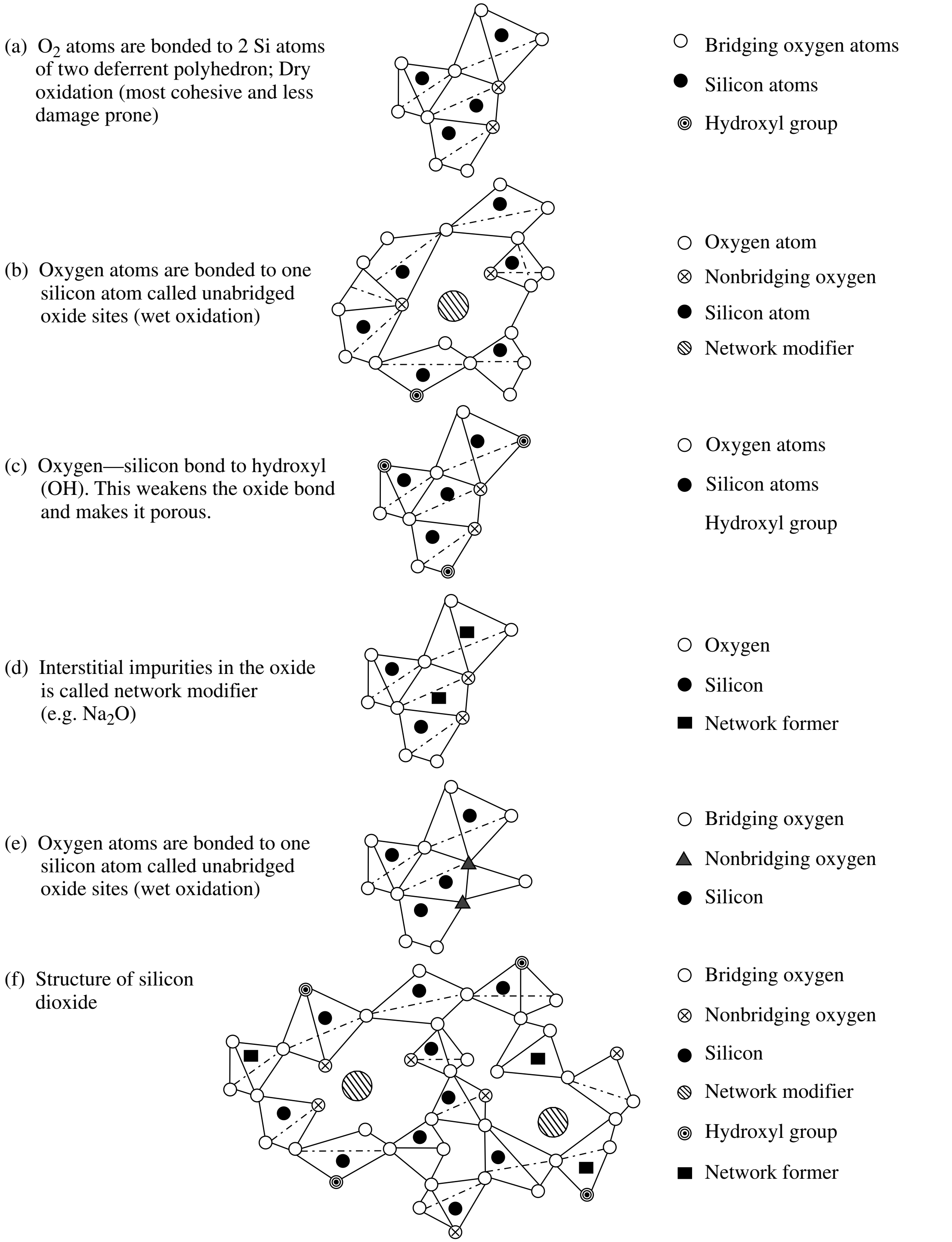


Fig. 4.1 Silicon dioxide structure



(a) O_2 atoms are bonded to 2 Si atoms of two deferrent polyhedron; Dry oxidation (most cohesive and less damage prone)

(b) Oxygen atoms are bonded to one silicon atom called unabridged oxide sites (wet oxidation)

(c) Oxygen—silicon bond to hydroxyl (OH). This weakens the oxide bond and makes it porous.

(d) Interstitial impurities in the oxide is called network modifier (e.g. Na_2O)

(e) Oxygen atoms are bonded to one silicon atom called unabridged oxide sites (wet oxidation)

(f) Structure of silicon dioxide

Fig. 4.2 Structure of silicon dioxide

cation due to the formation of a dipole under the electric field and the piezoelectric effect. In contrast to crystalline oxide, the amorphous oxide has a very short range of the structural orientations and does not possess any definite periodic structure. This type of oxide is called **amorphous** or **fused quartz**. Polycrystalline oxide lies between the crystalline and the amorphous forms of silicon dioxide.

If the silicon atoms are replaced in their sites by the dopant or impurity atoms, the oxide is referred to as the **network former**. These network former oxides have a lower melting point as compared to pure silicon oxide. This property of oxide is exploited in the process of IC fabrication, especially to perform the surface planarisation of the wafer. Here, the oxide is heated to a temperature of around 1020ºC to make it viscous, and then, it spreads uniformly over the wafer surface. This process is called **oxide reflows**. Generally, the network former oxide is doped with boron or phosphorous dopant and the resulting oxides are called **Boron Silicate Glass (BSG)** and **Phosphorous Silicate Glass (PSG**) respectively. The oxide that contains the ionic form of metal impurities (i.e. Na^+, K^+, Pb^{2+}, etc.) converts the bridge network oxide into the nonbridged oxide network, and this type of oxide is called the **network modifier**. Usually, these metal ions reside interstitially in the sites of the oxide. As a result of the conversion of the bridged oxide to the nonbridged oxide, the oxide network bonds weaken and this leads to a porous oxide which allows impurities to pass through.

4.3 OXIDATION EQUIPMENT AND PROCESS

The oxidation of the silicon wafer is done in the oxidation furnace. Usually, the oxidation of the wafer is done in the temperature range of 900ºC to 1200ºC. Sometimes, oxidation at a lower temperature around 700ºC is preferred for a thin oxide film. The oxidation furnace contains a fused quartz tube, which is heated by a segmented resistive heater wound from the outside of the quartz tube, and is called the **heating zone**. A typical oxidation system is depicted in Fig. 4.3. In order to start the oxidation, nitrogen gas is flown in the oxidation furnace and then, by adjusting the current through the segmented heaters, a uniform temperature range (±1ºC) is attained at the centre of the furnace. This uniform temperature range of the furnace is called the **centre zone.** The temperature at both the ends of the furnace is comparatively lower than that of the centre zone. From one end of the furnace tube, the wafer is inserted for oxidation. This end of the furnace tube is called the **mouth** and is usually referred to as the **front end**. The opposite end of the furnace tube is connected to the gas and water-vapour delivery systems and this end of the furnace is called the **rear end**. When the required oxygen temperature is attained, the wafer is kept

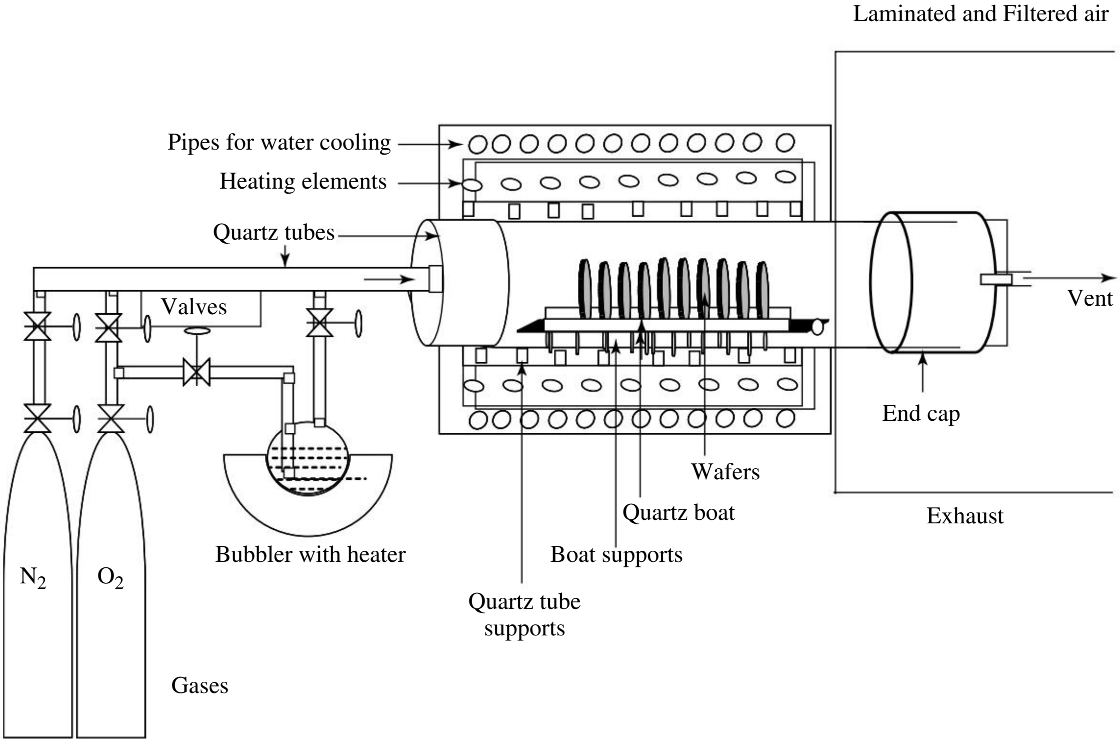


Fig. 4.3 Wafer oxidation system

at the mouth of the furnace for a few minutes and then, it is slowly pushed into the centre zone. This step of oxidation is called **wafer loading**. Prior to keeping the wafer at the mouth of the furnace, the wafers are placed in a specially made, slotted quartz jig, called the **quartz boat**. To avoid a sudden temperature change during the oxidation process, the wafer should not be pushed inside the furnace fast; otherwise, this may lead to defects in the silicon wafer and in the worst case, the silicon wafer may shatter into pieces. The oxidation is carried out using the dry oxidation process by closing the flow of the nitrogen gas and then passing the oxygen gas into the furnace for a few minutes to grow the thin silicon dioxide. Thereafter, the wet oxidation process is applied by diverting the oxygen gas through the water bubbler into the furnace. The oxygen gas carries the water vapour into the furnace and wet oxidation starts. Depending on the oxide thickness, the time for wet oxidation is determined; generally, for a 1-micrometre thick silicon oxide, it takes more than an hour. When the wet oxidation process is completed, the oxygen gas is disconnected from the bubbler and is passed directly into the furnace for dry oxidation for the second time. This entire oxidation process is called the **dry-wet-dry oxidation process**. After the completion of the dry-wet-dry oxidation process, the oxidation gas is stopped and the nitrogen gas is passed into the furnace. The wafers are kept in the nitrogen environment for some time at the oxidation temperature. This helps to reduce the wafer defects as well as the oxidation charges that had manifested in the process of oxidation. This particular process is called the **wafer-annealing process** or simply the **annealing process** or **wafer annealing**. Thereafter, the oxidised wafers are slowly withdrawn from the centre zone to the mouth of the furnace and kept there for a few minutes to avoid a sudden temperature change. Then, the wafers are taken out of the furnace mouth and transferred to the next process. The whole process starting from the wafer loading into the boat to the unloading from the boat after oxidation is called the **oxidation process**. The parameters of the oxidation process such as oxidation temperature, oxidation time, wafer loading and unloading time are collectively called the **oxidation process recipe** or in short, the **oxidation recipe**.

4.4 KINETICS OF OXIDATION

In the oxidation process, the oxygen atoms react with the silicon atoms and form silicon dioxide. If the silicon surface is already oxidised then the oxygen atoms diffuse through the grown (or growing) oxide layer and react with the silicon atoms at the oxide/silicon interface and convert the silicon to silicon dioxide. Thus, the interface moves inside the silicon wafer and the silicon dioxide thickness increases as shown in Fig. 4.4.

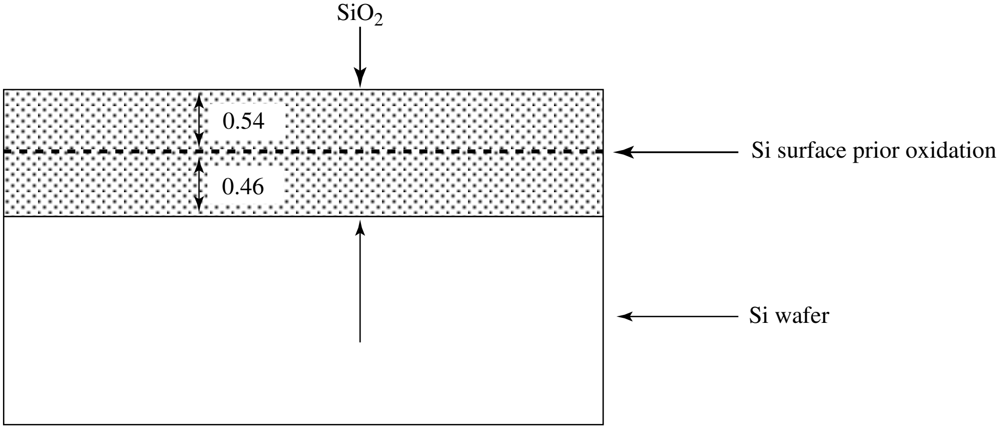


Fig. 4.4 Silicon dioxide growth mechanism

Example 4.1

Estimate the consumption of the silicon layer by the oxidation process.

Answer

Volume of 1 mol of Si = Si molecular wt./Si density
= 28.09 g/mol /2.33 g/cm^3 = 12.06 cm^3/mol

Volume of 1 mol of SiO_2 = SiO_2 molecular wt./SiO_2 density
= 60.08 g/mol/2.21 g/cm^3 = 27.18 cm^3/mol

1 mol of Si is converted to 1 mol of SiO_2 = Vol. of 1 mol of Si/Vol. of 1 mol of SiO_2

Volume/area = Thickness of Si/Thickness of SiO_2
= 12.06/27.18 = 0.44 (thickness of SiO_2)

This means that growing 1000 Å thick silicon dioxide will consume a 440 Å thick layer of silicon. It is to be noted that the thickness of the grown oxide is more than the thickness of the silicon layer consumed. This increase of volume is due to the incorporation of oxygen atoms. As a consequence, the height of the oxidised surface also increases above the original wafer surface of silicon. In the dry oxidation process, around 2.2×10^{22} molecules per cm^3 of oxygen atoms are incorporated, whereas in wet oxidation, twice the number, i.e. $2(2.2 \times 10^{22})$ water molecules (H_2O) are incorporated in a unit volume of the oxide.

To estimate the oxide thickness prior to silicon oxidation, the understanding of a mathematical model is essential. This oxidation model (for that matter any process model) minimizes large numbers of experimental iterations that saves time and money. A reasonably good oxidation model was developed by Deal and Grove in the early 1960s. Unfortunately, no perfect mathematical model has been developed till today. Hence, the Deal-and-Grove model is used till today and in special cases, the model is modified.

4.4.1 Deal-and-Grove Model

As mentioned previously, the oxygen atoms react with the silicon atoms and form silicon dioxide. If silicon is already oxidised then the oxygen atoms diffuse through the oxide layer and react with the silicon atoms at the oxide/silicon interface and form silicon dioxide. Thus, the oxidation chemical reaction can be expressed as

$$Si + O_2 = SiO_2$$

$$Si + 2H_2O = SiO_2 + 2H_2$$

To describe the Deal-and-Grove oxidation model, let us assume that there is an oxide layer (t_{ox}) present on the wafer. Let us also assume that a number of oxygen atoms (say flux F_1) from the nondepleting (no change of oxygen concentration) bulk oxygen gas, is transported into the gas bulk region I. Let us consider that some of the oxygen atoms diffuse through the oxide layer (say flux F_2) in the region II and reach the oxide/silicon interface where they react with the silicon atoms, and the rate of consumption of oxygen (reaction of oxygen with silicon) is flux F_3, as shown in Fig. 4.5. At the state of equilibrium (steady state), all the three fluxes are equal and can be represented mathematically as

$$F_1 = F_2 = F_3$$

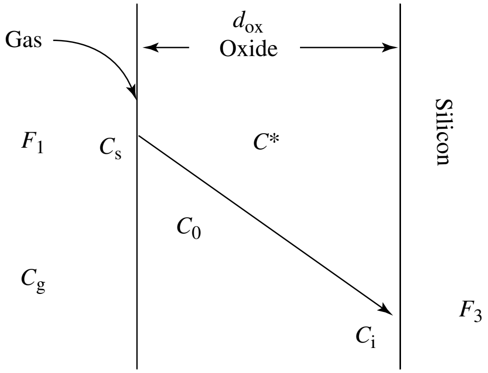


Fig. 4.5 Model of thermal oxidation

Let us suppose that C_g is the undepleted oxygen concentration in the bulk during the oxidation process and C_s is the concentration of oxygen at the gas/oxide interface. It is obvious that C_s is less than C_g due to the fact that the oxygen diffuses (flux F_1) into the oxide during oxidation. Then, flux F_1 can be expressed as

$$F_1 = h_g(C_g - C_s) \tag{1}$$

where h_g is a constant and is called the **gas-phase mass-transport coefficient**.

If C_o is the oxygen concentration just inside the oxide at the gas/oxide interface, and its partial pressure is P_s then at equilibrium condition, and applying Henry's law, the oxygen concentration in solid *proportional to* partial pressure P_g of bulk oxidant and can be expressed as

$$C_o = HP_s \tag{2}$$

where H is called the **Henry constant**.

so the the oxygen concentration in the oxide can be written as

$$C^* = HP_g \tag{3}$$

where H is the Henry constant.

From the ideal gas law, the partial pressure C_g of the oxygen in the region I, and the partial pressure P_g of the oxygen in the region II can be related by the Boltzmann constant and temperature as

$$C_g = \frac{P_g}{kT} \tag{4}$$

Similarly, the partial pressure of oxygen in the bulk oxide near the gas/oxide interface can be expressed as

$$C_g = \frac{P_s}{kT} \tag{5}$$

Now from the above equations (1), (2) and (3), one can get

$$F_1 = h_g(C_g - C_s) = \left(\frac{h_g}{kTH}\right)(C^* - Co) = h(C^* - Co) \tag{6}$$

where, $h = h_g/kTH$ called the **gas-phase mass-transfer coefficient in oxide**.

Using Fick's diffusion law, flux F_2 can be written as

$$F_2 = D\frac{C_o - C_i}{t_{ox}} \tag{7}$$

where D is the **diffusion coefficient**, t_{ox} is the thickness of the oxide, and C_i is the oxygen concentration at the oxide/silicon interface.

When the oxygen atom reaches the oxide/silicon interface, it reacts with the silicon atom. Hence, the consumption or flux of the oxygen F_3 at the oxide/silicon interface can be expressed in terms of the oxygen concentration, and the oxygen and silicon reaction rate as

$$F_3 = K_s C_i \tag{8}$$

where K_s is the rate constant of **surface reaction** of the silicon.

At equilibrium,

$$F_1 = F_3$$

Hence,

$$h(C^* - C_o) = K_s C_i \tag{9}$$

Equating F_2 and F_3, we get,

$$D\frac{(C_o - C_i)}{t_{ox}} = K_s C_i \tag{10}$$

$$C_i = 1 + \left(K_s \frac{t_{ox}}{D}\right)\frac{C_o}{} \tag{11}$$

Replacing C_o from the above equation,

$$C_i = \frac{C^* - \left(\frac{K_s C_i}{h}\right)}{1 + \left(\frac{K_s t_{ox}}{D}\right)} = \frac{C^*}{1 + \left(\frac{K_s}{h}\right) + \left(K_s \frac{t_{ox}}{D}\right)} \tag{12}$$

Solving for C_o, we get,

$$C_o = \frac{\left(1 + \left(K_s \frac{t_{ox}}{D}\right)\right) C^*}{1 + \left(\frac{K_s}{h}\right) + \left(K_s \frac{t_{ox}}{D}\right)} \tag{13}$$

The rate of oxidation depends mainly on two limiting parameters. The first limiting parameter is that how fast the oxygen atoms diffuse through the oxide and reach the oxide/silicon interface; and the second limiting parameter is that how fast the oxygen atoms react with the silicon atoms at the oxide/silicon interface. These limitations are explained below.

Limiting Case 1

If the diffusion coefficient D is very small, it leads to lesser availability of oxygen at the oxide/silicon interface with respect to the surface rate reaction.

Then, $C_i \to 0$ and $C_o \to C^*$

This situation of oxidation is called **diffusion-controlled oxidation**.

Limiting Case 2

If the diffusion coefficient D is very large, the oxygen concentration C_o reaches the oxide/silicon interface very fast, but the rate of oxidation is limited by the surface rate reaction. Then C_i can be written as

$$C_i = C_o = \frac{C^*}{1 + \left(\frac{K_s}{h}\right)} \tag{14}$$

This situation of oxidation is called **reaction-controlled case**. This means that oxidation rate does not depend on the equilibrium oxygen concentration C^*, but depends only on the reaction rate constant K_s.

Oxide Growth Rate

As stated previously, consider that N_1 is the number of oxygen molecules incorporated into a unit volume of the oxide layer (SiO_2 is around 2.2 × 10^{22} molecules /cm^3). Hence, in the dry oxidation process where pure oxygen is used, $N_1 = 2.2 \times 10^{22}$ oxygen molecules / cm^3 is required to form silicon dioxide; whereas in the wet oxidation process, $N_1 = 2(2.2 \times 10^{22})$ molecules/cm^3 of H_2O is required to form silicon dioxide.

Let us imagine the case where the oxide layer t_{ox} is already present on the silicon wafer and the oxidation process is going on. The rate of oxidation (oxidation growth rate) can be expressed in terms of the flux F_3 as

$$F_3 = N_1\left(\frac{dt_{ox}}{dt}\right)$$

And using equations (8) and (12), one can get

$$= \frac{K_s C^*}{1 + \left(\frac{K_s}{h}\right) + \left(\frac{K_s t_{ox}}{D}\right)} \tag{15}$$

As assumed at zero oxidation time, the initial oxide thickness t_{ox} has grown at time t_{i}, so one can write

$$N_1 \int_{t_i}^{t_0} \left[1 + \frac{K_s}{h} + \frac{K_s t_{ox}}{D}\right] t_{ox} = K_s C^* \int_0^t dt \tag{16}$$

The above equation can be written in the form shown below:

$$\frac{t_0 - t_{ox}}{B} + \frac{t_0 - t_{ox}}{B/A} = t \tag{17}$$

$$t_{ox}^2 + At_{ox} = B(t + \tau) \tag{18}$$

$$A = 2D\left(\frac{1}{k_s} + \frac{1}{h}\right) \tag{19}$$

where $$B = \frac{2DC^*}{N_1} \tag{20}$$

and $$\frac{B}{A} = \frac{C^*}{N_1\left(\frac{1}{K_s} + \frac{1}{h}\right)} \cong \frac{C^* K_s}{N_1} \tag{21}$$

A, *B* and *B/A* are constant, the parabolic rate constant and the linear rate constant respectively. Equation (18) can be expressed in a similar form

where $$\tau = \frac{t_{ox} + At_{ox}}{B} \tag{22}$$

In Eq. (20), the oxide thickness τ is expressed in terms of time that corresponds to the oxidation time for a particular process recipe (dry or wet oxidation). For example, to grow the oxide thickness t_{ox}, the time taken by dry oxidation may be significantly higher than the time taken by wet oxidation; hence, the oxidation times will be different.

Equation (18) is in the form of a mixed parabolic-linear equation; hence, this equation is called **mixed linear parabolic**. A typical graph of Eq. (20) is shown in Fig. 4.6.

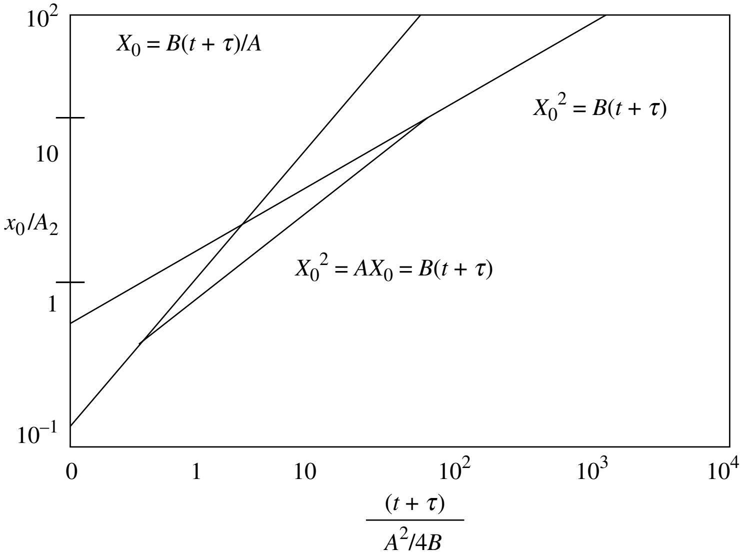


Fig. 4.6 Thermal oxidation rate with oxidation time

To find the oxidation time, Eq. (18), one can write

$$t_{ox} = \frac{-A \pm \sqrt{A^2 + B(t + \tau)}}{2A} \tag{23}$$

$$t_{ox} = \frac{A}{2}\sqrt{1 + \frac{(t + \tau)}{A/4B}} - 1 \tag{24}$$

The linear-parabolic Eq. (18) has two limiting cases. This depends on whether the oxidation time is smaller or greater than $A^2/4B$.

First Case *When, $t >> \tau$*

This means that the oxidation time is appreciably greater than the rate of constants and the initial oxidation time (τ).

The equation then reduces to

$$t_{ox}^2 \approx B(t + \tau)$$

This equation represents parabolic law; hence, B is called the **parabolic rate constant**.

Second Case *When* $(t + \tau) << A^2/4B$

$$t_{ox} = \frac{B}{A}(t + \tau) \tag{25}$$

This equation represents a linear equation; hence, B/A is called the **linear rate constant.**

In the actual oxidation process, oxide thickness is controlled by the oxidation time, keeping all other parameters constant. The rate constant parameters A, B and B/A are different for dry oxidation and wet oxidation. The typical rates of growth of dry and wet oxidation of (100) silicon are shown in Fig. 4.7 and Fig. 4.8 respectively; and the approximate values of these rate constants for dry and wet oxidation process with respect to temperature are listed below in Table 4.1 and Table 4.2 respectively.

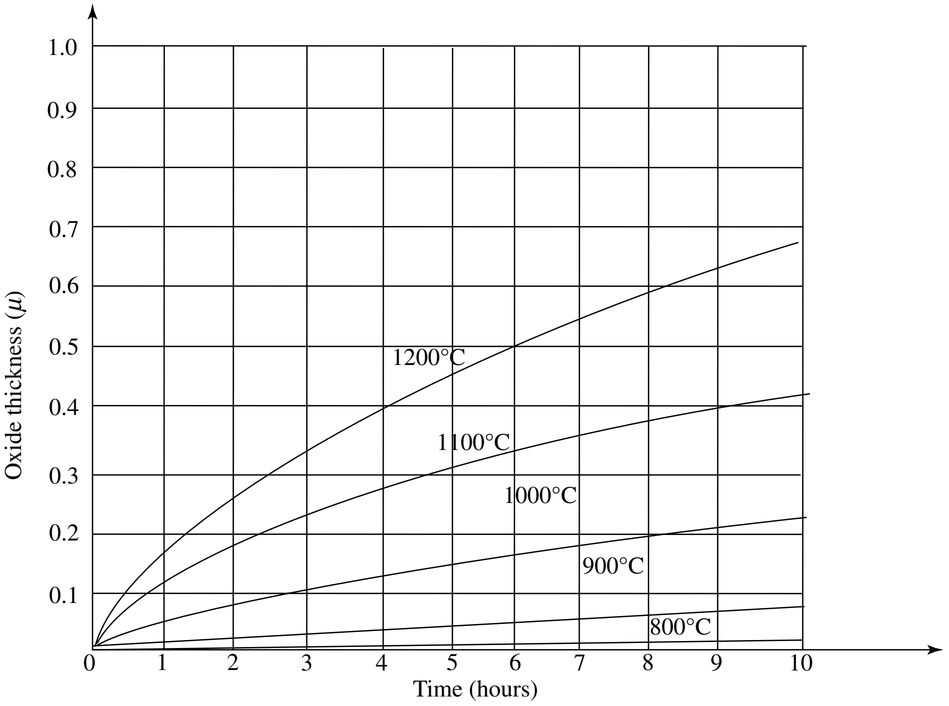


Fig. 4.7 Typical rate of dry oxidation for (100) silicon based on Deal–Grove model

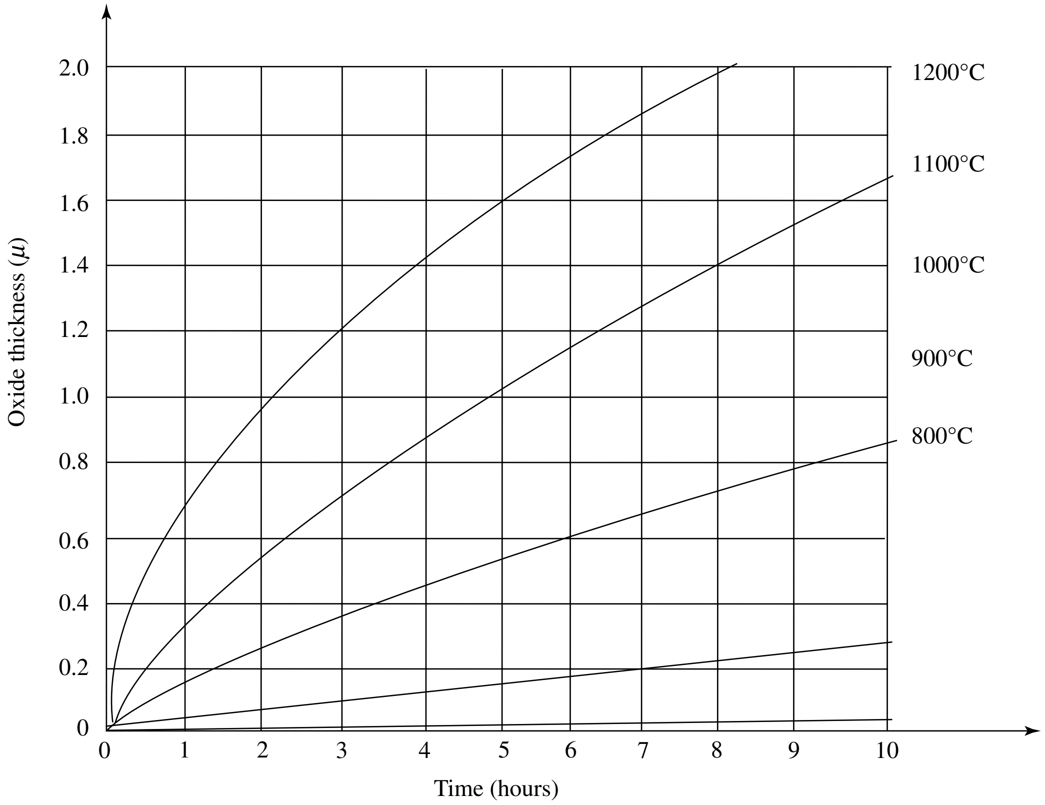


Fig. 4.8 Typical rate of wet oxidation for (100) silicon based on Deal-Grove model

Table 4.1 Approximate rate constants for wet oxidation

Oxidation temperature	Parabolic rate constant		Linear rate constant	
Temperature (°C)	A (μm)	B (μm^2/h)	B/A (μm/h)	τ(h)
1200	0.05	0.72	14.40	0
1100	0.10	0.50	4.60	0
1000	0.23	0.29	1.30	0
920	0.5	0.20	0.41	0

Table 4.2 Approximate rate constants for dry oxidation

Oxidation temperature	Parabolic rate constant		Linear rate constant	
Temperature (°C)	A (μm)	B (μm^2/h)	B/A (μm/h)	τ(h)
1200	0.04	0.05	1.12	0.03
1100	0.09	0.03	0.30	0.08
1000	0.17	0.01	0.07	0.37
920	0.24	0.00	0.02	1.40
800	0.4	0.0	0.0	9.0

Example 4.2

Find out the silicon oxide thickness when (100) silicon wafer is subjected to wet oxidation for 1 hour at 1100°C, followed by dry oxidation for 3 hours at the same oxidation temperature.

Answer Let us consider that the oxidation rate is linear. From the graph, it is seen that the thickness in wet oxidation at 1100°C is around 700 Å; and in dry oxidation, the time taken to grow 250 Å thickness at 1100°C is around 45 minutes.

The total time taken for the oxide thickness growth due to dry oxidation is 3 hours 45 minutes. The total oxide thickness, from the dry oxidation graph, for 3 hours 45 minutes at 1100°C is around 2.10 μm.

The silicon dioxide thickness of the wet oxidation parameter (coordinate) is first converted to the dry oxidation time parameter (coordinate, i.e., τ)

4.5 COMMENT ON THE DEAL-AND-GROVE OXIDATION MODEL

It has been experimentally found that the Deal-and-Grove model matches well with the oxidation temperatures from 700ºC to 1300ºC, but it miserably fails at lower oxidation temperatures. The Deal-and-Grove model fits well with the experimental data between 3000 Å to 20,000 Å oxide thickness. This model also fits well in the range of 0.2 to 1 atmosphere of oxidation pressure. In early days, ICs were made of thick oxide and the oxide was grown at a higher temperature and atmospheric pressure, barring a few exceptional applications. Therefore, the Deal-and-Grove oxidation model was good enough to predict the oxide thickness in those days. But this oxidation model severely suffers, as in recent years, as the transistor geometries have been shrinking as also the oxide thickness (see Chapter 11).

Activation energy (E_a) calculations from the Deal-and-Grove model for the reaction rate constant (K_s) comes out to be around 2 eV, which is close to the experimental data of the activation energy of silicon bond breaking (~1.83 eV/molecule). The diffusivity (D) of oxygen is a function of temperature and is expressed as $D_0\ e^{(-E_a/KT)}$; where D_0 is another constant; thus, the parabolic rate constant (B) is proportional to $e^{(-E_a/KT)}$. From the experimental calculation for dry oxidation, the activation energy (E_a) is found to be ~1.24 eV, which is very close to the activation energy (1.18 eV) of oxygen diffusion into the silica (silicon dioxide). The calculation from the Deal-and-Grove model for wet oxidation shows that the activation energy of 0.71 eV is needed for water to diffuse into the oxide, which is found to be close to the activation energy (0.79 eV) for diffusion of H_2O inside the silica. It is important to be noted that in the context of oxidation rate, the equilibrium constant C^* for H_2O and O_2 is 3×10^{19} cm^3 and 5×10^{16} cm^3 respectively. Hence, the oxidation rate of wet oxidation is much higher than that of dry oxidation. The Henry law was developed initially for atomic diffusion, but the Deal-and-Grove model is now used for molecular diffusion which is justified.

Oxidation rate at temperatures much below 700ºC is significant in today's IC generation and this oxidation rate could not be predicted by the Deal-and-Grove model. It is observed that around 200 Å thick oxide is instantaneously grown when the wafer is inserted in the oxidation furnace even at lower oxidation temperatures. In addition, this model does not give accurate oxidation rate for more than 1 atmospheric pressure and less than 0.2 atmospheric pressure. It is experimentally found that at higher pressure, the oxidation rate is faster; the reason being that the partial oxidation pressure increases with the pressure of oxygen gas and that leads to faster oxide growth rate. The Deal-and-Grove model also does not fit in the mixed oxidation ambient such as, in the presence of the impurities or the dopant. For example, the oxidation rate in the environment of HCl does not match with the experimental data. Furthermore, the oxidation rate is higher in the presence of the dopant or the impurities. It is essential to mention that chlorine-based oxidation is required for a good-quality oxide film, especially for the MOS gate. In addition, the Deal-and-Grove model fails to predict the oxidation rates for different crystal orientations. Furthermore, this model does not fit well for the higher silicon-modulated surface.

The dry oxidation is very sensitive to the presence of water molecules. In dry oxidation, the water molecules should be less than 1 ppm (parts per million). In the presence of water molecules, the oxide gets converted from a bridged to a nonbridged network.

This leads to a porous and less cohesive oxide. The water molecules may come from gas, container, processing room and out of diffusion from oxidation furnace, etc. The presence of a greater number of sodium atoms (10^{22} atoms/cm^3) also increases the oxidation rate to around twice. To get rid of the sodium atoms, the oxidation is done in the presence of chlorine gas or a chlorine-based compound, such as HCl. When HCl is added to dry oxygen in the range of 1 to 5% in volume, an increase of 30% in the oxidation rate is found. The HCl molecules react with oxygen and produce water and chlorine molecules. The released chlorine molecules react with Na and get converted into volatile NaCl that goes out of the oxidation furnace along with the unused oxygen gas. Fortunately, the water molecules produced by the reaction do not affect the quality of the oxide as only a few molecules are produced. The reduction of the sodium atoms improves the breakdown voltage of the oxide and also reduces the **stacking faults**. The chlorine gas also getters many inorganic impurities and improves the oxide quality. In place of HCl, the chemical trichloroethylene (TEC) was also used. Unfortunately, TEC being a toxic chemical is not used in the present day. In wet oxidation, HCl is not used because of two reasons: it reduces the oxide growth rate and corrodes the equipment.

The oxidation rate increases if the wafer has impurities and defects. This increase of oxidation in the presence of impurities and defects was not predicted by the Deal-and-Grove model of oxidation. It has been mentioned previously that the starting wafers are not fully defect free. The defects also increase due to the high-temperature process steps, high-energy particle strike (ion implantation) and many other IC processing steps. If a large number of defects are present in the wafers, the rate of oxidation increases as much as twice. For instance, the source and drain formation requires high doses of boron and phosphorous ($>10^{16}$ atoms/cm^3). In this order of dopant concentration, the oxidation rate increases significantly. In addition, the dopant profile changes due to dopant redistribution at the oxide/silicon interface. This redistribution of the dopant is defined by the **segregation coefficient** and is generally denoted by *k* or *m* and is defined as

$$K = \frac{\text{Equilibrium concentration of dopant in silicon}}{\text{Equilibrium concentration of dopant in silicon dioxide}}$$

The dopant redistribution occurs because of the difference of dopant solubility in the silicon and silicon dioxide. It has been found that the oxidation rate is faster in the presence of boron than in the presence of phosphorous. The reason being that the boron atoms get sucked by the oxide and create more nonbridged oxide and the oxide becomes porous which leads to more oxygen diffusion through the porous oxide.

4.6 OXIDE CHARGES

It is found that the electrical charges reside in the oxide as well as at the oxide/silicon interface. These charges appear due to different types of material compositions, ionic bond formation and the presence of alkaline ions. These charges not only change the threshold voltage (V_{th}) of the MOS transistor, but also affect the other electrical parameters of the transistor. Therefore, these charges should be either eliminated or brought down to the minimum possible level.

There are four types of electrical charges present in the oxide and the oxide/silicon interface, namely, fixed charge (Q_f), interface trap charge or interface charge (Q_{it}), oxide trap charge (Q_{ot}), and mobile ionic charge (Q_m). These charges are depicted in Fig. 4.9. The unit of these charges is **the number of charges per area**, except for the interface charge, whose unit is defined by the **number of states per unit square per unit electronvolt**.

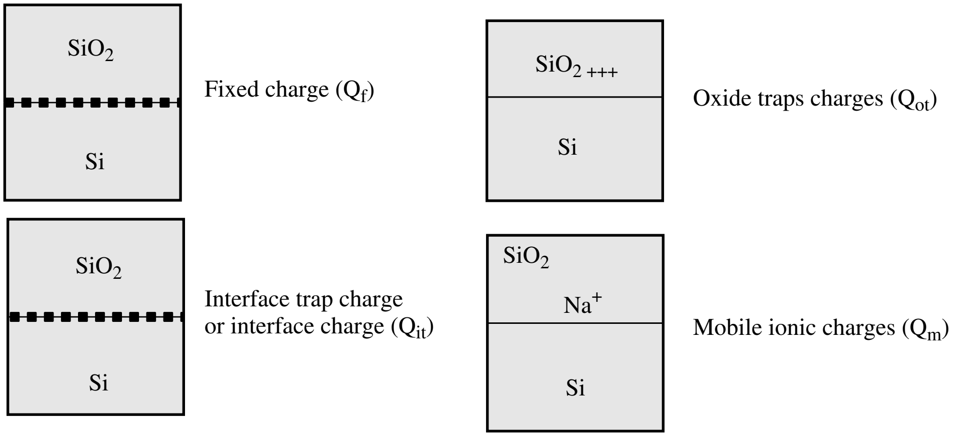

SiO_2
Si
Fixed charge (Q_f)
SiO_2 +++
Si
Oxide traps charges (Q_{ot})
SiO_2
Si
Interface trap charge
or interface charge (Q_{it})
SiO_2
Na^+
Si
Mobile ionic charges (Q_m)

Fig. 4.9 Oxide charges present in oxide and oxide/silicon interface

4.6.1 Fixed Charge (Q_f)

The fixed charges (Q_f) are just like a sheet of positive charges that reside in the oxide around 30 Å above the oxide/silicon interface. Unlike other charges, these charges are fixed and do not change or discharge and remain constant during the transistor operation. For this reason, these charges are called **fixed charges**. The experimental findings suggest that these charges are generated due to the incomplete oxidation of silicon atoms which leads to the formation of ionic silicon atoms in the oxide. The presence of the fixed positive charges pushes the transistor threshold voltage (V_{th}) towards negative voltage, irrespective of the type of MOS. After the wafer is annealed in N_2 or in a mixture of N_2 and H_2 gases at 450°C for 30 minutes, the fixed charge reduces to $\sim 5 \times 10^{10}/cm^2$ for (111) and $1 \times 10^{10}/cm^2$ for (100) crystal orientation. As the (100) wafer orientation charge is 5 times less than the (111) orientation charge, therefore, ICs are preferred to be made on the (100) oriented silicon wafer.

4.6.2 Interface Trap Charge or Interface Charge (Q_{it})

The interface traps charges or interface charges (Q_{it}) reside much closer to the oxide/silicon interface. These charges are generated due to two reasons. These are: due to the silicon lattice periodicity disruption at the oxide/silicon interface and the incomplete silicon bonds. Generally, the incomplete silicon bonds are referred to as **dangling bonds**. These interface charges are similar to that of fixed charges, except that the interface charges may be positive, negative or even neutral. The nature and number of these charges depends on the MOS fabrication process recipe. Hence, to fabricate an IC, one has to choose or optimise the process recipes carefully prior to MOS fabrication. The interface charges are highly concentrated near the valence band (E_v) and the conduction band (E_c) edges in comparison to the middle of the forbidden gap. It has been experimentally found that when the fixed charges increase, the interface charges also increase. In addition, it has also been found that the order of the charge density of the interface charges and the fixed charges are the same. Hence, it is believed that both the charges are generated from common source(s). It has been found that the interface charge density (not fixed charge) creates allowed energy states in the forbidden gap and is expressed in terms of the number of states/cm^2.eV. These interface charges are dependent on the crystal orientation. After the annealing of the wafer in N_2 or in the mixture of N_2 and H_2 gases at 450°C for 30 minutes, the interface charge can be brought down to around $1 \times 10^{10}/cm^2$.eV for (100) orientated silicon crystal.

4.6.3 Oxide Trap Charge (Q_{ot})

The oxide trap charges are located within the oxide. The hole and electron pairs are generated due to the breaking of the oxide bonds due to the exposure of high-energy radiation (particles), such as the X-ray, high electron bombardment during ion-implantation, etc. This high-energy radiation creates electrons and holes. The electrons move inside the silicon, leaving behind holes in the oxide. Therefore, these charges are always positive in nature. These charges are not a function of the crystal orientation, as they remain in the oxide and are also produced there. These charges can be completely eliminated, if the wafer is annealed in N_2 or in a mixture of N_2 and H_2 gases at 450°C for 30 minutes.

4.6.4 Mobile Ionic Charge (Q_m)

The mobile ionic charges reside in the oxide. As these mobile ions move inside the gate oxide, they are called **mobile charges**. The presence of alkaline ions such as Na and K elements are mainly responsible for the generation of these charges. These mobile charges are due to ions; hence, the charges are always positive in nature. The source of generation of these mobile charges is the chemical, the equipment, glassware and other IC processing materials. The mobile charges have a significant effect on the transistor threshold voltage, especially when they move closer to the oxide/silicon interface. These charges are independent of the crystal orientation. The mobile charges can be eliminated, if oxidation is done in the presence of chlorine.

4.7 SILICON OXIDE CHARACTERISATION

It is mentioned under MOS process flow in Chapter 3 that silicon dioxide plays a very vital and crucial role as the transistor structural component in an IC. Therefore, silicon dioxide must have good qualities like less stress, higher refractive index, higher density, higher breakdown, good composition, less pinholes and lower defect densities. Furthermore, the thickness of silicon dioxide should be known accurately prior to MOS transistor fabrication, as it also governs the threshold voltage of the transistor. Hence, the oxide quality should be evaluated thoroughly.

4.7.1 Silicon Oxide Thickness Measurement

There are a number of techniques used for oxide thickness measurement. These techniques can be categorised into three distinct divisions on the basis of measurement principles, namely physical, optical and electrical techniques.

Physical Measurement Techniques

Generally, the physical measurement techniques are destructive in nature, hence, extra wafers are needed to measure the oxide thickness. One of the most versatile physical measurement techniques is the mechanical stylus method. Its principle of measurement is based on the mechanical displacement of the stylus at the oxide/silicon edge, as shown in Fig. 4.10. A sharp vertical step (sharp edge) of the oxide is made using lithography and oxide etching techniques. A flexible stylus (mechanical cantilever beam) is placed over the silicon wafer and it is moved from the oxidised silicon towards the bare silicon (non oxidised silicon). This causes the displacement of the stylus due to oxide height. The stylus measurement technique can measure the oxide (or any other film) thickness up to a few nanometres.

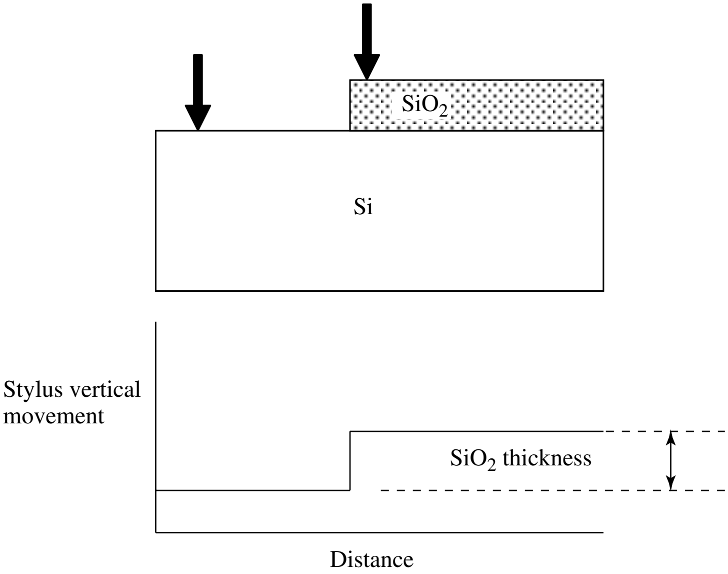


Fig. 4.10 Oxide thickness measurement by mechanical stylus

Optical Measurement Techniques

To measure the oxide thickness or, for that matter, any transparent dielectric film, optical techniques are fast and convenient. Barring a few optical measurement techniques, most of the optical measurement techniques are nondestructive in nature. Generally, two optical principles are used to measure the oxide thickness. These techniques are based on the principle of interference of light and polarisation. These optical techniques are powerful and thickness measurement can be done on line (on the processing wafer); hence, no extra wafer is needed. Unfortunately, these optical techniques need to solve complex mathematical equations, but in the present days, microprocessors are used for this purpose without any loss of time. Furthermore, the precise equipment settings, measurements and calculations—all are done by microprocessor-based systems with great accuracy. In principle, these optical measurement techniques are based on phase information. Therefore, the thickness of the oxide is measured in between the light phase from an angle of 0° to 360°. If the oxide is thicker in terms of 360° angle then one must have prior knowledge of the approximate thickness of the oxide.

Light Interference Technique

The accurate value of oxide thickness is measured using the light interference technique. In this technique, a parallel monochromatic light is projected onto the oxidised wafer. The projected light gets partially reflected from the air/oxide interface and the rest of the light is reflected from the oxide/silicon interface. These reflected lights are superimposed on each other. The light reflected from the oxide/silicon interface travels twice that of the oxide thickness than its counterpart, i.e. the light reflected from the air/oxide interface, as shown in Fig. 4.11. Thus, a phase difference is introduced between these lights. When these lights are superimposed, fringes are formed. The maxima and minima of light interference (fringes) can be expressed as

$$\lambda = \frac{2n_i\, t_{ox} \cos\phi}{m}$$

where, λ is the monochromatic light wavelength, t_{ox} is the oxide thickness, n_i is the refractive index of the oxide, m is the order of the fringe (m = 1, 2, 3… for maxima and m = 1/2, 3/2, 5/2 … for minima),

and $$\phi = \sin^{-1}\left[\frac{n_o \sin\alpha}{n_i}\right]$$

where angle α is angle of the incident light.

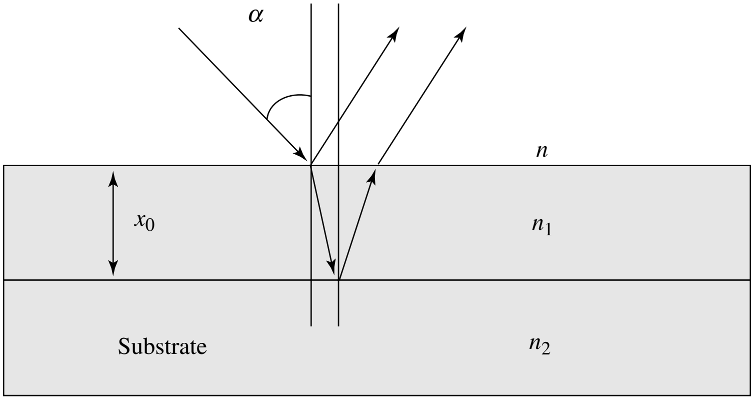


Fig. 4.11 Optical method of oxide thickness measurement

Keeping the incident angle α constant, and changing the wavelength λ of the light, the positions of the maxima (bright) and minima (dark) fringes can be determined. By counting the numbers of fringes from a reference point, one can calculate the oxide thickness using the above-mentioned equations. One can also measure the oxide thickness by keeping the wavelength constant and varying the angle of incidence of light. If the oxide thickness is measured using the short light wavelength, one can even measure a fraction of nanometre of the oxide thickness.

Look-up Techniques

A wild guess of the oxide (or nitride) thickness can be made by looking at the colour of the oxidised wafer. When a uniform oxidised wafer is seen under fluorescent light, a bright and uniform colour is visible. This phenomenon is a simple case of the interference of light. When light falls on the oxide film, a part of light is reflected back from the air/ oxide interface and the remaining part passes through the oxide; and then gets reflected from the oxide/silicon interface. When these two lights are superimposed on each other, a uniform colour appears across the wafer, because of the uniform oxide thickness. One can see this uniform single colour by naked eye. The change of colour depends on the oxide thickness, as shown in Table 4.3.

Table 4.3 Change of colour w.r.t. oxide thickness

Colour	Silicon dioxide (A^0)	Silicon nitride (A^0)
Silver	270	200
Brown	530	400
Yellow-brown	730	550
Red	970	730
Deep blue	1000	770
Blue	1200	930
Very pale blue	1500	1100
Silver	1600	1200
Light yellow	1700	1300
Yellow	2000	1500
Orange red	2400	1800
Red	2500	1900
Dark red	2800	2100
Blue	3100	2300
Blue-green	300	2500
Light green	3700	2800
Orange-yellow	4000	3000
Red	4400	3300

Ellipsometric Technique

To measure the oxide thickness, the principle of polarisation of light is exploited to get an optical measurement technique called the **ellipsometric technique**. When light passes through the oxide, it changes its degree (or state) of polarisation. This degree of polarisation is related to the phase of light and that in turn is related to the oxide thickness. The theoretical derivation of the ellipsometer equations is very complicated and involves many optical parameters of the materials. The short form of the ellipsometer equation can be expressed as

$$\tan \psi \cdot \exp(j\Delta) = \frac{R_p}{R_s}$$

where Δ is the phase, ψ is the amplitude of the two orthogonal components of the polarised light, R_p and R_s are the reflections of the two orthogonal (polarised) components p and s, and are functions of the optical parameters such as the refractive index, angle of incidence as shown in Fig. 4.12. These reflections of light are orthogonally polarised to each other and they lie between the angles of 0° and 90°. The scrip Δ represents the change in phase

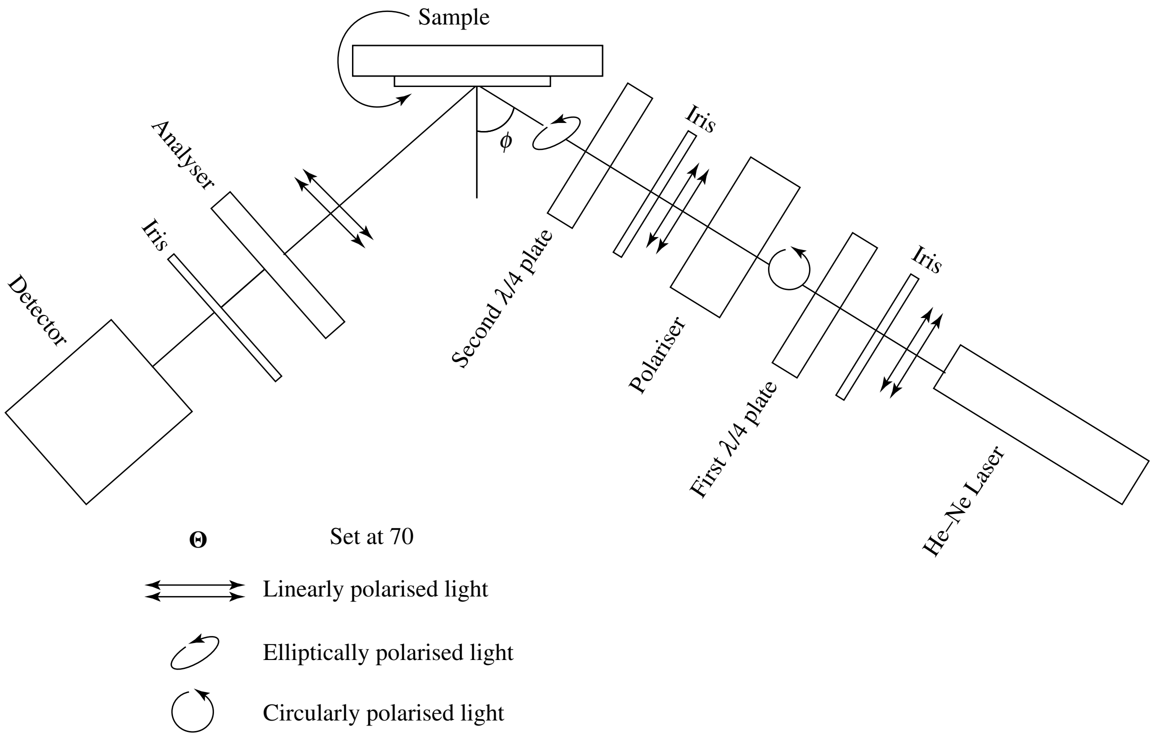


Fig. 4.12 Ellipsometer method for oxide thickness measurement

during the reflection of light and it lies between 0° and 360°. These two parameters ψ and Δ are experimentally measured by the ellipsometer. After measuring ψ and Δ and using Snell's law with Fresnel's coefficient, the thickness of the oxide is calculated. The thickness of the oxide film can be measured in the order of 1 nanometre. On the other hand, if the oxide thickness is known then the unknown value of the refractive index of the oxide film can be obtained.

Other physical measurement systems are Scanning Tunnelling Microscope (STM), Atomic Force Microscope (AFM), Scanning Electron Microscope (SEM) and their derivatives. These methods take longer measurement time, are costly, and need elaborate arrangements, but give very accurate oxide thickness; hence these measurement techniques are not used in routine oxide thickness measurement.

Electrical Technique

The oxide thickness can be measured by the electrical technique. Small metal film dots, around 1 millimetre in diameter, are deposited on the grown oxide and a global metal film is deposited on the back of the silicon wafer. Thereafter, electrical potentials are applied on one of these metal dots and the back of the wafer. This results in the formation of a capacitance across the oxide. This structure is called the **MOS capacitor** as shown in Fig. 4.13. The oxide capacitance of the MOS structure can be expressed as

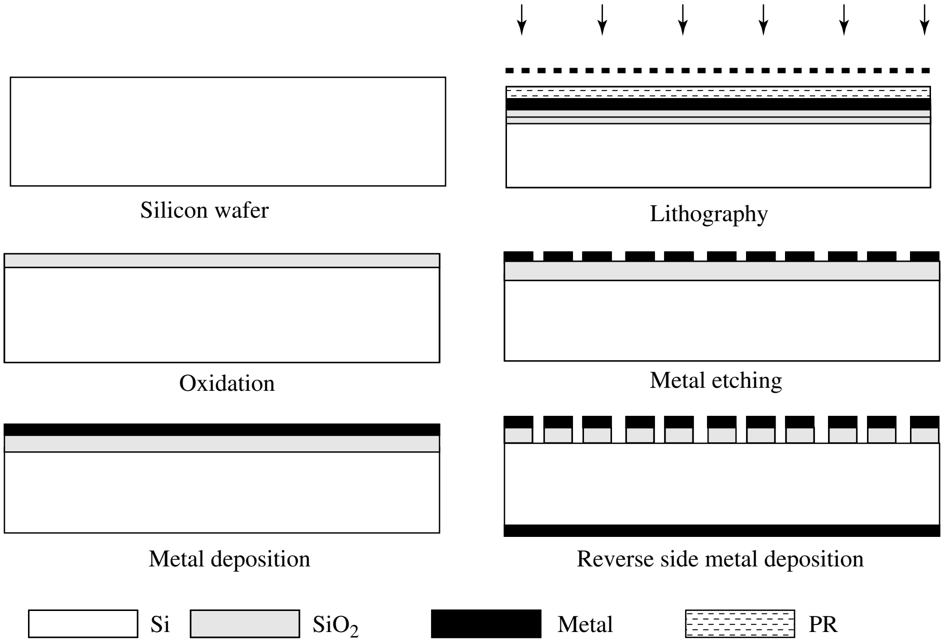


Fig. 4.13 Fabrication of MOS capacitor

$$C_0 = \frac{\varepsilon_0 \, \varepsilon_s}{t_{ox}}$$

where ε_0 is the silicon dioxide dielectric constant and ε_s is the permittivity.

This technique of oxide thickness measurement does not give accurate results; the reason being that the value of the oxide capacitance is low and that may lead to erroneous capacitance measurement. In addition, the oxide capacitance depends on the oxide quality and many other electrical parameters. Hence, this capacitive measurement technique is generally not used for oxide thickness measurement, but it reveals many useful parameters of the MOS transistor, especially the MOS electrical parameters. Therefore, it is essential to describe this technique.

4.8 ELECTRICAL CHARACTERISATION OF MOS CAPACITANCE

The MOS transistor is an electrical device; therefore, any electrical information during the IC fabrication process is most essential. Apart from the thickness information by the electrical Characterisations, the MOS capacitor gives other valuable electrical information such as the oxide charges, doping concentration, threshold voltage, and other electrical parameters. Through this electrical Characterisation of the MOS capacitor, the IC fabrication process recipes are optimised prior to the actual IC fabrication. One of the frequently used electrical Characterisations of MOS capacitor is the capacitance–voltage measurement technique.

4.8.1 Electrical Characterisation of MOS Capacitor by the Capacitance-Voltage (*C-V*) Measurement Technique

In the *C-V* measurement technique, the oxide charges are Characterised by studying the change of capacitance with voltage across a MOS capacitor. The MOS structure is realised by depositing circular metal dots of around 1 mm diameter (may be smaller) on the gate oxide grown onto the wafer, as shown in Fig. 4.13. These metal dots can be made using two techniques. The first technique is a simple one to make metal dots on the wafer and is called the **shadow metal deposition technique**. In this technique, a circular perforated hole of ~1 mm metal sheet is kept over the gate oxidised wafer, and the metal is deposited through the holes on the wafer. In the second technique, the gate oxide is metallized on the entire wafer and then circular dots are made by the lithography and metal etching techniques. To make the MOS capacitor, the back of the wafer is also metallised. The typical schematic diagram of the *C*-*V* measurement set-up is shown in Fig. 4.14. In the *C*-*V* measurement technique, an ac signal of 1 MHz and around 50 mV is superimposed on the variable dc voltage and then the voltage is fed to one of the metal dots. The voltage that is applied on the gate is called the **gate voltage**. When the back of the wafer is connected to the ground, a MOS capacitor is formed across the gate oxide.

Let us consider that the MOS capacitor structure is made on an *N*-type silicon wafer and a positive dc voltage is applied onto a metal dot, and no ac signal is superimposed; then, the free electrons in the wafer will be attracted towards the gate oxide to compensate the gate electric field. This will result in an accumulation of the electrons just below the oxide/silicon interface. This type of capacitor is called the **accumulation capacitor**. The accumulation capacitor will depend on the thickness of the oxide, its dielectric constant and the gate voltage. Let us consider that the maximum accumulation capacitance is C_o. If the positive dc gate voltage is swept with slow sweep rate towards the negative voltage, the

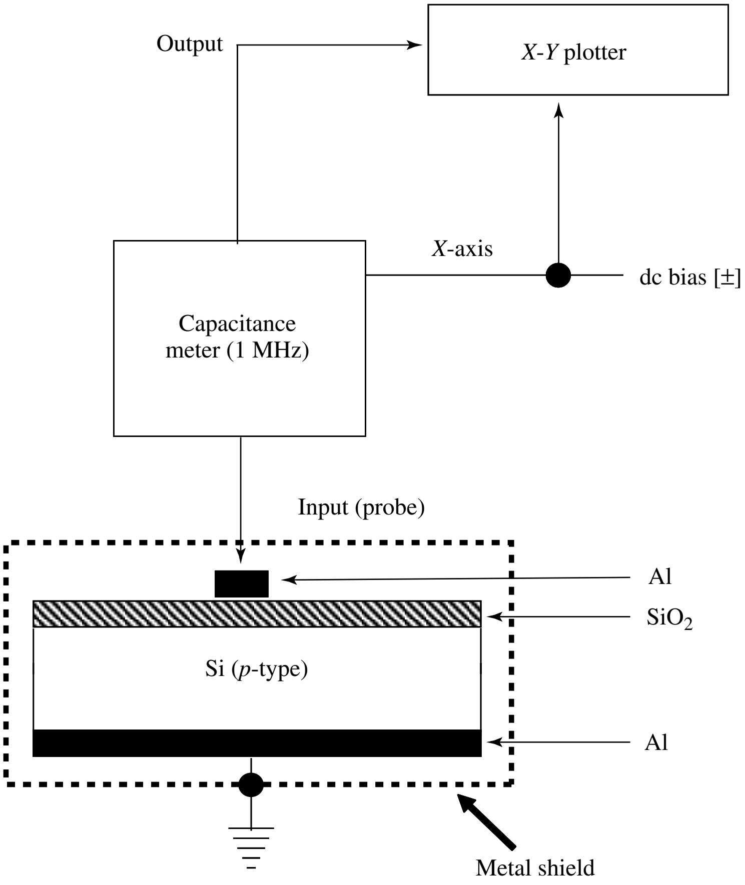


Fig. 4.14 A typical high-frequency *C*-*V* curve measurement system

accumulation of electrons will gradually reduce and a depletion layer will start forming. The depletion layer will keep on increasing with the gate voltage till the depletion capacitance C_s is pegged to fixed depletion width. At this point, the total capacitance will go down to the minimum capacitance value $\mathbf{(C_o + C_s)_{Min}}$. This condition is called the **threshold voltage** of a MOS transistor. Beyond this gate voltage, the holes will get attracted from the bulk silicon and start getting collected under the gate oxide; and finally, an inversion stage will reach. At this condition, the gate voltage is balanced by the inversion layer. Hence, the MOS capacitance will come back to the original accumulation capacitance C_o. This inversion layer under the gate oxide acts as a MOS transistor channel in between the source and the drain, if the source and the drain are formed at the ends of the inversion layer. A typical *C-V* curve is shown in Fig. 4.15.

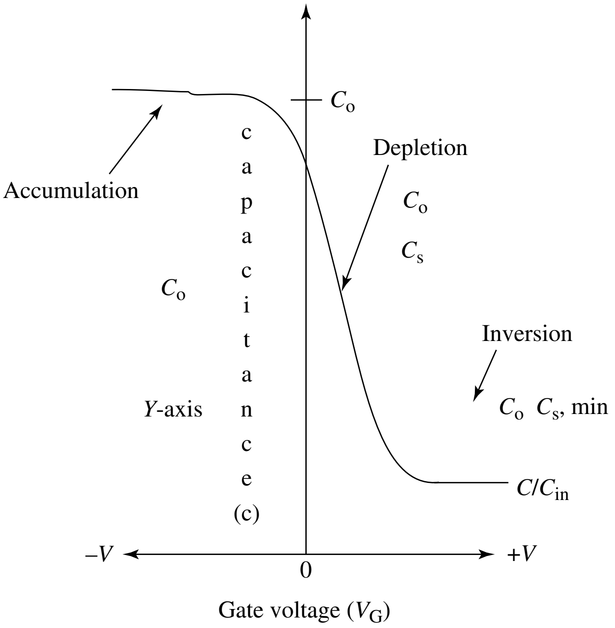


Fig. 4.15 A typical high-frequency *C-V* curve

As mentioned before, there are four types of charges present in the oxide and the oxide/silicon interface. To Characterise the oxide and the interface charges, the *C-V* measurements act as a simple and powerful technique. There are two *C-V* measurement techniques, namely high-frequency *C-V* and low-frequency *C-V*, used for MOS capacitor Characterisation. In the first technique, i.e. high-frequency *C-V* measurement technique, the ac signal is superimposed on the dc in the range of 100 Hz to 1 MHz. In the second technique, a low ac frequency in the range of 10 Hz to 100 Hz is superimposed on the dc voltage. Generally, the sweep speed of the dc voltage ranges from 1 to 10 volts/second from negative to positive voltage for *P*-type silicon, whereas for *N*-type silicon, the dc sweep voltage is from positive to negative. These two *C-V* measurement techniques reveal a large number of electrical parameters which helps in MOS transistor fabrication.

4.9 CASE I: NON-IDEAL CASE

4.9.1 High-Frequency *C-V* Measurement Technique

Let us consider the non-ideal case where no charges are present in the oxide and the oxide/silicon interface. Generally, high-frequency *C-V* measurements are carried out by 1 MHz frequency superimposed on the dc gate voltage, which is applied onto the MOS gate and thereafter, the dc voltage is swept from positive to negative or vice-versa according to the starting silicon wafer. For instance, if the MOS capacitor is made on the *N*-type wafer then the dc voltage sweep is carried out from the positive to negative potential. The gate being a positive voltage, the electrons get accumulated below the oxide/silicon interface and a maximum accumulation (oxide capacitor) capacitance C_o is formed. The value of the accumulation capacitance C_o depends on the oxide thickness and oxide dielectric constant. When the gate voltage starts sweeping towards the negative voltage then the accumulation layer vanishes and a depletion layer is formed, and the capacitance reaches to its minimum value $(C_o + C_s)_{\text{min}}$. Till this stage, there is no effect of the 1 MHz ac signal frequency on the MOS capacitor. When the gate voltage increases further, the holes present in the silicon bulk start moving towards the gate to balance the gate voltage. If the voltage of the ac frequency cycle changes faster as compared to the hole mobility, the holes start oscillating at their mean position in response to the ac signal voltage and the minimum capacitance remains at the fixed level. The *C-V* curves of high ac frequency measurement set up are shown in Fig. 4.15.

4.9.2 Low-Frequency *C-V* Measurement Technique

In the low-frequency *C-V* measurement, superimposition of the ac signal frequency is kept in between 10 Hz and 100 Hz. When the gate voltage is swept from positive to negative, an accumulation condition is reached and the depletion layer pegs to its minimum width. Till this stage, the *C-V* curve remains similar to the high-frequency *C-V* curve; thereafter, it starts to defer from the high-frequency *C-V* curve. At low frequency, the gate voltage modulation (due to ac cycle) is slow as compared to the hole mobility. Hence, holes have sufficient time to move at the oxide/silicon interface and they form the inversion layer resulting in an inversion capacitance C_o. The value of the inversion capacitance is the same as the accumulation capacitance C_o, i.e. the oxide capacitance. It is essential to mention that as the inversion condition is reached, the depletion layer is pegged to the maximum width with an increase in negative gate voltage. A typical *C-V* curve is shown in Fig. 4.16.

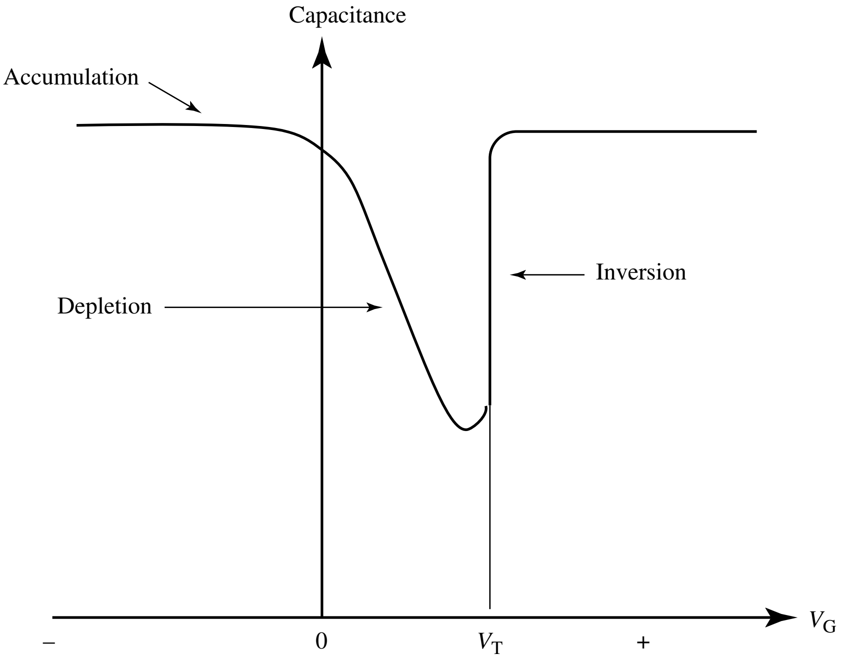


Fig. 4.16 Typical low-frequency *C-V* of *N*-type silicon

4.10 CASE II: IDEAL CASE

In the ideal case, the oxide charges are invariably present in the oxide and the oxide/silicon interface. These oxide charges are one of the main reasons of shifting of the threshold voltage of the MOS transistor. As the threshold voltage is a very important electrical parameter of the MOS transistor, so it is essential to measure the threshold voltage accurately prior to MOS fabrication. An easy way to measure the threshold voltage of a MOS transistor is through the high-frequency *C-V* measurement. If the threshold voltage is not as desired then the transistor's threshold voltage is adjusted during MOS transistor fabrication. This fabrication process step is called **threshold voltage adjustment**, as described briefly in Chapter 3 and further explained in detail under ion implantation in Chapter 9.

4.11 PROCEDURE OF OXIDE AND OXIDE/SILICON INTERFACE CHARGE MEASUREMENTS

The typical *C-V* curves of the MOS capacitor on *P*-type silicon in the presence of oxide charges are shown in Fig. 4.17. If there are fixed charges present in the oxide, say which are positive in nature, the *C-V* curve will horizontally shift towards negative voltage. The total amount of these fixed charges can be estimated by the horizontal voltage shift in the *C-V* curve multiplied by the accumulation capacitance (C_o). Apart from the fixed charges, the trap charges also possess positive charges and the presence of this positive charge threshold voltage also shifts the *C-V* curve towards more negative voltage. Fortunately, trap charges vanish after wafer annealing, therefore, one can differentiate trap charges from fixed charges. The interface trap charges make the slope of the *C-V* curve gentle, as shown in Fig. 4.17c. The shift of the *C-V* curve from the ideal *C-V* curve depends on the type of trap charges (either positive or negative voltage axis) present in the oxide/silicon interface. From the voltage shift and slope of the *C-V* curves, the interface charges are estimated.

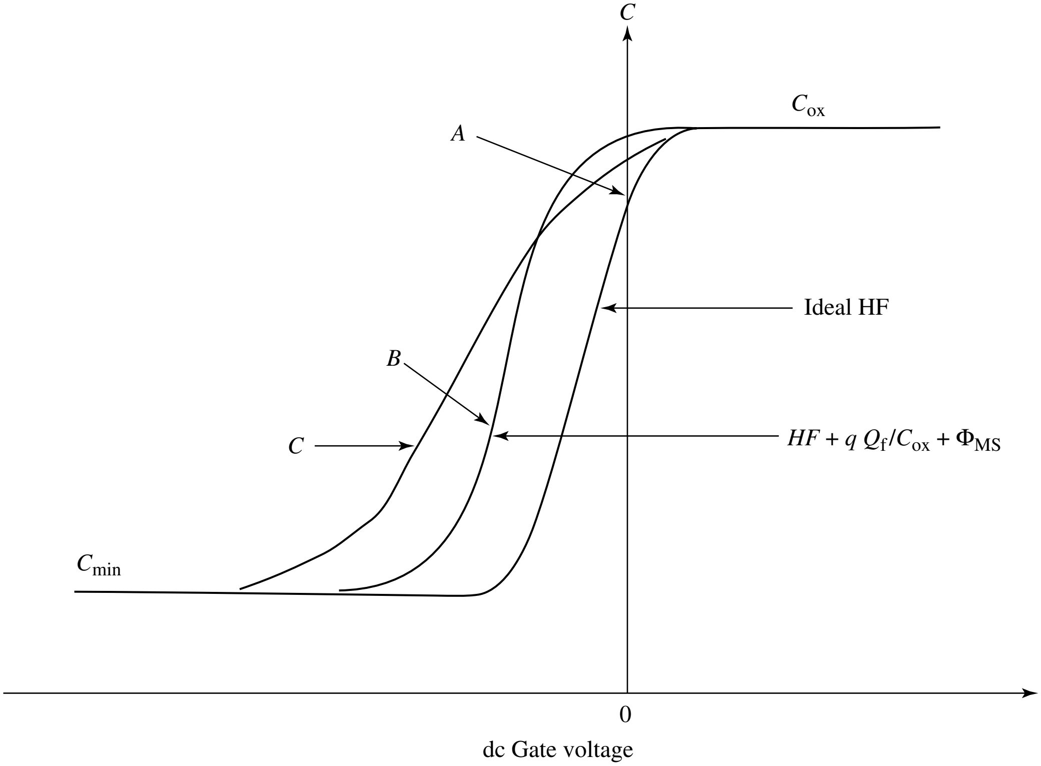


Fig. 4.17 *C-V* curves of *P*-type silicon in presence of oxide charges

The mobile charges are mainly due to the presence of sodium and potassium ions in the oxide and they manifest as positive charges. These mobile charges are estimated by the *C-V* curves with a special technique called the **bias-heat stress technique** as shown in Fig. 4.17a. In the bias-heat stress technique, the *C-V* curves are taken at different temperatures at two bias conditions of the MOS capacitor. In this technique, normal *C-V* measurements are carried out and then the MOS capacitor is heated up to 250°C under a positive bias of 30 volts for about 30 minutes. Thereafter, the MOS capacitor is cooled down to room temperature and again *C-V* measurement is taken without removing the positive gate bias. At the elevated temperature, the mobile ions move towards the oxide/silicon interface in response to the positive gate voltage and remain there. Then, the MOS capacitor is again heated up to 250°C with a negative gate bias of 30 volts for 30 minutes. Thereafter, the MOS capacitor is cooled down to room temperature and *C-V* measurement is taken without removing the gate bias. At this elevated temperature and negative gate voltage, the mobile ions get attracted and therefore, get collected at the top of the metal surface. This causes a shift in the *C-V* curve, as shown in Fig. 4.17b. By analyzing the shift in *C-V* curves, the mobile charges are estimated.

4.12 OXIDE BREAKDOWN MEASUREMENTS

The silicon dioxide breakdown can be measured with the modified *C-V* measurement technique. Here, the current-voltage (*I-V*) of the oxide is measured on the MOS capacitor rather than the *C-V* measurement. In this technique, current passing through the oxide with the gate voltage is measured. Initially, there is no current (except the leakage current) flowing through the oxide when low voltage is applied. When the voltage (electric field) exceeds a certain value, the oxide breaks down that causes a drastic current flow across the oxide. Generally, the electrical breakdown of a good-quality oxide is in the range of $5–10 \times 10^6$ mV/cm. If there are defects in the oxide then the oxide breakdown takes place at an earlier stage. By mapping the breakdown voltage across the wafer, one can find out the defect density in the oxide film.

4.13 OTHER SILICON OXIDATION TECHNIQUES

Silicon dioxide can be grown by anodisation and plasma techniques. These techniques are to get good thin oxides. One of the advantages of these techniques is the low-temperature process, which is much needed for the recent (ULSI) MOS fabrication technique. The anodisation technique is not used because it is not compatible with IC fabrication. Details of silicon oxidation by plasma technique are discussed in Chapter 11.

4.13.1 Silicon Oxidation by Anodisation Technique

In the anodisation technique, the silicon wafer is oxidised in an electrolytic cell, where the wafer is connected to the anode, and the cathode is connected to a noble metal (preferably platinum metal) as shown in Fig. 4.18. In the electrolyte process, an electrochemical potential is developed in between the silicon and the electrolytic solution. To establish the equilibrium of charges, electrons come out from the silicon surface leaving the holes on the silicon surface. The overall reaction can be expressed as

$$Si + 2h^+ + 2H_2O \rightarrow SiO_2 + 2H^+ + H_2$$

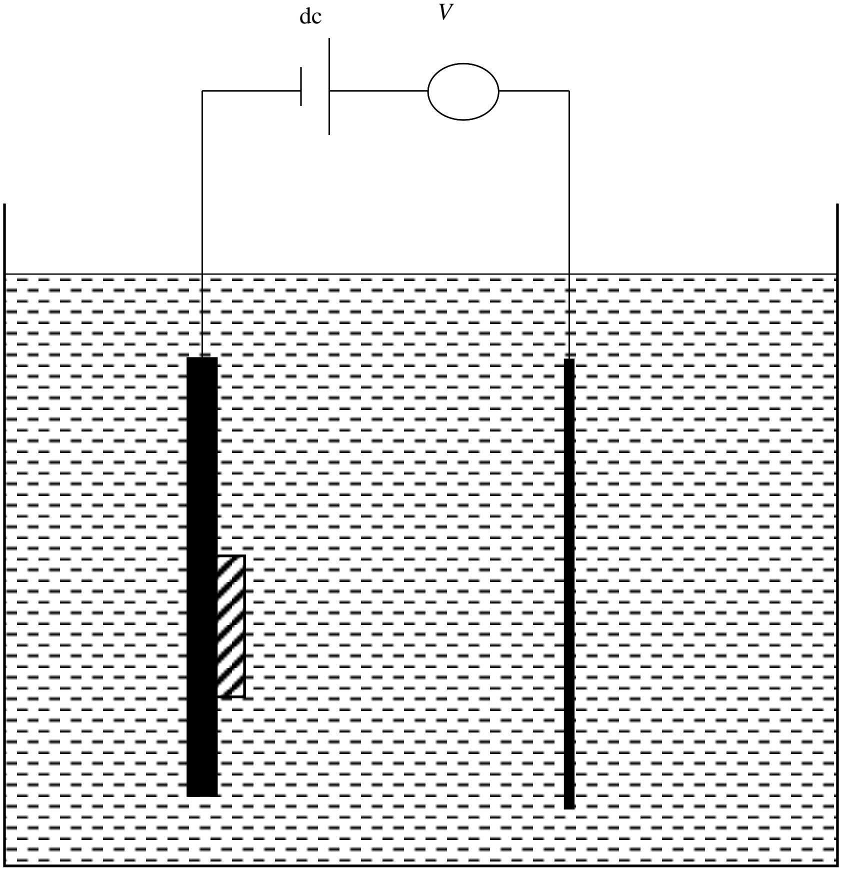


Fig. 4.18 Silicon oxidation by anodisation technique

Some of the electrolytic solutions used for silicon anodisation are

1. 0.1 $MH_3BO_3 + Na_2B_4O_7 + H_2O$
2. $H_3PO_4 + H_2O$
3. N-methylacetamide (NMA) + KNO_3
4. Ethylene Glycol + $KNO_3 + H_2O$
5. Tetrahydrofuryl Alcohol (THFA) + NH_4NO_3

Out of these electrolytes, ethylene glycol gives a uniform oxide, better reproducibility, higher purity and relatively denser oxide.

The anodic oxide is inferior to the thermally grown oxide. It has a large number of pinholes, bad oxide/silicon interface and less density. In addition, anodic oxidation cannot be batch processed and a thick oxide cannot be grown. Furthermore, anodic oxidation is not compatible with the IC fabrication process; therefore, this technique is not used for IC fabrication. On the other hand, anodic oxidation is carried out at room temperature and the rate of oxidation is independent of the crystal orientation of the doped wafer and is in the range of 10^{15} to 10^{22} atoms/cm^2.

4.13.2 Silicon Oxidation by Plasma Oxidation Technique

Silicon can be oxidised by the electrical discharge (plasma condition) technique, as shown in Fig. 4.19. Two separate parallel-plate electrodes are housed in a vacuum chamber. The silicon wafers are kept on the negative electrode and a high dc or RF voltage is applied between the electrodes in the presence of oxygen gas at low pressure. At high voltage, electrical discharge (plasma) takes place and oxygen species are formed. In the plasma condition, the oxygen species react with the silicon atoms and silicon oxide is formed. The growth mechanism of plasma oxidation is not yet known. It is not clear whether the oxygen species go inside the silicon dioxide film or the silicon atoms move towards the oxide/silicon interface and then react with the oxygen species. The plasma oxidation has the advantages of low-temperature deposition and it is also a clean oxidation process. With the plasma oxidation process, a thick as well as good-quality oxide can be obtained, but it suffers from non-uniformity of oxide thickness across the wafer. Details of plasma oxidation are given in Chapter 11.

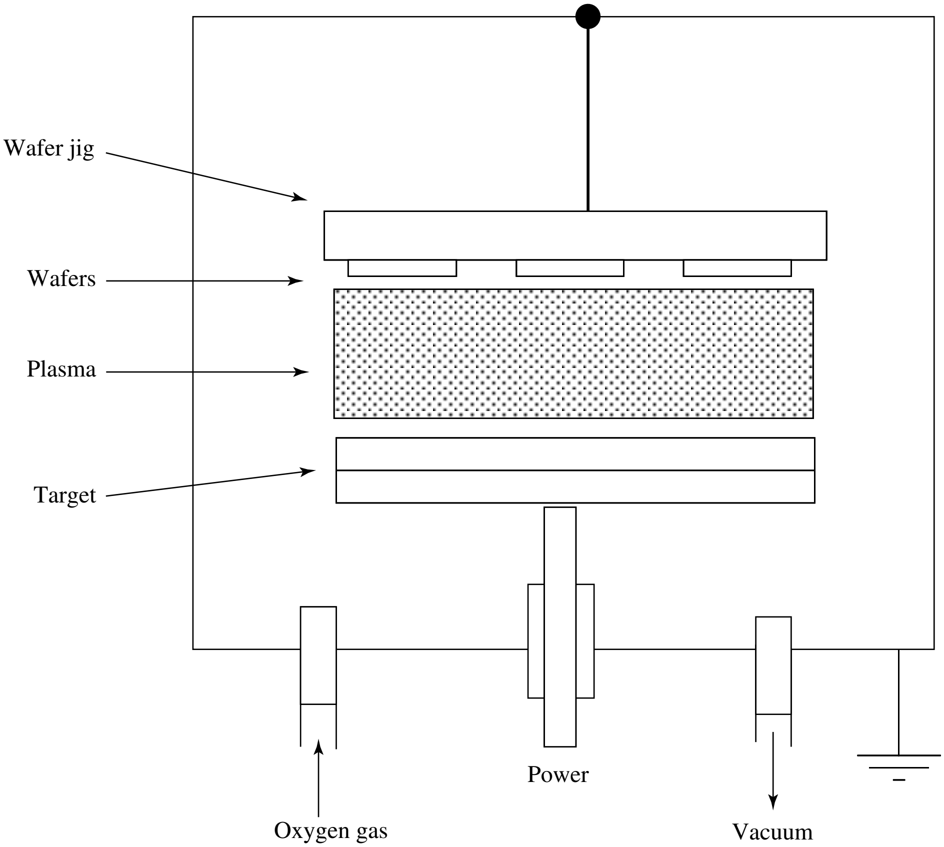


Fig. 4.19 Silicon oxidation by plasma technique

Summary

Silicon dioxide is used as a gate dielectric besides being used for electrical isolation, surface planarisation, as a barrier for impurities, for wafer passivation and as a pad oxide for nitride deposition. Therefore, silicon dioxide must have certain qualities such as lower stress, high refractive index, high density, higher breakdown, good composition, less pinholes, lower defect density and amorphousness. These qualities of the oxide can be easily achieved when silicon is heated in a furnace at high temperature in the presence of pure oxygen gas. The wet oxidation process is used for a thicker oxide for silicon oxidation, but the oxide quality is not good. There are other two techniques used for oxidation, namely the anodisation technique and plasma oxidation. The anodisation technique is not used in IC fabrication; and in the plasma oxidation technique, the silicon is oxidised in electrical discharge (plasma) condition. This technique has useful application in IC fabrication because of low-temperature processes.

A reasonably good oxidation model was developed by Deal-and-Grove, in the early 1960s, but it does not satisfy many oxidation conditions. In spite of that, the Deal-and-Grove model is used till today on case basis, and modified if required.

There are four types of electrical charges present in the oxide and the oxide/silicon interface, namely, fixed charge (Q_f), interface trap charge or interface charge (Q_{it}), oxide trap charge (Q_{ot}), and mobile ionic charge (Q_m). Out of these, Q_f and Q_{it} are inherently generated during silicon oxidation and the other two charges can be got rid off by taking proper care and by annealing.

There are a number of techniques used for oxide thickness measurement. These techniques can be categorised in three distinct categories on the basis of measurement principles, namely physical, optical and electrical techniques. Out of these, optical measurement is mostly used.

The *C-V* and *I-V* are essential to characterise the MOS capacitor prior to bulk IC fabrication. These measurements tell us about the process quality, threshold voltage, oxide charges and many other semiconductor electrical parameters.

References

- J D Plummer, M Deal and P B Griffin; *Silicon Fundamental Technology: Fundamentals, Practice and Modeling*, Prentice Hall, 2000
- S M Sze; *VLSI Technology*, Second Edition, McGraw-Hill, 1988
- S K Gandhi; *VLSI Fabrication Principles*, Second Edition, Wiley, 1994
- D Nagchoudhuri; *Principles of Microelectronic Technology*, Wheeler, 1998
- S A Campbell; *The Science and Engineering of Microelectronic Fabrication*, Oxford University Press, 1996
- S Grove; *Physics and Technology of Semiconductor*, John Wiley & Sons, 1967

Multiple-Choice Questions

4.1 Which type of oxide film is preferred for MOS transistor fabrication?
(a) Amorphous (b) Crystalline (c) Polycrystalline.

4.2 Oxidation rate is faster in which orientation?
(a) (1,0,0) (b) (1,1,0) (c) (1,1,1).

4.3 Which rate constant parameter is responsible for fast oxidation?
(a) Parabolic rate constant (b) Linear rate constant
(c) None of these two parameters

4.4 Fast wet oxidation rate is due to
(a) water diffusion in the grown oxide
(b) concentration of water in the grown oxide
(c) none of these parameters.

4.5 Colour of the oxide is seen due to
(a) interference of light (b)diffraction of light
(c) polarisation of light

4.6 Low-frequency *C-V* measurement is generally done at
(a) kHz (b) MHz (c) GHz

4.7 Where do the mobile charges reside?
(a) In the oxide (b) In the oxide/silicon interface (c) Inside the silicon

4.8 Where do the oxide trap charges reside?
(a) In the oxide (b) Close to the oxide/silicon interface(c) Inside the silicon.

4.9 Where do the fixed charges reside?
(a) In the oxide (b) Close to the oxide/silicon interface(c) Inside the silicon.

4.10 The boron silicate glass melts at around
(a) 950°C (b) 1020°C (c) 1050°C

Descriptive Problems

4.1 Why is dry oxidation better than wet oxidation?

4.2. Why is it that the initial oxidation is linear and it becomes parabolic thereafter?

4.3 How are the interface charges measured in the presence of fixed charges by the capacitance–voltage measurement technique?

4.4 A silicon wafer of *N*-type, (100), 8–10 Ω-cm is oxidised at 1100°C for 230 minutes in wet oxidation. Calculate the oxide thickness.

4.5 What is the procedure to grow the MOS gate oxide? If a silicon wafer with specifications: *N*-type, (100), 8–10 Ω-cm is oxidised at 1100°C for 800 Å and 100 Å thick gate oxidation, calculate the oxidation time.

4.6 1μ thick silicon is oxidised by wet oxidation; thereafter 1000 Å oxide is removed by etching. How much time will it take to oxidise the silicon to again reach 1 μ thickness of silicon dioxide?

4.7 Calculate the gate oxide thickness of a MOS transistor for a 1-volt threshold voltage with surface charge Q_S.

CHAPTER 5 Mask

5.1 INTRODUCTION

MOS transistors are made of localised diffused areas below the silicon surface and stacks of localised film layers over the silicon surface. Many layers of these localised diffused areas and films are made by transferring prefabricated opaque (black) and transparent patterns made on a glass plate. This opaque and transparent pattern on a glass plate is called the **mask**, as shown in Fig. 5.1. These localised patterns of the mask are translated on the wafer by lithography and etching

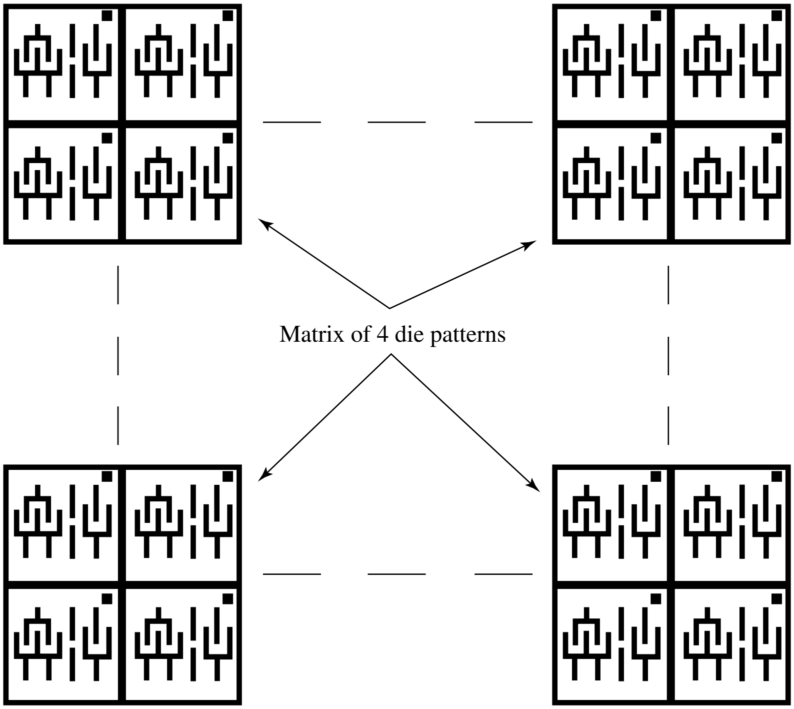


Fig. 5.1 Mask

processes. A brief description of the lithography and etching processes has been given in Chapter 3, and the details of these processes are covered in Chapters 6 and 7 respectively.

The mask patterns are made of photographic emulsion on the photographic plate or metal film on a transparent glass plate. An individual mask contains a matrix of identical black and transparent patterns called the **die patterns**, and the area occupied by the patterns is called the **die**. Each die pattern contains three important features, namely **IC patterns**, **alignment marks** and **scribe tracks** as shown in Fig. 5.2. The alignment marks are used to align a particular level of the mask with respect to other levels of the mask, during the lithography process. These mask alignment marks are made (patterned) at one corner of the die. Generally, the die patterns are surrounded by straight opaque or transparent lines (patterns) from all sides and these lines are called scribe tracks. Finally, these scribe tracks are used for separating one die from another (i.e. one IC from another IC) by cutting (scribing) the wafer. For this reason, these lines are called scribe tracks as shown in Fig. 5.2. It is essential to mention that the die is made either in square or rectangular patterns as they are easy to generate. Thus, the scribe tracks are always in straight lines in the *x-y* direction, as shown in Fig. 5.1. Another use of the scribe track is the alignment of the first mask with respect to the wafer cut, as shown in Fig. 5.3. The process of mask-to-wafer alignment is explained under photolithography in Chapter 6. This step of lithography

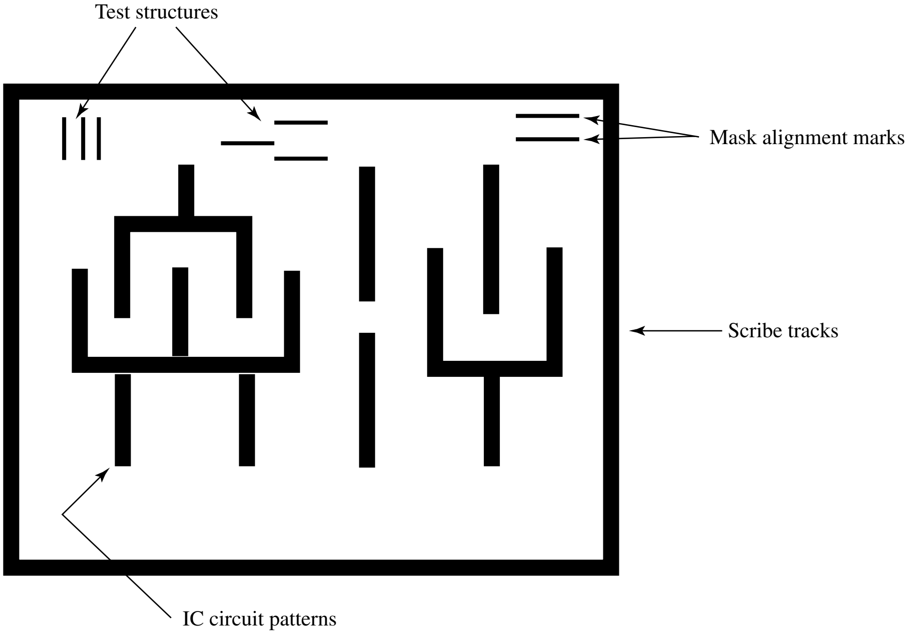


Fig. 5.2 Mask containing IC circuit patterns, mask alignment marks and scribe tracks

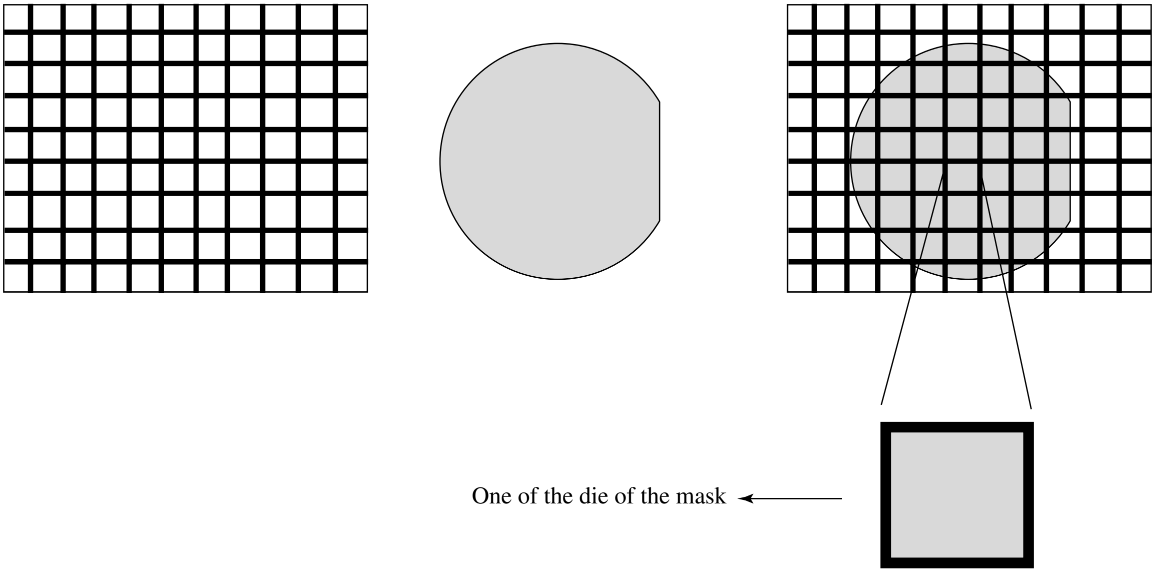


Fig. 5.3 The first mask is generally an active mask aligned with the cut of the wafer with respect to the scribe track

ensures that the MOS transistors are always made on the (100) plane and perpendicular to the (110) orientation, as the oxide and oxide/silicon interface charges are less on the (100) plane and the IC can be cut or broken in the (110) orientation. The mask may also contain a few test patterns (structures) inside the die, as shown in Fig. 5.2. These test patterns are called **test structures**. These test structures are used to monitor the IC fabrication process quality. Generally, these test structures are avoided to be put in the mask, as it consumes the silicon area. It has been mentioned before that each and every process is optimised separately prior to the sequential IC fabrication process. These individual optimised processes are called **unit processes**. Unfortunately, a sequential process, which is called **process integration**, slightly changes its integrated process recipe from the optimised unit process recipe. Hence, the unit processes have to be re-optimised by evaluating the test structures. Once these integrated recipes are optimised, IC fabrication is done in bulk.

Inside the die, the patterns may be of different sizes and shapes depending on the transistor layout and the electrical requirements of the IC. To fabricate an integrated circuit, a large number of masks are needed and the total number of masks required to make a complete integrated circuit is defined by the **mask count**. For instance, to make a MOS transistor, a minimum of four masks are required, as described in Chapter 3. This is called the **level of the mask**. Each level of the mask is named after its specific role in

MOS/IC fabrication. For instance, to make a polysilicon gate MOS transistor, the first level of mask is called the **active mask** which signifies that the MOS transistor will be made at this particular area on the silicon wafer. Similarly, the second level of the mask is called the **gate mask** and it is used to make the gate of the MOS transistors. The third level of mask is called the **contact mask** and it is used for metal contacts, and the fourth level of mask is called the **metal mask** and this mask is used to pattern the metal for electrical wirings and contacts (see Chapter 3).

To get well-aligned matrices of the dies, two cross (plus) marks are patterned on either side of the scribe track in all the mask levels. These patterns are called **reticle alignment marks** or **fiducial marks** or **die alignment marks** and are placed at the centre but outside of the scribe track in say, *x*-direction. One of the reticle alignment marks is shown in Fig. 5.4. The reticle alignment marks are aligned in a straight line with the help of a microscope and cross-wire arrangement and the matrix of the die is made on the mask. If all mask levels are not aligned then none of the masks can be aligned with respect to each other during the lithography process, as shown in Fig. 5.5. During die matrix fabrication, these reticle alignment marks are covered, so that they do not appear in the mask.

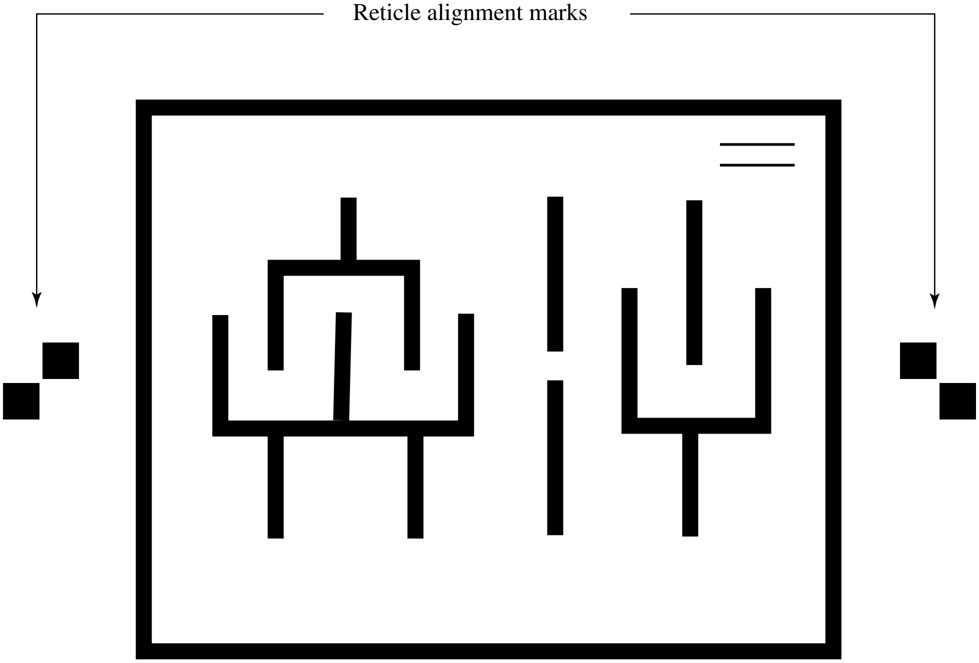


Fig. 5.4 Reticle alignment marks or "fiducial marks" or "die alignment marks"

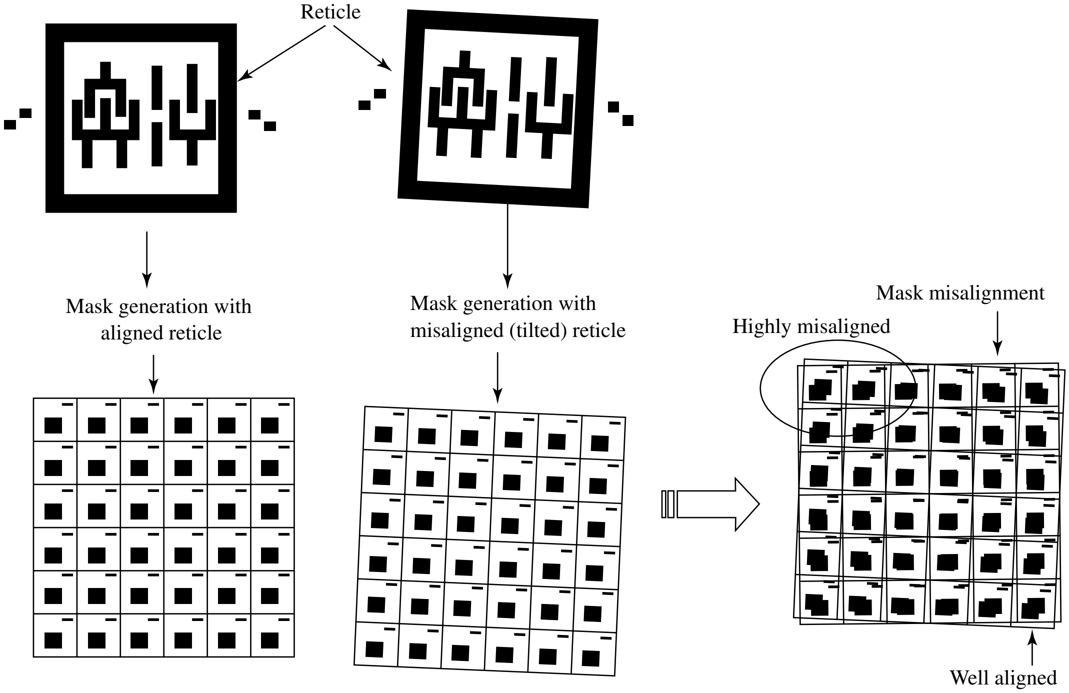

Reticle
Mask generation with aligned reticle
Mask generation with misaligned (tilted) reticle
Mask misalignment
Highly misaligned
Well aligned

Fig. 5.5 Importance of reticle alignment marks

5.2 PROPERTIES OF MASKS

A mask has a large number of criteria. The patterns of the die in the mask should be consistent and they should be distinctly separated (high resolution) from each other. The ratio of light passing through the transparent patterns and the opaque patterns should be extremely high (called **high contrast**). The edge definition (edge sharpness) of the die patterns should be very high. The corners of the die patterns should be perfect and they should not be rounded. This roundness of corners is aggravated during lithography and etching processes. To avoid the roundness, the corners of the die patterns are modified during mask fabrication. In addition, closely placed patterns are made relatively wider in the mask, so that the desired dimensions of the pattern on the wafer after the lithography and the etching processes can be obtained. This process of mask fabrication is called **mask engineering** or **mask biasing**. The mask should be hard, so that it can withstand the mechanical pressure during the lithography process and mask handling. Apart from these criteria, the glass plate should be extremely flat, and should have low defect density, high transmission for UV light, chemical resistance, and compatibility to the exposure system.

In addition, the glass should have low thermal expansion so that it does not deform due to the heat produced during lithography exposure. The mask should have long life. The mask should be extremely clean and should not have dirt, particulates and pinholes. In addition, the mask should be bigger than the wafer size, so that the wafer can be fully utilised. To meet all these criteria of the mask makes it a very costly affair.

5.3 TYPES OF MASKS AND MASK FABRICATION TECHNIQUES

Masks are classified into two categories depending on the total pattern area (exposed area) with respect to the unexposed area. One category of mask is called the **bright field mask** and the other category is called the **dark field mask**. The bright field mask has lesser opaque area than the transparent area; hence, this type of mask is called the bright mask as shown in Fig. 5.1. In contrast, the dark field mask has more opaque area than the transparent area; hence, this category of mask is called the dark field mask, as shown in Fig. 5.6. As the MOS is scaling down and the density of the MOS is increasing drastically, therefore, one can hardly categorise the masks.

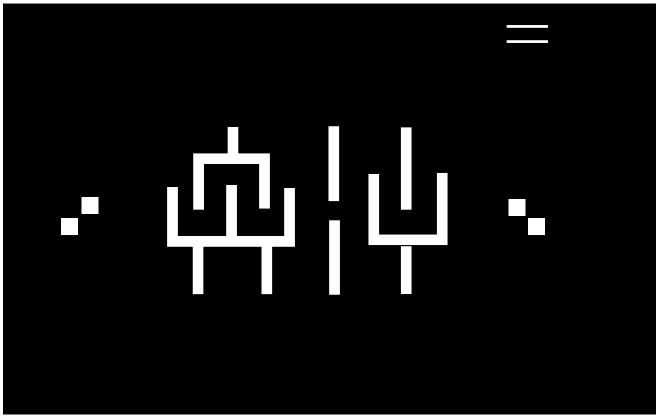


Fig. 5.6 Dark field artwork; 200 times bigger than die patterns

There are two types of masks used for IC fabrication, namely, **emulsion masks** and **metal masks**. In the emulsion mask, the photographic emulsion material is coated on the glass plate and patterns of the mask are registered on the photographic emulsion material using the photography technique. For this reason, this type of mask is called the

emulsion mask. Photographic emulsion is made of gelatin, silver halide and dye, as shown in Fig. 5.7. Gelatin acts as the base material for the emulsion. The silver halide impregnated into the gelatin converts into silver grains and becomes black (opaque) when the emulsion is exposed to light, and the remaining areas where light has not fallen, become transparent after developing. Hence, one can get opaque and transparent patterns on the photographic plate. The dye is used to increase the sensitivity of the emulsion to a particular wavelength of light. For IC fabrication, the dye is used in the emulsion to make it sensitive to green light. To reduce the halation effect, an annihilation layer is coated on the emulsion and the on-back of the glass plate. Generally, in IC fabrication, the resolution of the emulsion photographic plate is 2000 lines/mm and this emulsion photographic plate is called **high-resolution photographic plate.**

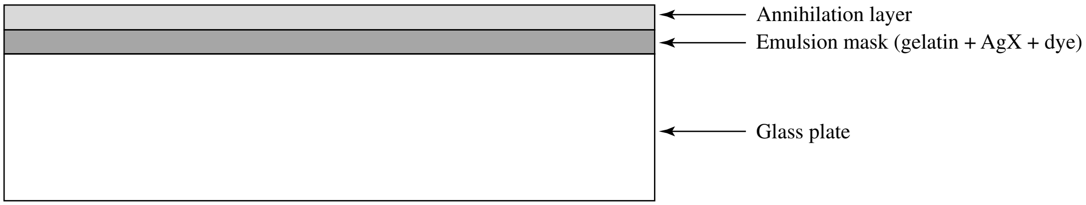


Fig. 5.7 Figure 5.7 High Resolution Plate (HRP) emulsion mask

In the metal mask, the patterns are made out of metal film deposited on the glass plate; therefore, this type of the mask is called the **metal mask** as shown in Fig. 5.8. Masks are made using the lithography and metal etching processes. Out of the two types of masks, the emulsion masks are easy to fabricate and are less costly, but they suffer with regard to pattern resolution. Generally, the maximum workable resolution that is advisable is 1.5 μm. For this reason, the emulsion masks are restricted to larger transistor dimensions, and not for VLSI and ULSI applications. On the other hand, the patterns of submicron geometries can be obtained in the metal masks. Therefore, in recent days, the metal masks are used for VLSI and ULSI applications.

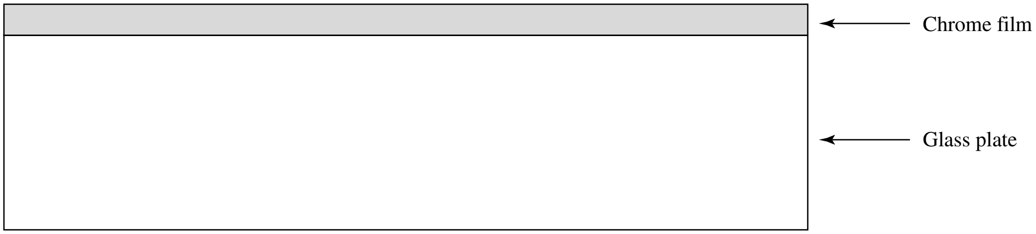


Fig. 5.8 Chrome-plate mask

5.3.1 Emulsion Mask Fabrication

The emulsion masks are made in two steps. In the first step, big-sized die patterns are made on a photographic plate called reticle. Thereafter, the image of the reticle is exposed onto another photographic plate in the matrix form. Usually the size of the die pattern in the reticle is 10 times bigger than the mask die size. The reticle consists of die patterns, wafer alignment marks, scribe tracks and reticle alignment marks. To make the reticle, these reticle patterns are made onto a plastic sheet called the **rubylith**. One side of the rubylith is coated with a thin red-coloured plastic film. The patterns of the reticle are made on the rubylith by cutting the red thin film. The machine which is used to cut the red thin film is called the **computer-controlled co-ordinator**. After cutting the red plastic film, the desired patterns are retained and from the rest of the places, the red plastic is peeled off. The patterns which are peeled off become transparent to light. This patterned rubylith sheet is called **artwork** as shown in Fig. 5.9. Depending on the required mask, the artwork is made either for a bright field mask or a dark field mask. The artworks of a bright field mask and a dark field mask are shown in Fig. 5.6 and Fig. 5.9 respectively. The mask which is made from the artwork of a bright field mask will have less red-film patterns. On the other hand, the mask which is made from the artwork of a dark field mask will have more patterns. When the artwork is illuminated from one side with green light, it has very high contrast when looked from the other side of the artwork. Generally, the artwork is made of 200 times (200×) the actual size of the die, and thereafter, it is reduced to 20 times (20×) on a photographic plate by a photographic lens system. Thereafter, the photographic plate is developed, fixed, washed and finally dried. The photographic plate that

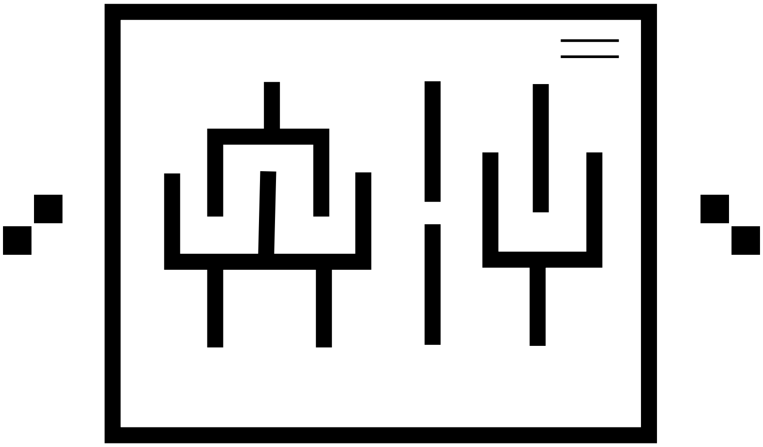


Fig. 5.9 Bright field artwork; 200 times bigger than die patterns on rubylith

contains the image of the artwork is called the **reticle**. The process by which the reticle is made is called the **first reduction process** and the camera which is used to make the reticle is called the **first reduction camera**, as shown in Fig. 5.10.

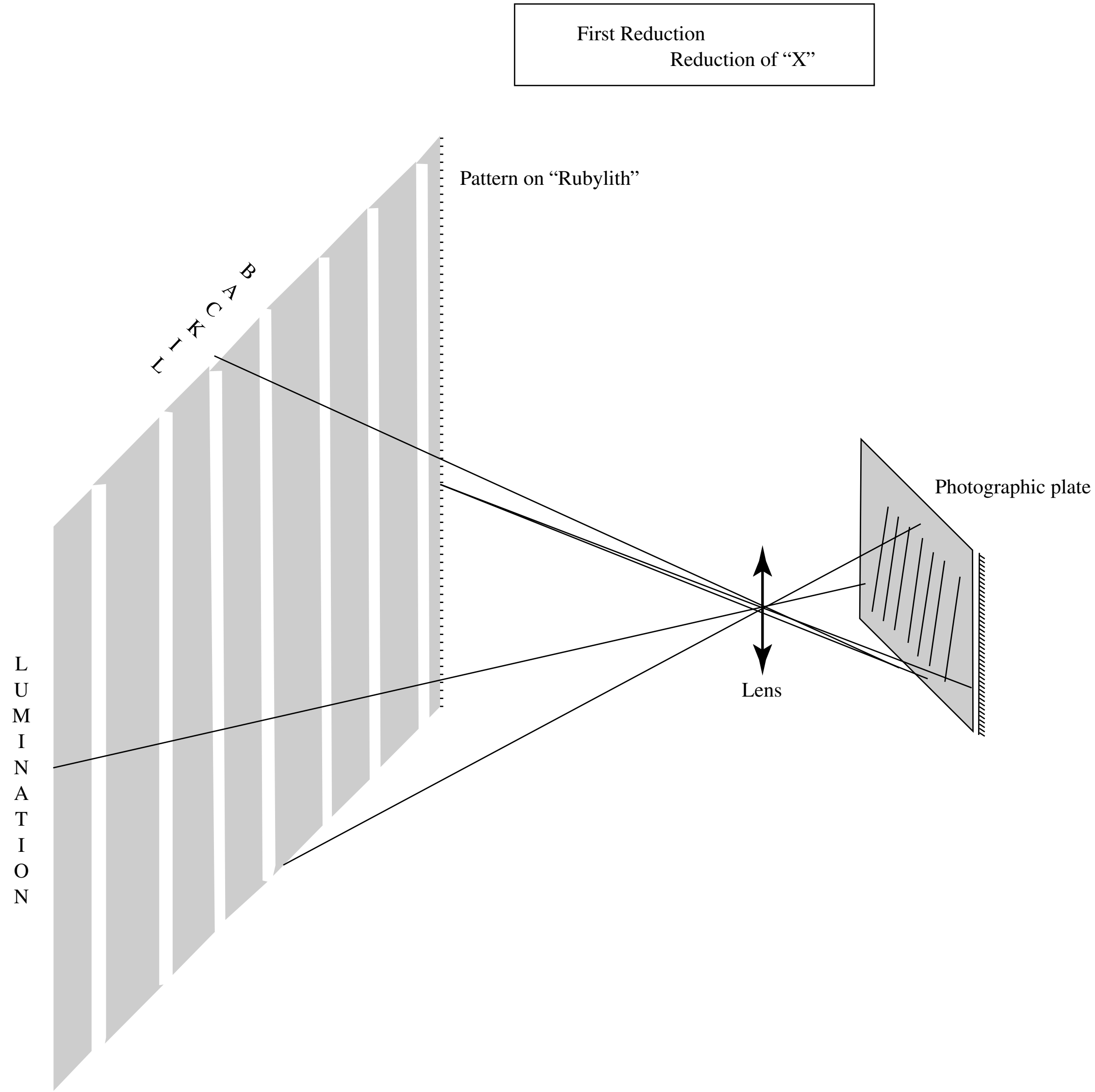


Fig. 5.10 First reduction camera

Once the reticle is made, it is placed in a special type of jig called the **reticle holder**. Thereafter, the reticle holder is placed under the microscope and reticle alignment marks are aligned in a straight line with the help of the microscope and a cross-wire arrangement. Once the reticle alignment marks are aligned in a straight line, the reticle is securely fixed with the jig. Then, the reticle jig is placed in between the photographic plate and a 10× reduction lens. This type of camera is called the **step and repeat camera**. The photographic plate is mounted onto the *x-y* translation stage. To get the matrices of the die patterns on the photographic plate, the translation stage first moves in the *x*-direction and the reticle is illuminated in steps, by a xenon flash lamp fitted at the bottom of the reticle. The stepping distance is chosen such that the scribe track of the die is partially overlapped with the subsequent die. Once the photographic plate is completely exposed in the *x*-direction, the translation stage moves back to its starting point and then the stage moves in the *y*-direction. If the scribe track is square in shape then the stepping distance in both the *x* and *y* directions is the same. Once the *x-y* stage moves to the desired distance in the *y*-direction, the xenon lamp flashes and the die is exposed; then the *x-y* stage moves again in the *x*-direction and exposes the second row of the matrix, as described above. This process repeats till the full photographic plate is exposed completely. The desired stepping distance of the *x-y* stage and the flashing of the xenon bulb are precisely controlled by a computer. This process of mask fabrication is called second reduction or **step and repeat process**, as shown in Fig. 5.11. Once the matrix of the reticle is exposed, the photographic plate is developed, fixed, washed and finally dried. This processed photographic plate is then called the **mask**.

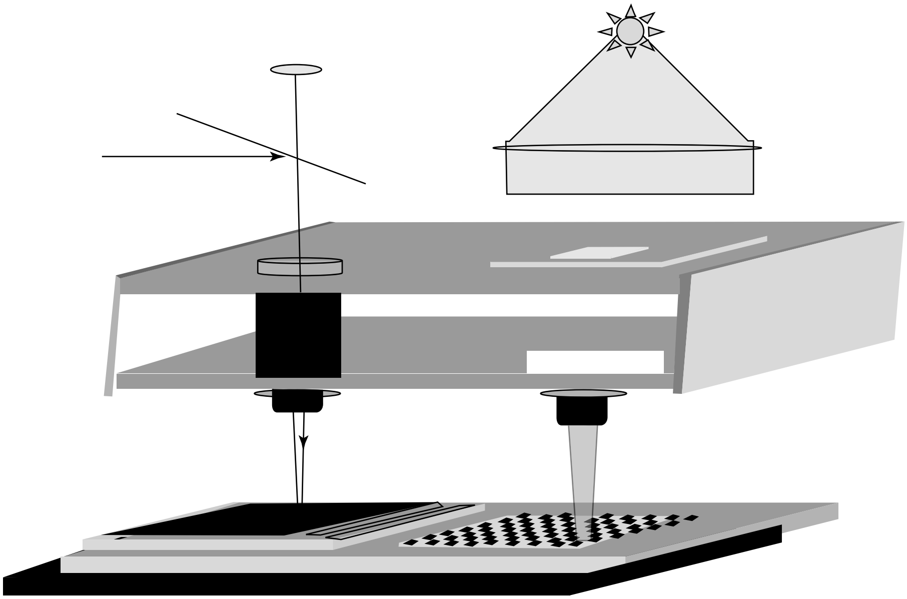


Fig. 5.11 Step and repeat camera

The emulsion reticle or mask can also be made on the photographic plate without going through the laborious artwork generation and first reduction process. An equipment is available to make the reticle or mask directly on the photographic plate. This equipment is called the **optical patterns generator**. The concept of the optical patterns generator is similar to that of the step and repeat camera, except that shutter is placed in place of the reticle. Two shutters are used; one that opens or closes in the *x*-direction and another that opens or closes in the *y*-direction. The apertures of these shutters are imaged on the photographic plate and exposed to the xenon lamp to generate the die patterns. Thereafter, the photographic plate is developed, fixed, and dried. The shutter can be opened or closed in the *x* and *y* directions independently, so that different pattern sizes or shapes can be obtained on the photographic plate.

Unfortunately, the emulsion masks are not suitable for VLSI and ULSI applications. This is because the gelatin of the photographic emulsion is soft, and swells after water absorption. In addition, the large grain size of the emulsion grain limits the resolution of the emulsion mask significantly.

5.3.2 Metal Mask

A thin metal film is deposited on the glass plate and thereafter, a photoresist is coated over the metal film. The patterns on the PR are made using the lithography and metal-etching techniques. The details of the lithography process and the metal-etching techniques are described in Chapters 6 and 7 respectively. Generally, the metal mask is made of chrome film. This is because chrome metal is hard, less oxidising and its adhesion with glass is excellent. For these reasons, the mask is made of a chrome metal film and is called the **chrome mask** as shown in Fig. 5.8, and it has been used in IC fabrication in recent years.

To make the chrome mask or reticle, a 1000 Å thick chrome film is deposited on one side of the glass plate and then, a photoresist is coated over the chrome film. Thereafter, the mask or reticle patterns are made onto the photoresist. After the photoresist is developed, the unexposed photoresist remains on the glass plate and the exposed photoresist dissolves in the developer, from the rest of the glass plate. Thereafter, the chrome-film-plated glass is put into the chrome-etching solution. The uncovered chrome film is etched away and the chrome patterns that are protected by the photoresist, remain unaffected. Once the chrome etching is complete, the photoresist is removed completely from the entire glass plate; whereas the chrome film patterns remain on the glass plate. The chrome patterns on the glass plate block the light and except the areas of chrome patterns, light passes through the plate.

The chrome mask or the reticle patterns can be directly made by the pattern-generation technique. Here, the photoresist-coated chrome film glass plate is kept at the focal plane of the lens and illuminated by the ultraviolet light. If the dimensions of the MOS transistors are bigger, the mask patterns can be made directly by this technique and if the dimensions of the transistors are small (in the submicron range) then the mask is made using the reticle.

The metal masks are also made using the **electron beam technique**. In this technique, the die patterns are generated (exposed) by an electron beam; therefore, this process of mask making is called the **electron beam** or the **e-beam mask-making technique.** In this technique, the mask patterns are made by a well-focussed electron beam on the resist-coated chrome plate. The resist material is different than the optical photoresist. The construction of the e-beam equipment is similar to the oscilloscope, except that in the place of the phosphorous screen, a resist-coated chrome plate is used as shown in Fig. 5.12. There are two modes of exposing the pattern, namely, the **raster scanning mode** and the **vector scanning**

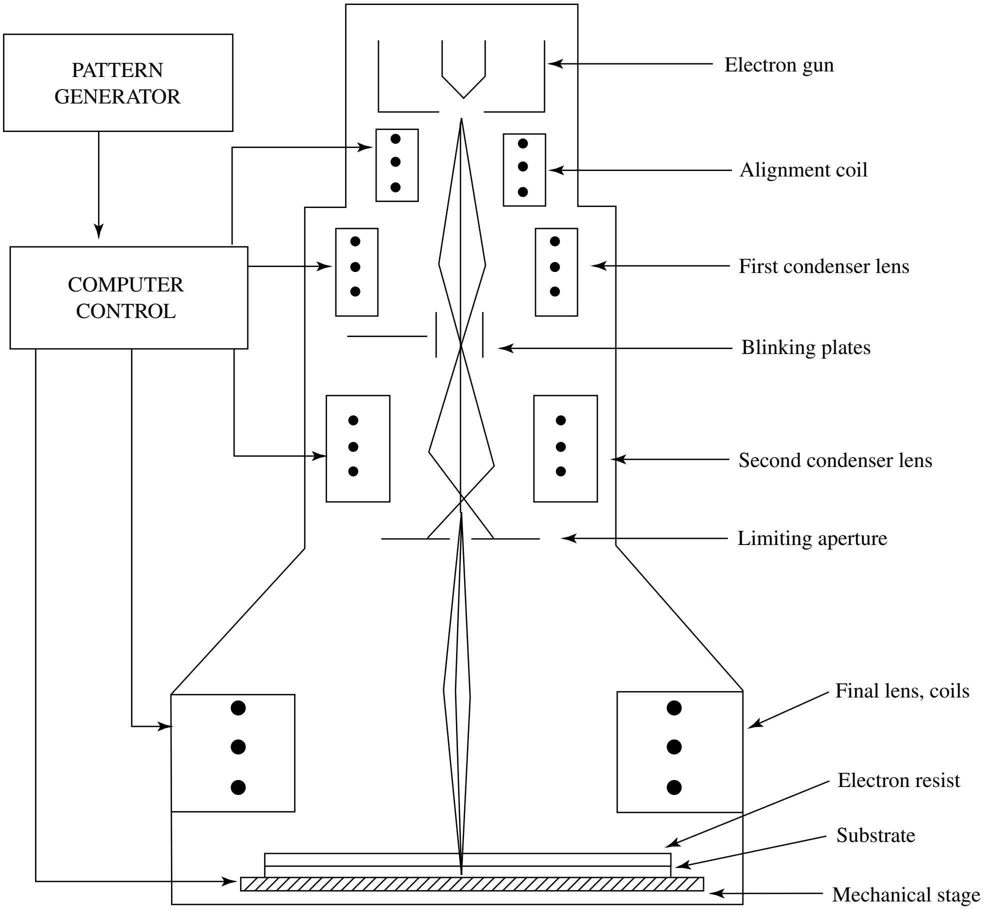


Fig. 5.12 Schematic of an electron beam machine

mode. In the raster scanning mode, an electron beam is kept fixed and the resist-coated metal glass plate mounted on the *x-y* stage moves in straight (raster) lines in the *x* or *y* directions. The electron beam exposes the resist at the desired pattern and at other times, the electron beam is blocked (gated off) completely. This mode of scanning is similar to that of the line printer. In the vector scanning mode, the resist-coated chrome plate is kept fixed and the electron beam moves and makes a pattern; thereafter, the *x-y* stage moves to the next pattern and then the e-beam is exposed again. Unfortunately, mask fabrication by the electron-beam technique takes a long time and that reduces the mask throughput (mask or reticle per unit time) significantly; but the masks of patterns with submicron geometries (high resolution) can be generated using this technique. To increase the mask throughput to some extent, the electron-beam diameter is changed according to pattern geometry. For instance, for big pattern geometries, the e-mean diameter is increased during exposure.

In recent days, the electron-beam technique is used extensively for mask fabrication for VLSI and ULSI applications. It is also very suitable for custom-made IC applications, where large volumes of ICs are not required. Unfortunately, mask making by the electron-beam technique is costly, it requires frequent maintenance of the equipment and the mask throughput is very less; but this is the only option for mask making of the VLSI and ULSI applications.

It is essential to mention that the mask becomes defective during each cycle of its use, especially in the intimate contact mode of lithography, as explained in lithography in Chapter 6. Therefore, after a few lithography processes, the masks are changed. The number of times a mask is used for lithography before it is discarded is called the **mask lifetime**. Hence, to reduce the cost of the IC, a combination of the emulsion mask and the chrome mask is used. Usually, the reticles of all mask levels are fabricated and then from these reticles, the chrome masks are made by the step-and-repeat camera technique and the reticles are kept reserved. Another option is to keep the original set of masks and from these set of masks, another set of masks are made by duplicating. The original set of masks is called the **master masks**. To increase the mask life, the mask is not put into intimate contact mode, but kept a small distance away from the wafer. This mode of lithography is called the **proximity mode**. Apart from the proximity mode technique, another technique is used, where the mask is kept a long distance away from the wafer and an image of the mask is projected onto the wafer. This technique is called **projection lithography**. Details of the intimate contact mode, the proximity mode and the projection mode of lithography are described in Chapter 6. It is said that the mask is the heart of IC fabrication, so it should be handled carefully and delicately.

Summary

The mask is made of localised opaque (black) and transparent patterns. These localised patterns of the mask are translated by lithography and etching processes on the wafer.

There are two types of masks used for IC fabrication, namely emulsion masks and metal masks. The emulsion masks are made of a high-resolution photographic emulsion using the photography technique and its workable resolution is around 1.5 μm; whereas the submicron geometries can be obtained from the chrome metal mask. Generally, the chrome mask is made using electron-beam lithography. In recent days, the chrome masks are used for VLSI and ULSI applications. But in the near future, direct writing on the wafer by e-beam lithography will be used.

Individual masks contain a matrix of identical black and transparent patterns called the die pattern, and the area occupied by the patterns is called the die. Each die pattern has three important features, namely IC patterns, alignment marks and scribe tracks. The alignment marks are used to align a particular level of the mask with respect to the other levels of the mask during the lithography processes. The scribe track is used for the first mask alignment with wafer cut and separating one die from the other die (i.e. one IC from the other IC) by cutting (scribing) the wafer.

To fabricate an integrated circuits, a large number of masks are needed and the total number of masks required to make a complete integrated circuit is defined by the mask count, and each mask is defined by the level of mask.

A mask pattern should have high resolution, sharp edge and perfect corners. To get high fidelity of patterns on the wafer, the mask patterns are modified and this process is called mask engineering. The mask glass plate should be extremely flat, and should have low defect density, high transmission for UV light, chemical resistance, compatibility to the exposure system, low thermal expansion, should be extremely clean, should be able to withstand the mechanical pressure during the lithography process and mask handling, and should not have dirt, particulates and pinholes; also, the mask should be bigger than the wafer size, so that the wafer can be fully utilised.

References

- David J Elliott; *Integrated Circuit Fabrication Technology*, McGraw-Hill, 1982
- J D Plummer, M Deal and P B Griffin; *Silicon Fundamental Technology: Fundamentals, Practice and Modeling*, Prentice Hall, 2000

- S M Sze; *VLSI Technology*, Second Edition, McGraw-Hill, 1988
- S M Sze; *Semiconductor Devices Physics and Technology*, John Wiley and Sons, 1985

Multiple-Choice Questions

5.1 The mask contrast should be
(a) high (b) low (c) does not matter

5.2 Emulsion mask is made using
(a) lithography technique (b) photographic techniques

5.3 Ideally, the thermal expansion of the mask and the silicon wafer should be
(a) same (b) high (c) low

5.4 What is the thinness of the emulsion of an emulsion mask?
(a) 1–2 microns (b) less than 1 micron
(c) more than 2 microns

5.5 What is the thickness of the metal of a metal mask?
(a) 1000 Augustan (b) 1 micron (c) more than 1.5 microns

5.6 Which side of the mask should be kept during lithography?
(a) Emulsion/metal should be in intimate contact with the silicon surface
(b) Mask glass should be in intimate contact with the silicon surface
(c) Does not matter

5.7 Reticle alignment mark is used for
(a) die alignment during lithography (b) wafer cut alignment
(c) array of die alignment

5.8 Mask alignment mark is used for
(a) die alignment during lithography (b) wafer cut alignment
(c) array of die alignment

5.9 Mask size with respect to wafer size should be
(a) bigger (b) smaller (c) does not matter

5.10 The photographic plate is exposed to which light?
(a) Green (b) Red (c) Ultraviolet

5.11 Metal mask plate is exposed to
(a) green light (b) red light (c) ultraviolet light

Descriptive Problems

5.1 A particular level of mask is aligned to only one level of mask or more than one level of mask. Explain with the help of the CMOS process-flow mask sequence.

5.2 The emulsion mask has less pattern resolution; explain the reasons.

5.3 Using light diffraction, explain why the projection exposure system is better for IC fabrication than the other exposure systems.

5.4 Write the difference between emulsion mask and chrome mask.

5.5 In how many ways can the emulsion mask be made?

5.6 What is the importance of reticle alignment marks?

CHAPTER

6

Lithography

6.1 INTRODUCTION

MOS transistors are made from patterns of mask-levels and all mask-level patterns should be well aligned. It has been found that the maximum number of failures in IC fabrication occurs in the lithography process step. For this reason, the lithography process should be carried out with great care and precaution. The lithography process starts with the photoresist (PR) film coating on the wafer. Thereafter, the mask patterns are translated on the PR. Details of mask patterning as well as mask fabrication have been described in Chapter 5. Thereafter, the wafer is put in the film etchant solution that etches the film. The role of the PR is to protect the film below from the film etchant. The unprotected film reacts with the etchant and gets removed (etched out) from the wafer. Once the film etching is complete, the PR is removed completely from the wafer. This process of removing the entire PR from the wafer is called **PR stripping**. When the PR is stripped from the wafer, the unprotected film patterns remain on the wafer. As discussed above, to make patterns on the film, one has to go through two distinct processes, namely, the **lithography process** and the **etching process**. In this chapter, the lithography process is described and in the subsequent chapter, the etching process is covered.

6.2 PHOTOLITHOGRAPHY PROCESS

The literal meaning of lithography (*litho* means stone) is to engrave or print on the stone. But in the context of IC fabrication, the meaning of lithography is to make PR patterns on the wafer. The lithography process steps are shown in Fig. 6.1 and are described as follows. In the first step, the silicon wafer is coated uniformly with the photoresist. Usually, PR coating is done by two techniques. In the first technique, the wafer is held in a jig, and then the PR is sprayed over the wafer using a PR spraying

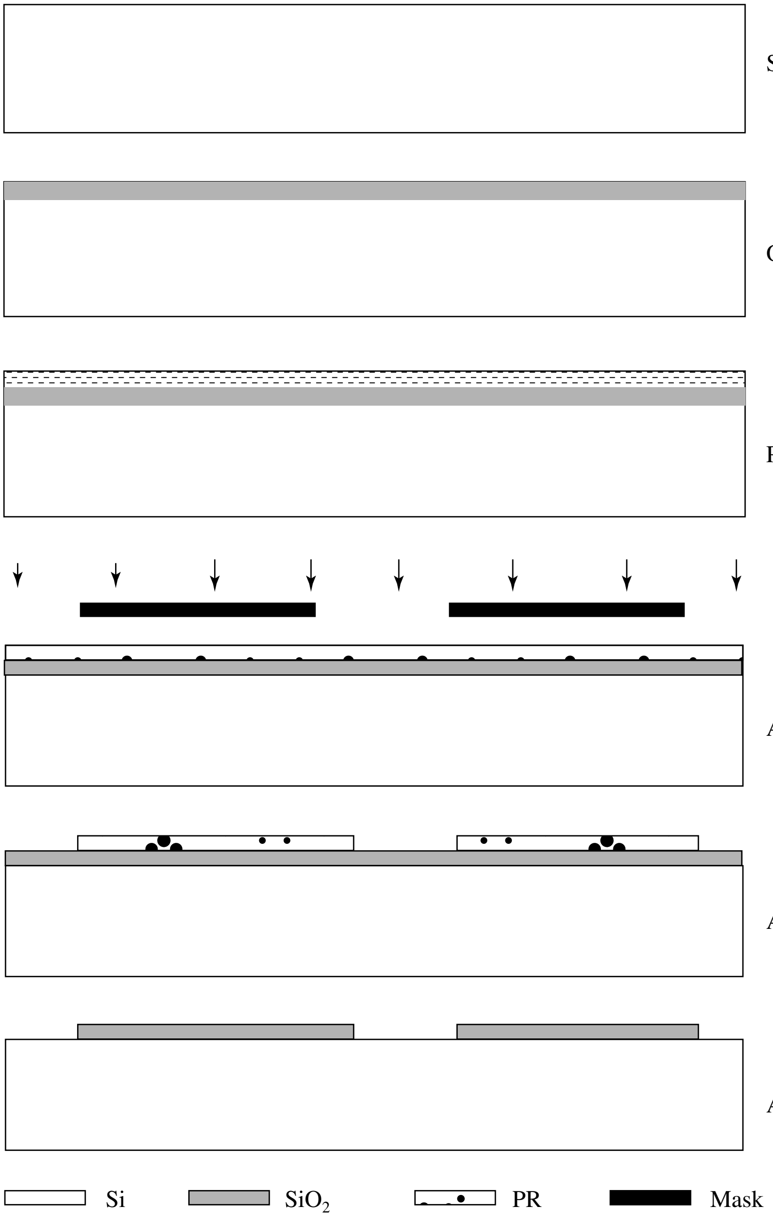


Fig. 6.1 Lithography process steps

machine. This process of PR coating is called **spray coating**. In the second technique, the wafer is firmly held in a jig by vacuum and the PR is put onto the wafer surface using a PR dispenser and then the wafer is spun at a high speed. At the high rotation speed, a thin layer of the PR remains on the wafer surface and the excess PR flows off the wafer due to centrifugal force. This process of PR coating is called **spin coating** and is shown in Fig. 6.2. The thickness of the PR depends on the wafer jig rotation speed as well as the PR viscosity. The spin-coating technique is more popular than the spray-coating technique. Normally, the thickness of the coated PR is around 1 micrometre. Prior to PR coating, HMDH is coated to promote adhesion between the PR and the wafer. After PR coating, the wafer is kept inside the oven for 45 minutes at around 95°C temperature. This process of heat treatment is called the **PR pre-baking process** or simply **pre-baking**. In the PR pre-bake process, the solvent present in the PR evaporates significantly and makes the PR harder, so that the PR does not stick to the mask during lithography. In addition, the PR adheres well to the wafer. Then, the patterns of the photographic mask are translated onto the PR. This process is carried out using a special type of equipment called the **mask aligner**. In the mask aligner, the wafer and the mask are brought closer to each other. To translate the first mask patterns onto the wafer (PR), one of the scribe tracks of the mask is perfectly aligned with the primary cut (110 orientation) of the wafer, as shown in Fig. 6.3. The alignment process is done by looking at the wafer cut and the mask scribe track simultaneously through a microscope. To avoid damage, both the mask and the wafer are kept separated around 1 to 2 micrometres at the time of the alignment process. Once the scribe track and the wafer cut are aligned, the wafer and the mask are made to come in contact and are then exposed by UV light from the top of the mask. Generally, the exposure time of the PR is in the order of a few seconds to a few minutes depending on the type of the PR. After PR exposure, the wafer is developed in the PR developer solution. Usually, the PR is developed in 60 seconds at room temperature. The exposed PR (positive PR) goes away in the developer and the unexposed PR remains on the wafer, as shown in Fig.

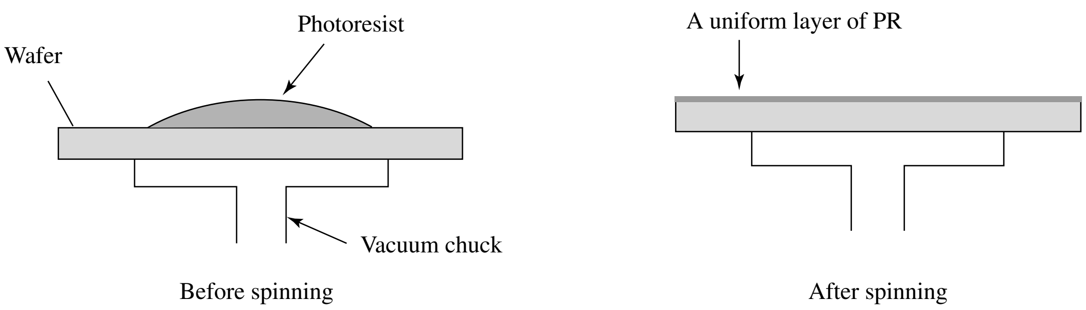


Fig. 6.2 PR coating on wafer

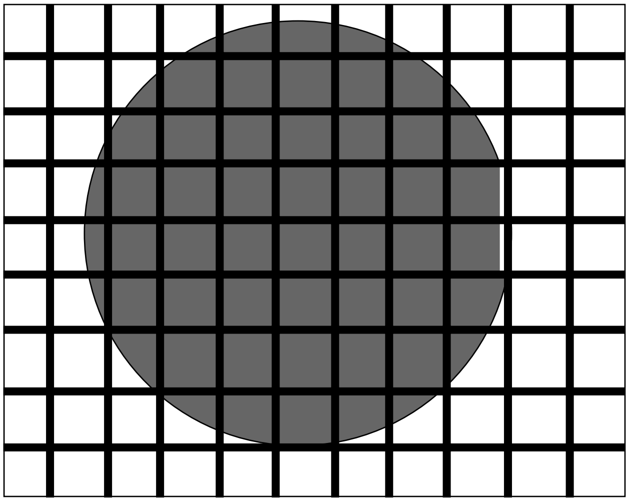


Fig. 6.3 Top view of first mask and wafer-alignment procedure

6.1. Once the PR developing is complete, the wafer is put into DI water. The DI water is used to stop further development of the PR and remove the developer from the wafer. Then the wafer is taken out from the deionised water (DI) and dried in pure nitrogen gas. Thereafter, the PR patterns are inspected thoroughly under the microscope. If any problem is encountered till this point, then the wafer can be processed again, after completely removing (stripping) the PR from the wafer. Then, the wafer is put for a second heat treatment in the oven for 45 minutes at around 110°C. This process of heat treatment is called **PR post-bake process** or simply **post-bake**. The post-bake process further hardens the PR. In addition, it promotes adhesion to the wafer and increases resistance to the etchant. Once the PR patterning is complete, the wafer is immersed in the film etchant solution. The film areas that are protected by the PR do not dissolve in the etchant solution. On the other hand, the unprotected patterns are etched out in the etchant solution. Once the etching of the film is complete, the role of the PR is accomplished; hence, the PR is stripped off from the wafer by a chemical or by the electric discharge (plasma) in the presence of a selective gas. The chemical that dissolves the PR is called the **PR stripper** or in short, the **stripper**. Once the PR is completely removed from the wafer, the replicas of film patterns of the first photographic mask patterns are left on the wafer. It is important to mention that the PR should be processed strictly as per the vendor's recommendations; otherwise the lithography process may lead to many problems.

For the second mask lithography, the patterns of the second mask are aligned with the patterns created by the first mask lithography on the wafer. Similarly, the lithography processes of all the levels of the masks are done in the same process, as discussed in Chapter 3.

6.3 PHOTORESIST

The photoresist (PR) plays a very important role in the lithography process. The photoresist is made of an organic polymer material, a photosensitive material (dye), and a solvent. The base material of the PR is an organic material. The photosensitive dye is added to the base material to increase the sensitivity of the PR to UV light. The base polymer and the dye are dissolved in an organic solvent to create the PR in liquid form. There are two types of photoresists used in lithography, namely, the **positive photoresist** and the **negative photoresist**. When the positive PR is exposed to UV light, the bigger polymer chains in the PR get ruptured and form smaller chains which decrease its molecular weight and the PR gets dissolved in the PR developer; whereas the unexposed bigger chains in the PR do not get dissolved in the developer, and remain on the wafer. In contrast, when the negative PR is exposed to UV light, the organic material of the negative PR forms bigger polymer chains from the smaller polymer chains and that increases its molecular weight. In the PR developer, the lower molecular weight polymer dissolves and the higher molecular chain remains unaffected.

The PR should have certain qualities such as high resolution, high resistance to the etchant, and high sensitivity to the exposed light. The resolution of the PR is a very important factor in lithography. PR resolution means that all the close PR patterns must be completely separated from each other. PR resolution is also discussed in Chapter 11 in detail. Apart from resolution, the PR should also have the ability to withstand the etching process; otherwise, the PR may corrode or get dissolved in the etchant solution. The PR should adhere well to the wafer; otherwise it will detach from the wafer. There are many reasons of poor adhesion of the PR to the wafer. One of the main reasons of poor adhesion is moisture. If the film or the wafer absorbs moisture for any reason, then the wafer should be heated up at around 400°C for 30 minutes in the N_2 ambient, to drive out (diffuse out) the absorbed moisture from the wafer. In addition, moisture in the lithography process room also plays a significant role in poor PR adhesion. Hence, the lithography room should not have more than 50% Relative Humidity (RH). For better PR adhesion, the hexamethyldisilazane (HMDS) solution is applied onto the wafer prior to the PR coating; therefore, HMDS is called the **PR adhesion promoter.** Generally, a freshly formed film does not contain moisture, so the PR can be directly applied onto the wafer, but the use of an adhesion promoter is always advisable.

Light sensitivity of the PR is one of the most important parameters. The light sensitivity of PR is defined as the **the minimum quantity of light energy that is required to make or break the polymer chains of the base material**. The photosensitivity of the dye enhances the absorption of light. This absorbed light energy converts into thermal energy of the base material and that in turn modifies the organic chain. Nowadays, chemical amplifier is used in place of dye to increase the PR sensitivity. It is essential to mention that PR resolution is also a function of light wavelength. The resolution of PR also increases with lower wavelength of light. Unfortunately, shorter light sources (deep UV) emit less intense light thus a chemical amplifier is essential. The unit of the sensitivity of the PR is mJ cm^{-2}. The typical PR sensitivity for **g-line and I-line spectra** ultraviolet light emitted from high pressure mercury source is around 100 mJ cm^{-2}; whereas, deep UV does not have good enough light intensity; and its sensitivity lies between 20–40 mJ cm^{-2}, so the PR needs a chemical amplifier.

6.3.1 Types of Photoresists

Lithography can be carried out by many different types of exposure sources, namely, the optical exposure source, the X-ray exposure source, the electron beam exposure source, and the ion-beam exposure source. The polymer which is used for the light source is called the **photoresist.** There are two types of photoresists used in optical lithography and these are called the **positive photoresist** and the **negative photoresist**. Both these types of PRs are in liquid form. The positive and negative PRs are made of different base organic materials and they are complementary to each other. For other types of exposure systems, such as the electron beam exposure source, and the X-ray exposure source, different types of polymers are used and these are simply called the **resist**.

6.3.2 Positive Photoresist

The main ingredient of positive PR is a particular type of organic material: resin. Resin has a large polymer chain and high molecular weight. The photosensitive material (dye) and the organic solvent are added in the resin, as per requirement. The positive PR is red in colour. The role of the photosensitive dye has been explained in Section 6.3. The solvent is used to dissolve the resin and the dye to change the viscosity of the PR. The absorbed thermal energy breaks the polymer chains of the resin that decreases the resin's molecular weight. The low molecular weight of the resin makes it dissolve in the developer; whereas the unexposed resin with high molecular weight chain remains unaffected. The developer of the positive PR is made of a mild alkaline solution. Once the PR is developed, the wafer is rinsed in pure water. As the positive PR is red in colour, the patterns of the PR can be seen easily on the wafer. The positive PR does not swell in the developer or in water; hence, it retains its pattern dimensions even after the process. The positive PR dissolves in the acetone very easily; therefore, acetone is used as a positive PR stripper. The positive PR can be resolved in the nanometre scale; thus, it is used for high density VLSI and ULSI applications. Unfortunately, the positive PR is much costlier than the negative PR.

6.3.3 Negative Photoresist

The main constituent of the negative PR is polyisoprene polymer. The polyisoprene polymer is a rubberised type of organic material and it is made of smaller polymer chains. The photosensitive material (dye) is added into the polymer polyisoprene, and then, both the materials are dissolved in the organic solvent. The role of the dye and the solvent are the same as described earlier in Section 6.3.2. When the negative PR is exposed, the big polymer chains are formed from the smaller polymer chains. The formation of big polymer chains from the small polymer chains is called **chain cross-linking**. The cross-linked chains have higher molecular weights and they do not dissolve in the PR developer, as shown in Fig. 6.4. In contrast, the unexposed PR, which has lower molecular weight, dissolves in the developer easily. Usually, xylene is used as a developer. The developer, rinse and stripper of the negative PR are made of organic solvents. The negative PR is not friendly to use as compared to the positive PR. In addition, the negative PR swells in water and loses its resolution significantly. Due to these limitations, negative PR is not used for high density VLSI and ULSI applications. The negative PR is colourless and it is hard to see the patterns on the wafer.

6.4 NON-PHOTORESIST (RESIST)

As mentioned earlier, the base materials of the photoresists are not suitable for the electron beam and the X-ray lithography. In place of the photoresist, other types of organic polymers such as polybutene-1 sulphone (PBS) and polymethyl methacrylate (PMMA) are used as the positive resist, and polyglycidyl methyl acrylate co-methyl acrylate (COP) and germanium selenide (GeSe) are used as the negative resist for electron-beam lithography. For the X-ray lithography, polybutene-1 sulphone (PBS) and polymethyl acrylate (PMMA) are used as the positive resist, and dichloropropyl acrylate + glyodyl methacrylate (DCOPA) resist is used as the negative resist.

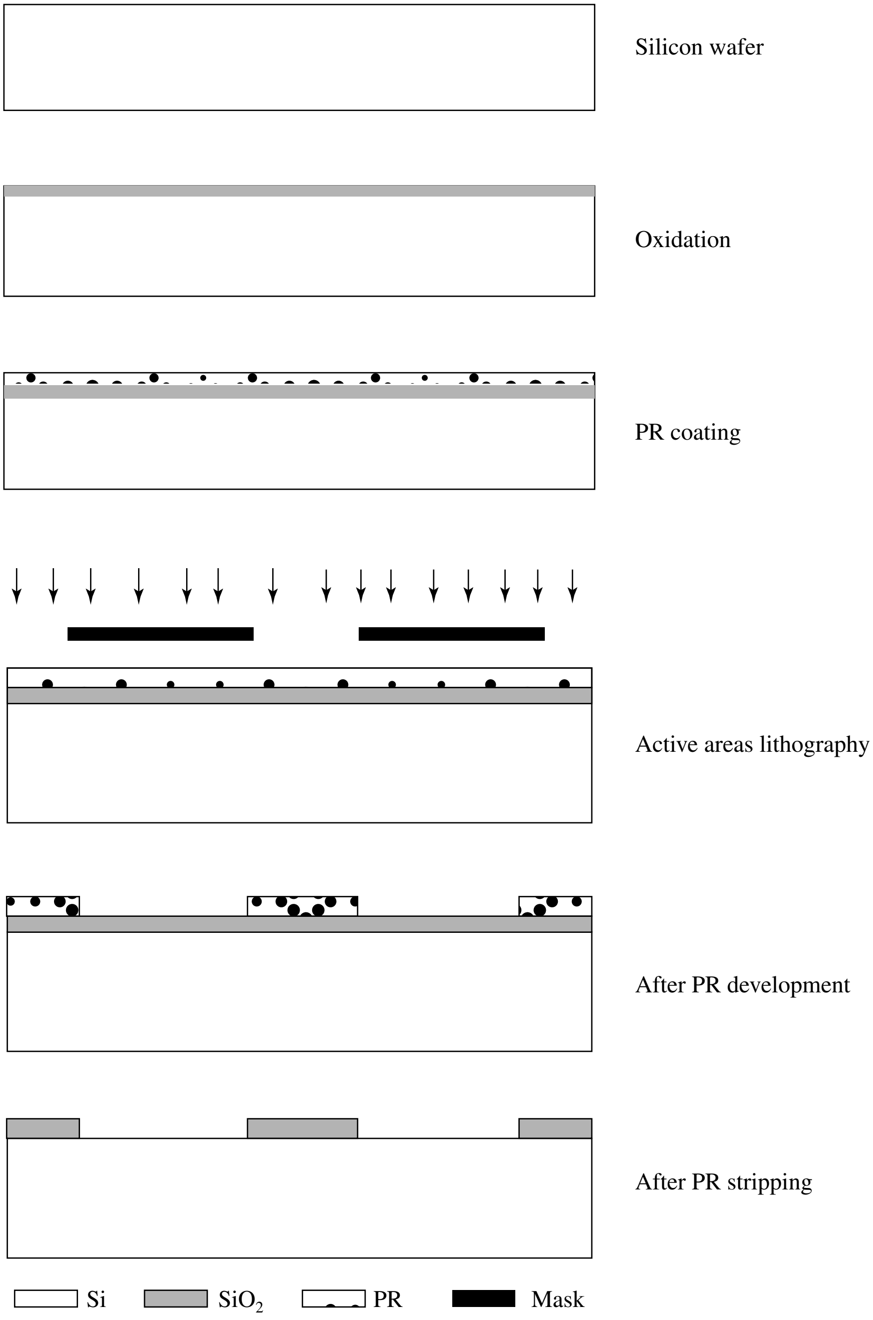


Fig. 6.4 Negative PR process

6.5 PR CLEANING PROCEDURE

It is essential that after the lithography process, the PR is completely removed from the wafer. Any trace of the PR on the wafer will be fatal for the IC yield and the process equipments, especially at high temperatures. At high temperature, the PR burns and turns into carbon. This carbonised PR is hard to remove by any type of chemical etching. Therefore, the PR has to be removed fully from the wafer prior to wafer processing. Generally, PR cleaning is done in two steps. In the first step, the PR is removed either by dissolving in an organic solution or by oxidising the PR in the plasma condition. This process of removing the PR is called **PR stripping**. The chemical which is used to remove the PR is called the **PR stripper** and the process by which the PR is removed in the plasma condition is called **Reactive Ion Etching (RIE)**. Reactive ion etching system and its operation is described in Chapter 10. In spite of a careful PR stripping, there are chances that a trace of PR is left on the wafer. This trace of PR is removed in the second step of PR cleaning: the wafer is kept in a concentrated sulphuric (H_2SO_4) acid container and slowly, hydrogen peroxide (H_2O_2) is added into that container. When H_2O_2 is added in the H_2SO_4, the solution becomes hot and reactive, and that dissolves the PR completely. Generally, the wafer is kept for around 15 minutes in the H_2SO_4 and H_2O_2 solution. It is important to mention that the silicon also gets oxidised in the solution. This oxide is removed by HF that further cleans the PR and any impurities or particulates from the wafer.

6.6 LIGHT SOURCE AND THE OPTICAL EXPOSURE SYSTEM

The light source and the optical exposure system play a very vital role in the PR patterning process. To understand the importance of the light source and the optical exposure system in the lithography process, some optical definitions and optical phenomenon are essential to explain. One of the most important definitions is the **optical resolution** (not PR resolution). As the density of the MOS transistors is phenomenally increasing in a chip and the size of the MOS transistors decreasing drastically, similarly the PR patterns are also decreasing. Therefore, the optical exposure system resolution, the wavelength of light and the PR resolution play a significant role in lithography.

The resolution or resolving power R of the optical system is expressed as

$$R = \frac{0.61\lambda}{NA} = \frac{k\lambda}{NA} \tag{1}$$

where k is an optical constant and its value is around 0.61, λ is the wavelength of the light and the nomenclature NA stands for the numerical aperture of the optical system. The numerical aperture of the optical system is defined as

$$NA = n \sin \theta \qquad (2)$$

where n is the refractive index of the medium and θ is the angle of incidence of the light on the wafer, as shown in Fig. 6.5. From Eq. (1), it can be seen that for higher optical resolution (to increase the resolution, R has to reduce), shorter wavelength and higher numerical aperture is required. For this reason, a continuous effort is going on to find out the short wavelength light sources and a higher numerical aperture optical system. Some of the light sources and wavelengths being used in optical lithography are listed in Table 6.1.

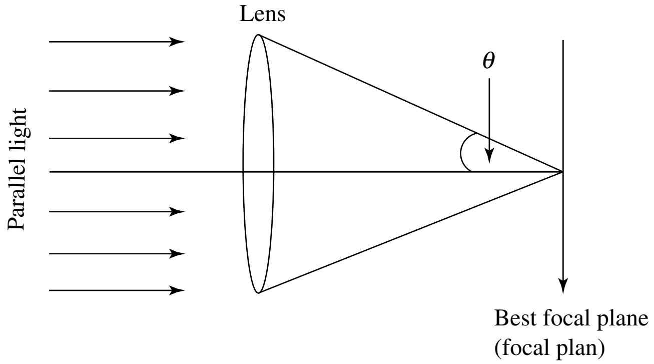


Fig. 6.5 Numerical aperture of the optical system

Table 6.1 Optical light sources and their wavelengths

Optical light source	Wavelength of light
High pressure mercury light g-line	436 nm (g-line)
High pressure mercury light i-line	365 nm (i-line)
Mercury–Zeon	270–290 nm
KrF Excimer laser	248 nm
ArF Excimer laser	193 nm

Furthermore, the focal length of the exposure source is related to NA. For visualisation, the focal length, numerical aperture and the resolution of the standard optical microscope is mentioned in Table 6.2.

Table 6.2 Particulars of standard optical microscope

Standard microscope focal length	Numerical aperture	Resolution (lines/inch)
50 mm	~0.1	~9,000
8 mm	~0.6	~40,000
2 mm	~1.3	~95,000

When the light illuminated circular aperture is imaged by the optical lens system at its focal plane, a bright central disc surrounded by dark and bright concentric (circulars) rings is formed, as shown in Fig. 6.6. The central bright disc is called the **Airy's disc**. According to the Rayleigh theory of resolution, the two nearby circular apertures can be resolved, if the first minima (dark ring) of the second aperture falls on the Airy's disc of the first aperture. This theory is called the **Rayleigh resolution criteria,** as illustrated in Fig. 6.7.

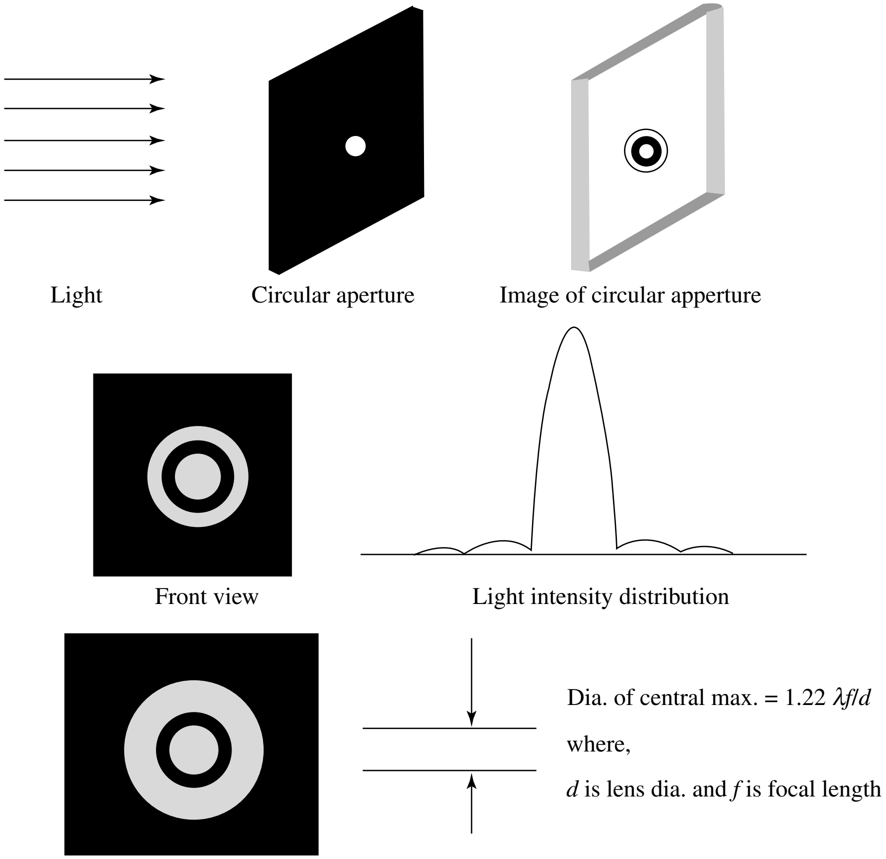

Light
Circular aperture
Image of circular apperture
Front view
Light intensity distribution
Dia. of central max. = 1.22 $\lambda f/d$
where,
d is lens dia. and f is focal length

Fig. 6.6 Optically illuminated circular aperture is imaged

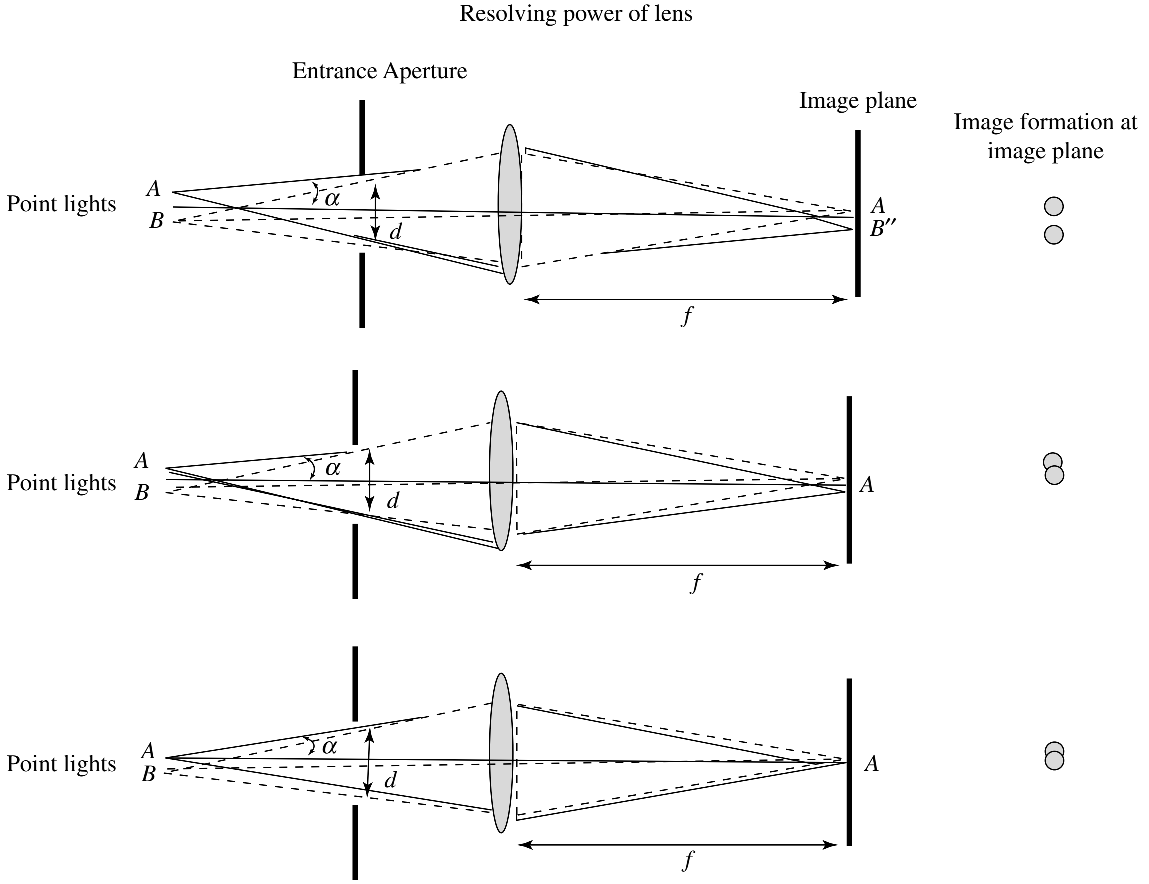
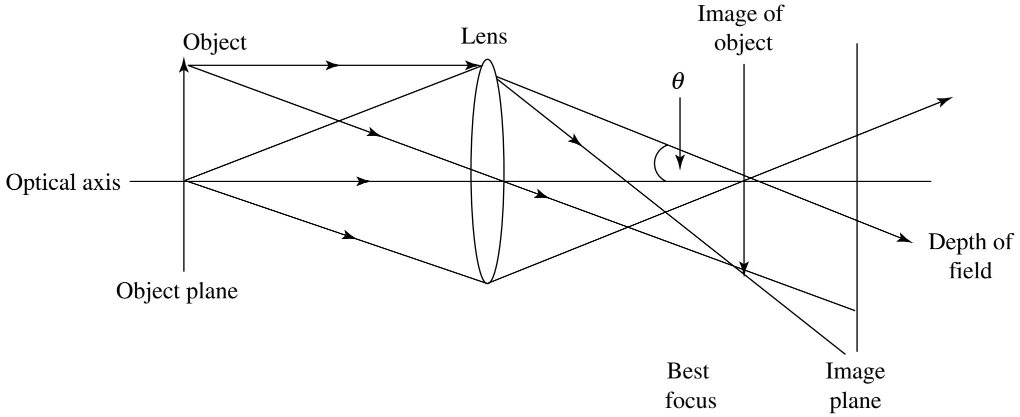


Fig. 6.7 Optically illuminated circular aperture is imaged and resolution

To get the best resolution, the images of the object patterns must be sharply focussed on the PR (that is image plane) of the object. For example, as explained in Chapter 5, the sharp images of the patterned rubylith are formed by the first reduction camera lens on the photographic plate. As per the optical definition, the plane that contains the rubylith is called the **object plane**, and the plane that contains the photographic plate is called the **image plane** as shown in Fig. 6.8. The magnification and demagnification (reduction) of the object depends on the relative positions of the object plane, the optical lens and the image plane.

The lateral shift on either sides from its best focus plane leads to degradation of the image quality and in turn the resolution. The maximum lateral shift of the image plane from its best focus plane, without losing much of the image quality (resolution), is called



where

λ = wavelength of light

n_a = numerical aperature

n = refractive index

θ = angle between converging rays and principal axis

Fig. 6.8 Optical image of object

the **depth of focus** as shown in Fig. 6.8. Similarly, the maximum tolerable lateral shift on either side of the object plane, without losing much of the image quality, is called the **depth of field** as shown in Fig. 6.8. Furthermore, the resolution is also limited by the object size. The maximum allowed object area (field) formed by the optical system, which does not cause much loss in the image quality, is called the **field of view**. In the optical projection system (first reduction system), the field of view implies the size of the artwork that can be projected on the focal plane without much loss in resolution in any part of the artwork. The field, in context of the optical lithographic system, is the maximum reticle or mask size that can be printed on the wafer.

It is well known that light does not travel in a straight line, when it strikes the edge of an object. This phenomenon is known as the **diffraction** of light, as shown in Fig. 6.6. The diffracted light degrades the resolution severely. Diffraction becomes more prominent when the object contains closely placed fine patterns. In addition, the diffraction increases as the wavelength of light decreases. On the other hand, the resolution increases as the wavelength of light decreases. The diffraction effect is minimised by designing the optical lens system. The optical system which has insignificant diffraction effect is called the **diffraction limited optical system**. The diffraction limited optical systems have been used in the recent VLSI and ULSI applications.

6.7 PATTERN TRANSFERRING TECHNIQUES AND MASK ALIGNER

In IC fabrication, the exposure techniques can be divided into two categories: the **optical technique** and the **non-optical technique**. In the optical technique, the PR is exposed to the optical light in the UV range. The optical lithography technique can be further divided into categories, namely, shadow and projection printing. In the non-optical technique, the resist is exposed by the X-ray, the electron beam and the ion-beam.

6.7.1 Optical Lithography Technique

Optical Shadow Printing

The basic principle of optical shadow printing is to cast the opaque and transparent patterns of the mask onto the wafer, more specifically, onto the PR. As the shadows of the mask patterns are casted onto the PR coated wafer, this technique is called **shadow printing technique**. Prior to PR exposure, the wafer and the mask patterns are well aligned. Once the alignment is done, the PR is exposed, developed and rinsed as described in Sections 6.3.2 and 6.3.3. The mask and wafer alignment and the PR exposure are done by a special type of machine, called the **mask aligner**. In the mask aligner, the wafer is held by a vacuum jig, which is attached to the *x-y-z* translation stages and Φ rotation mechanisms. The jig that holds the wafer is called the **wafer holder**. The mask is held just above the wafer by a vacuum chuck (jig) at a fixed position. This mask chuck is called the **mask holder**. The patterns of the mask and the wafer are aligned by adjusting the *x, y* and Φ stages of the wafer holder with respect to the fixed mask position. At the time of alignment, the mask and the wafer are kept a few micrometres apart, so that they do not rub each other and get damaged. Once the alignment is complete, the wafer is brought in intimate contact or very close to the mask by raising the wafer holder by the *z* movement. The alignment procedure of the first mask is shown in Fig. 6.9; where the scribe track of the mask is aligned with the cut of the wafer. Thereafter, the wafer is exposed to UV light from the top of the mask. After the exposure, the wafer is taken out from the mask aligner and the rest of the lithography steps are carried out, as described previously. It may so happen that the result of lithography is not satisfactory; in this case, the PR is stripped from the wafer completely and the wafer is processed again for lithography.

There are three modes of shadow printing lithography. These modes are: **intimate contact shadow printing mode** (**intimate contact mode**), proximity shadow printing mode (**proximity mode**) and **lift-off mode**. It is essential to mention that the shadow lithography was the first to be used in IC fabrication; hence, it is also called **conventional lithography**.

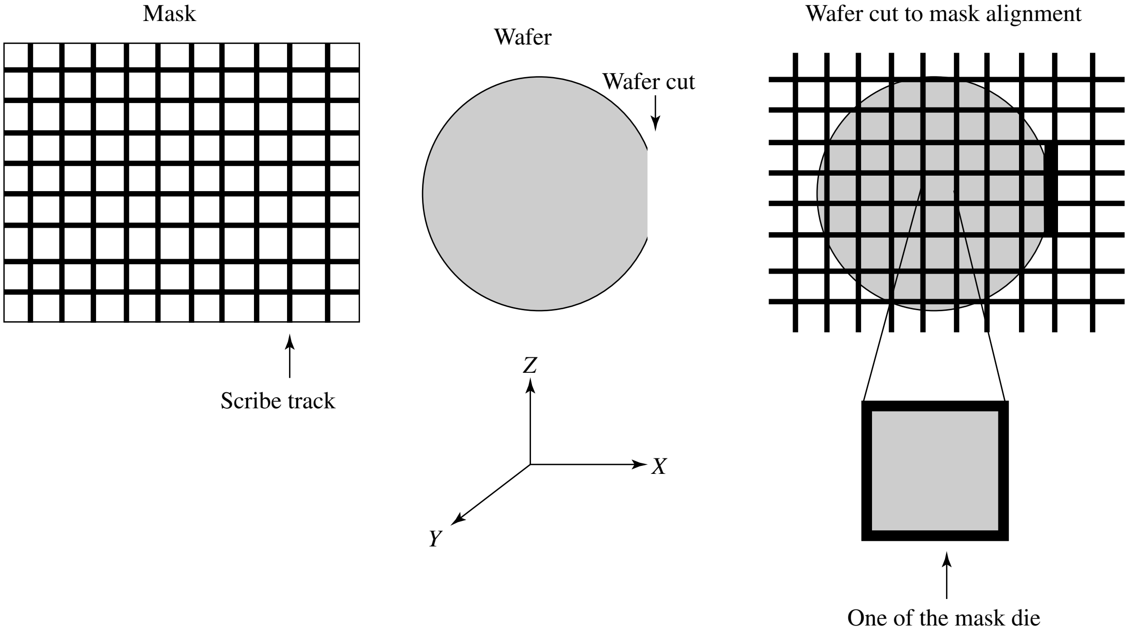


Fig. 6.9 Top view of first mask and wafer alignment procedure

Contact Mode Lithography

In the contact mode lithography, the wafer and the mask are made to be in intimate contact with each other at the time of exposure, as shown in Fig. 6.10. In general, the intimate contact is done using air pressure. The contact mode lithography has many advantages out of which some are worth mentioning such as high resolution, low cost, simple technology, convenience of use, high output, no limitation of the field of view (as the entire mask is printed), insignificant diffraction effects, and sharp edge definition (sharp vertical). Depending on the pressure exerted between the wafer and the mask during contact printing, the contact printing is subdivided into two categories, namely, the **hard contact printing** and the **soft contact printing** (see Chapter 11). In the hard contact printing, the wafer is pressed very hard against the mask, and in the soft contact printing, the wafer is gently pressed against the mask. Unfortunately, in the contact mode of lithography, the mask is damaged due to the mask and wafer contact. This reduces the mask lifetime significantly. As the cost of the mask is very high, so damage to the mask is a very serious issue. To overcome this disadvantage, the mask is duplicated onto another photographic plate (or chrome plate) by the contact printing technique. The duplicated mask is called the **working mask** and the original mask is called the master mask. Thereafter, the ICs are fabricated by the duplicated mask. Apart from the problem of mask damage, the contact

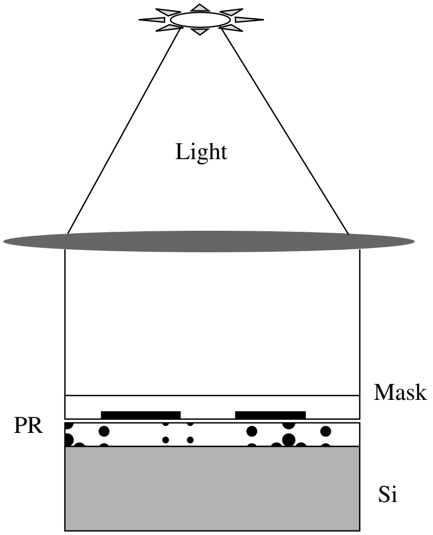


Fig 6.10 Contact printing mode

lithography requires a flat wafer surface for better contact and uniform pressure over the entire wafer to get better resolution. Furthermore, uneven pressure on the wafer may lead to wafer breaking.

To make a high edge definition, very high resolution, and to avoid wafer breaking, another version of contact mode lithography is employed. This mechanism is called **conformal lithography**. In conformal lithography, the mask is made of flexible glass or plastic sheet. The intimate contact is made by creating vacuum between the flexible mask and the wafer. The flexible mask bends (conforms) around the deformations, such as warps and bows on the wafer. This results in a very high resolution and high degree of edge definition with no fear of wafer breaking. The conformal lithography technique is not used in IC fabrication, due to the instability in the mask pattern dimensions caused by plastic deformation of the flexible mask. On the other hand, the conformal lithography is most suited for large area devices; where a very high edge pattern definition is required, such as the Surface **Acoustic Wave Devices (SAW)**.

Proximity Mode Lithography

In the proximity mode lithography, the mask and the wafer are separated by a few micrometres during light exposure. For this reason, this mode of lithography is called **proximity mode lithography**. The schematic diagram of the proximity mode lithography is shown in Fig. 6.11. It is obvious that the diffraction of light from the mask patterns decreases

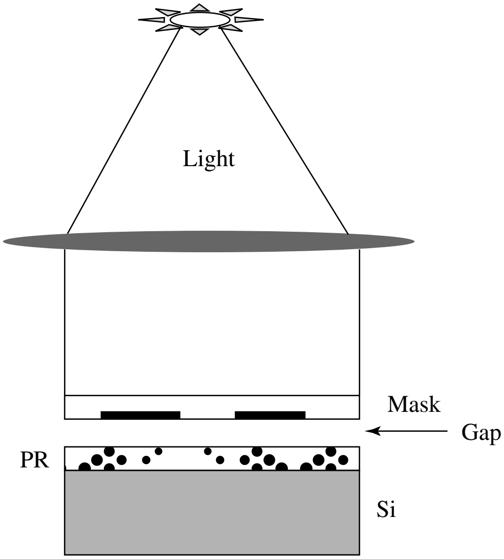


Fig. 6.11 Proximity printing mode

resolution. The diffraction of light increases as the separation (gap) between the mask and the wafer increases. The relationship between the proximity gap, wavelength of the light, and pattern resolution is expressed as:

$$R = \sqrt{(\lambda/g)}$$

where R is the resolution, λ is the wavelength of the light and g is the separation (gap) between the mask and the wafer.

In the proximity mode lithography, the mask is not damaged. In addition, lithography can be carried on the warped and bowed wafer without any fear.

Example 6.1

Find the resolution and depth of focus of the projection optical system, if k is 0.61 and the numerical aperture is 0.5 for the wavelengths of light as 4360 Å for high pressure mercury light g-line, and 1930 Å for ArF excimer laser.

Answer

$$R = k\frac{\lambda}{NA} = 0.61\frac{4360}{0.5} = 0.61 \ x8720 = 5319 \text{ lines/mm}$$

$$R = k\frac{\lambda}{NA} = 0.61\frac{1930}{0.5} = 0.61 \ x3860 = 2354 \text{ lines/mm}$$

$$\text{Depth of focus} = \frac{1}{2}\frac{\lambda}{(\text{NA})^2} = 0.5\frac{4360}{(0.5)^2} = 0.5\frac{3650}{0.25} = 8720\ \text{Å}$$

$$\text{Depth of focus} = \frac{1}{2}\frac{\lambda}{(\text{NA})^2} = \frac{1}{2}\frac{1930}{(0.5)^2} = 0.5\frac{1930}{0.25} = 3860\ \text{Å}$$

X-ray Lithography and X-ray Mask Aligner

X-ray lithography is done in the proximity mode of lithography where an X-ray source is used to expose the resist. The X-ray is produced by the bombardment of electrons onto an appropriate material called the **target**. Commonly, the palladium target is used for X-ray lithography. The palladium target emits the X-ray wavelength of around 4.4 Å that is almost 10^3 times shorter than the UV light wavelength. Therefore, the pattern resolution of the X-ray is expected to be 10^3 times higher than that of UV light. Unfortunately, the pattern resolution degrades due to the scattering and secondary generation of the X-ray inside the resist.

In the X-ray lithography process, the mask and the wafer are aligned with the visible light in a separate alignment system. Once they are aligned, the whole alignment system is transported under the X-ray beam. The mask alignment system and the X-ray exposure system are shown in Fig. 6.12. The X-ray takes more exposure time than optical lithography; hence, the throughput of the X-ray is lesser than the optical techniques. The X-ray lithography has not gained popularity, because X-ray mask fabrication is extremely difficult. The X-ray mask is made of a composite layer of many absorbing and non-absorbing (transparent) films on the silicon wafer membrane, where the silicon wafer provides mechanical support to the composite layers. Details of X-ray mask fabrication is explained in Chapter 11. The X-ray does not pass through the thick silicon wafer; therefore the silicon is thinned down to a transparent membrane by a silicon etching technique called **silicon micromachining**. The X-ray mask fabrication process is very elaborate and extremely complicated. The fabrication of the X-ray mask is shown in Fig. 6.13. The X-ray mask is fragile in nature and very costly. In addition, there is always a chance of mask distortion. In spite of these disadvantages, the X-ray lithography is a simple and low-cost technique. Most importantly, the X-ray is transparent to dust and many other foreign particulates that lead to fewer defects on the patterned resist. More about X-ray lithography is described in Chapter 11.

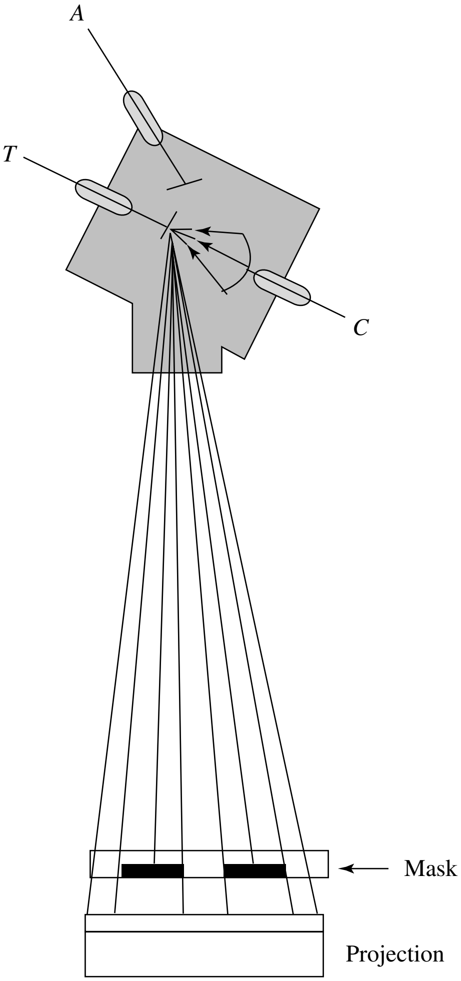


Fig. 6.12 X-ray exposure system and mask alignment system

Lift-off Lithography Technique

The lift-off lithography technique is a very powerful technique; especially the single mask process for the Micro-Electro-Mechanical-System (MEMS) and the Surface Acoustic wave (SAW). Extremely fine line patterns, vertical sidewalls, and sharp edges of the PR patterns with no undercutting (film does get etched under the sides of PR patterns) are the key advantages of lift-off lithography. Unfortunately, the lift-off technique is not compatible with the IC fabrication process line. This is because of its process-related issues. In the lift-off technique, the film is deposited after the PR patterning on the wafer, which is the reverse of the process done in the rest of the lithography techniques. The lift-off technique does not need the film etching process, but in the other lithography techniques, etching is a must. The process sequence of the lift-off is shown in Fig. 6.13. In the lift-off process,

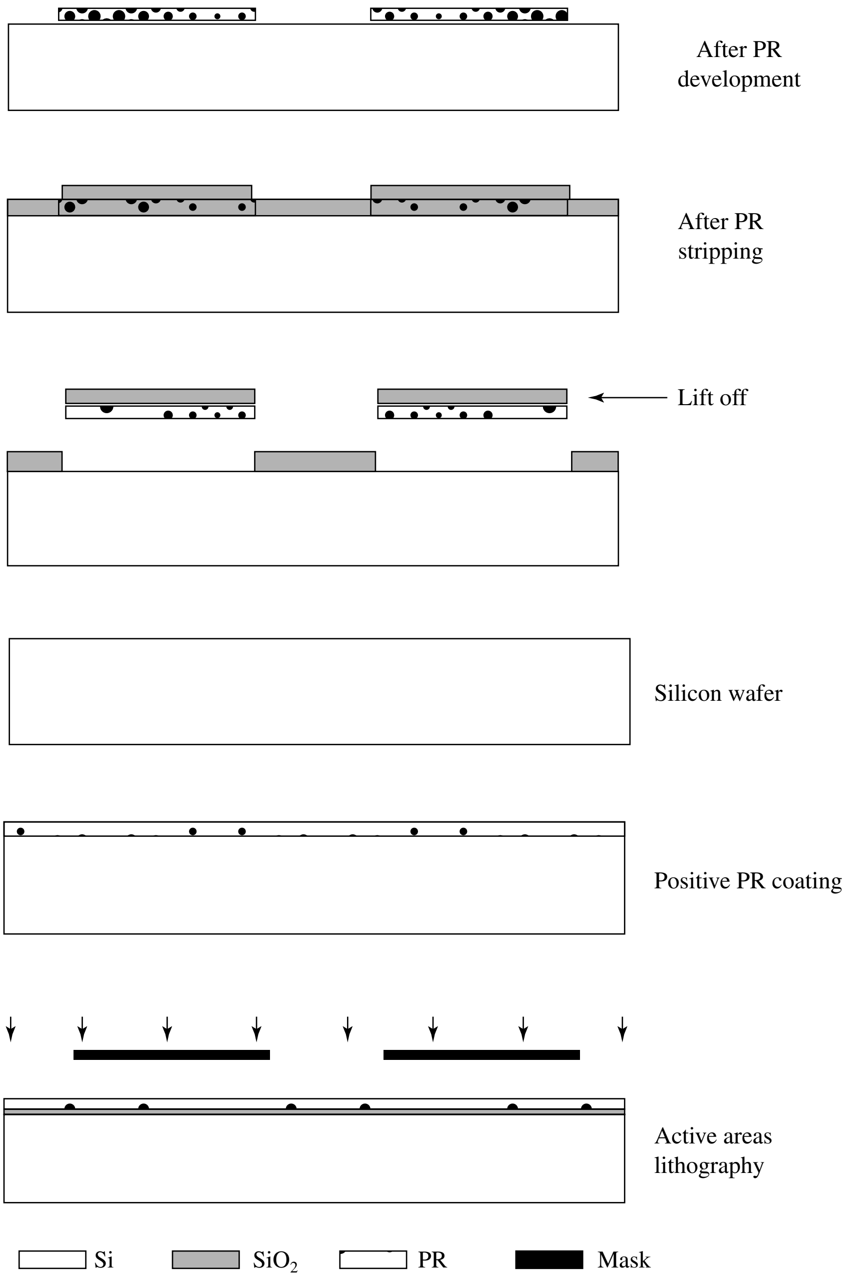


Fig. 6.13 Lift-off lithography process steps (Contd.)

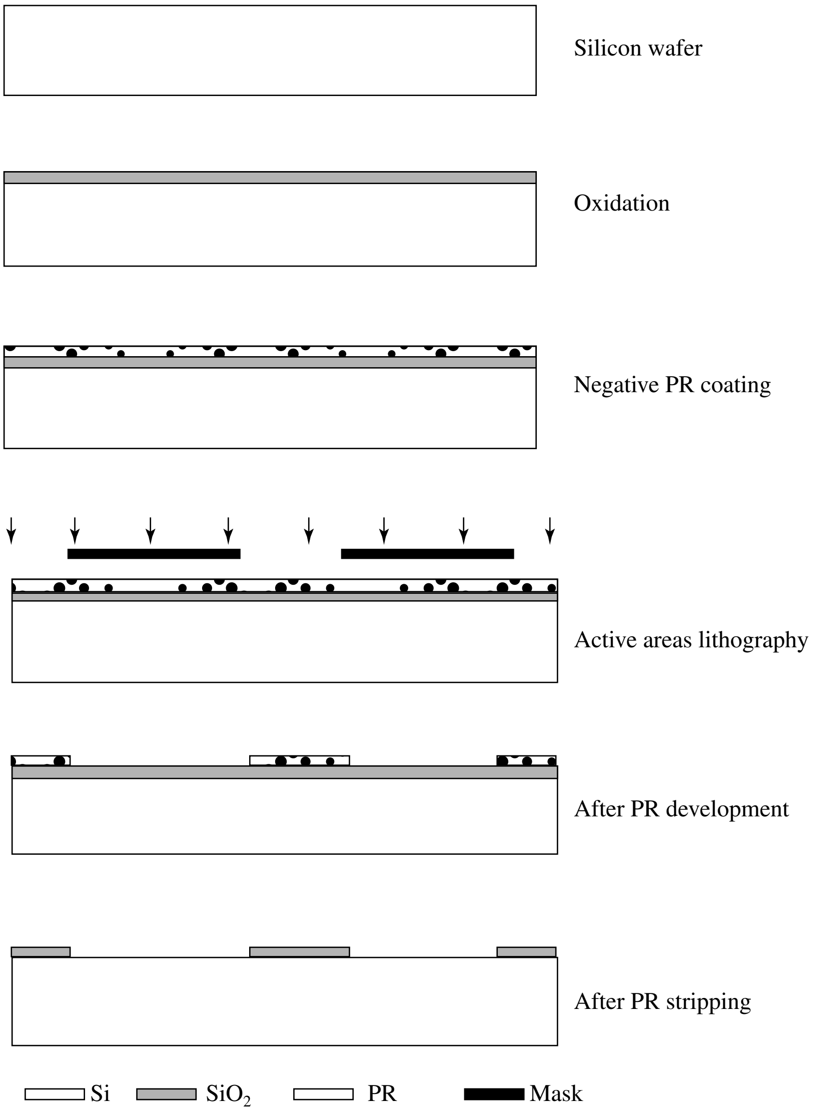


Fig. 6.13 Lift-off lithography process steps

the wafer is coated with around 2-micrometre thick PR. After patterning the PR, a very thin film is deposited on the entire wafer. This film is deposited on the patterned PR as well as on the bare silicon. Then the wafer is dipped into the chemical chlorobenzene for 30 minutes. In chlorobenzene, the PR swells and increases in volume that results in the breaking of the film at the PR/wafer edges. When the wafer is immersed in acetone, the PR is dissolved and the broken film patterns over the PR float (lift-out) on the acetone surface, and the film patterns deposited on the bare silicon remain as it is on the wafer.

6.8 OPTICAL PROJECTION LITHOGRAPHY TECHNIQUE

In the projection lithography, the image of the mask is projected on the PR coated wafer by the optical projection system. In this technique, an image of the reticle (or mask) is projected on the wafer and a matrix of reticles is made on the wafer by the step and repeat system. The optical projection systems are designed to have high resolution, high field of view, high depth of focus and other important optical parameters, besides being diffraction free. Therefore, the projection lithography systems are very costly as compared to the conventional contact lithography systems. Many types of projection systems are available in the market and they can be divided according to their exposure systems and image formation techniques. These optical projection systems and image-formation techniques are shown in Fig. 6.14; where the mask may be of *MX* (times). The notation *M* stands for the demagnification factor and *X* is the die size. For 1*X*, *M* = 1 is directly exposed onto the wafer, without any reduction in mask size during projection. In this optical system 1:1 (no reduction and magnification) projection system is designed, as shown in Fig. 6.14. In

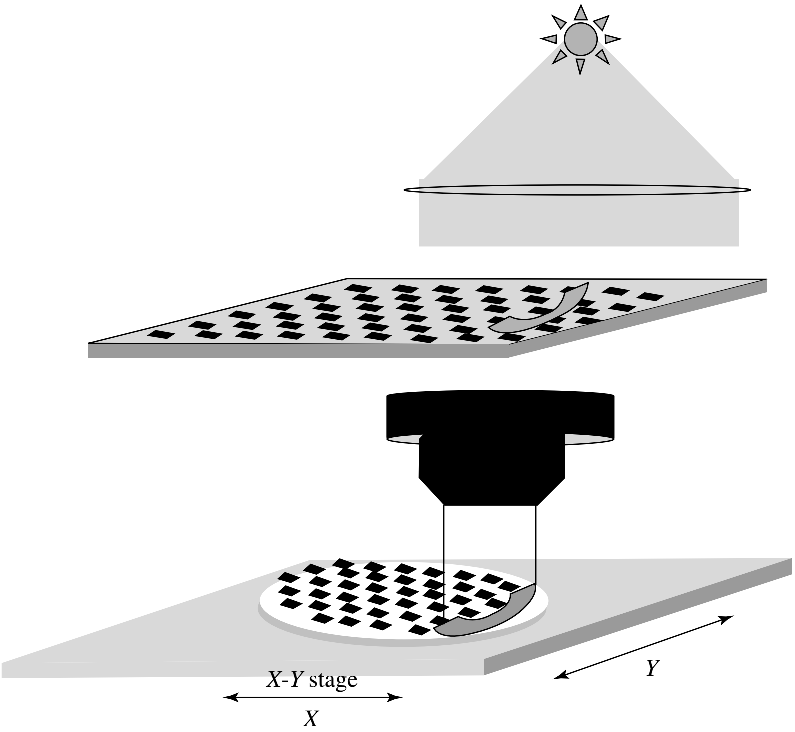


Fig. 6.14 1:1 projection system

the other category of the optical projection systems, reticle is used (not mask) as shown in Fig. 6.15. This projection system is also called the reduction projection system where the reticle is *M* times bigger than the actual die pattern of the IC. For example, in case of a reticle which is 10 times (*M* = 10) bigger than the die size, the projection system reduces the image of the reticle by 10 times on the wafer. This projection system is called the **1/10 projection system** and it is the same as the second reduction camera (step and repeat camera). The only difference is that the PR coated wafer is placed in place of the photographic emulsion and UV light is used to expose the PR rather than the optical light. 1/4*X* and 1/5*X* projection systems are also available in the market. In the optical projection lithography technique, the reticle (mask) is never damaged as the mask (or reticle) and the wafer are far from each other; but the mask (or reticle) gets defective due to light exposure and mask handling. One of the advantages of the reduction projection system is that any small defect in the reticle does not get resolved in the PR due to image reduction.

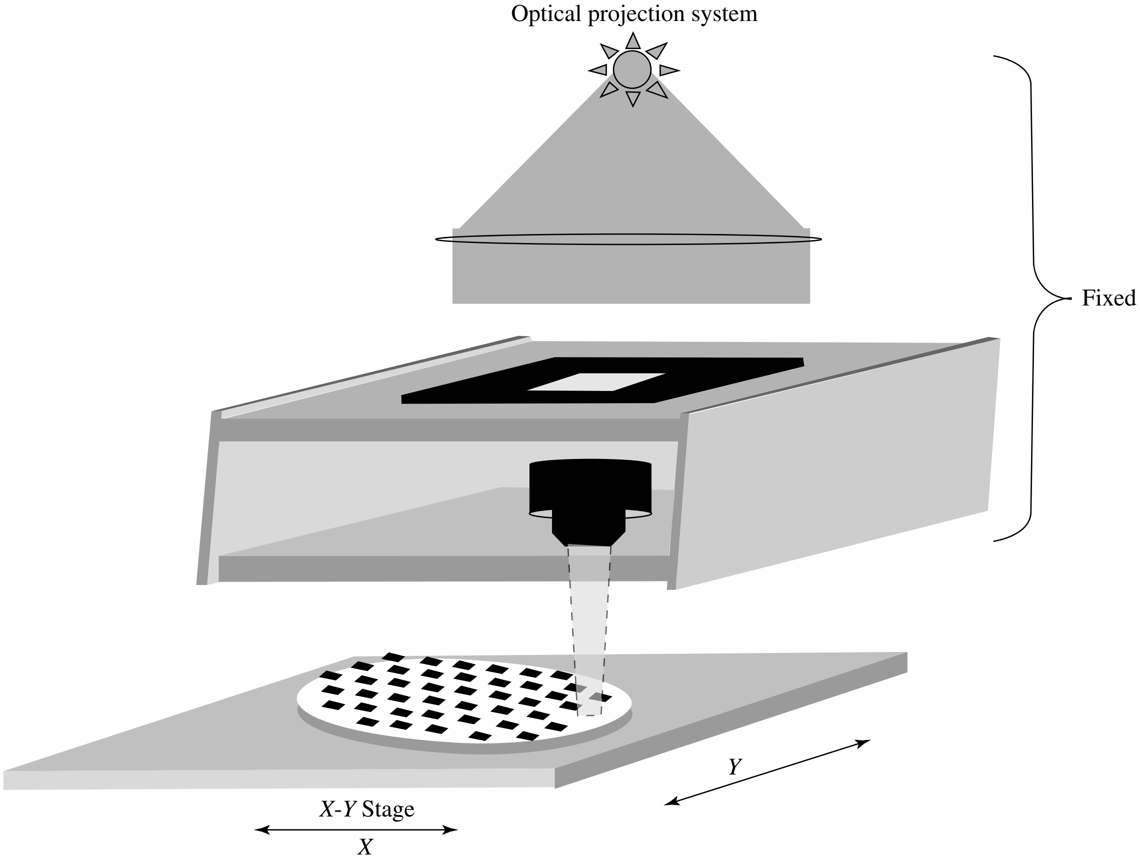


Fig. 6.15 *M*:1 projection system

6.9 NON-OPTICAL PROJECTION LITHOGRAPHY TECHNIQUE

6.9.1 Electron-Beam Lithography

The electron-beam (e-beam) lithography technique is an altogether different technique from the optical lithography technique, as shown in Fig. 6.16. The wavelength of the electron is much shorter than the X-ray wavelength; thus the resolution is higher than in the X-ray lithography. The wavelength of the electron is a function of the voltage of the electrodes, and can be expressed as

$$\lambda = \sqrt{\frac{e/m}{V}} = \sqrt{\frac{150}{V}} \times 10^{-8}\,\text{cm} = \frac{1.23}{\sqrt{V}} 10^{-8}\,\text{cm}$$

where V is the voltage between the cathode and the anode electrodes and e and m are the charge and mass of the electron respectively.

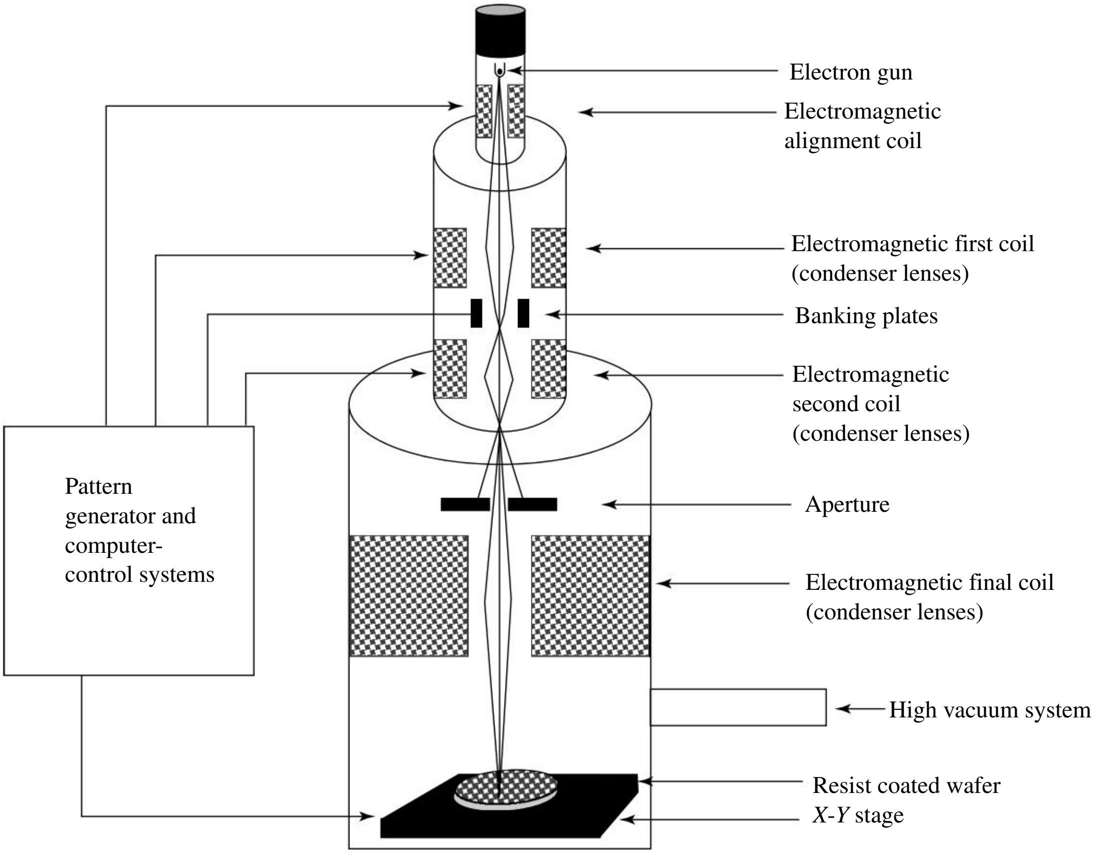

Electron gun
Electromagnetic alignment coil
Electromagnetic first coil (condenser lenses)
Banking plates
Electromagnetic second coil (condenser lenses)
Aperture
Electromagnetic final coil (condenser lenses)
High vacuum system
Resist coated wafer
X-Y stage
Pattern generator and computer-control systems

Fig. 6.16 Electron beam lithography system

The electron wavelength is around 10^4 orders shorter than the wavelength of the ultraviolet light at 15000 volts (see Problem 2). Hence, the theoretical resolution of the electron-beam lithography is around 10^4 orders higher than that of the UV lithography. The resolution is further increased due to lower numerical aperture. Generally, the numerical aperture of the electron beam is around 0.01, i.e., ~50,000 smaller than the 100 × optical lens. Furthermore, the electron beam lithography system has high depth of focus because of lower NA. Hence, the tolerance of wafer placement at the focal plane is not very stringent. The electron beam can be focussed from 0.01 to 0.05 micrometre diameter. Unfortunately, the electrons from the electron beam scatter in the resin, and that degrades the resolution of the patterns. In addition, electron beam lithography also suffers from backscattering and proximity effects. The proximity effect occurs due to the scattering of the electron beam, and that results in insufficient electron beam intensity at the corner of the patterns, which leads to corner rounding of the patterns. In spite of these drawbacks, the electron beam technique can resolve in the sub-micrometre range of patterns and is used extensively in VLSI and ULSI applications.

The electron beam lithography is most suitable for chrome mask fabrication. The details of electron beam mask making have been discussed in Chapter 5. Apart from chrome mask fabrication, the electron beam is also used to generate the patterns on the wafer directly. This process is called **Direct Writing on the Wafer** and in short, DWW. The DWW technique is most suitable for IC design validation and for custom made ICs where large volumes of ICs are not required. The IC made by the DWW technique does not require the masks. This saves a lot of time as well as the fabrication cost of the IC.

The electron-beam lithography is done in vacuum. As a result, there is a less chance of incorporation of the impurities during resist exposure. Unfortunately, the electron-beam lithography equipment develops faults very frequently, so it needs constant maintenance. In addition, the electron beam takes large exposure time; therefore, the mask or wafer throughput (number of wafers or masks per hour) is very low. The electron-beam lithography system is fully automated, elaborate, and costly. Apart from these limitations, the e-beam lithography suffers from many other limitations such as low electron beam current, beam density variation, and beam charging, etc.

Electron beam exposure is done in two modes, namely, raster scanning and vector scanning modes These modes have been explained in Section 5.3.2 and are being described again. In the raster scanning mode, the electron beam scans in line from one end of the mask (or wafer) to the other end. This scanning mode is similar to the cathode ray oscilloscope. At the time of pattern exposure, the e-beam falls on the resist; otherwise, the electron beam is kept out of (gated off) the mask or by using the electron-beam blanking mechanism. In the vector scanning mode, the electron beam exposes patternwise. It exposes one pattern completely and then it moves to expose the other pattern. In general, electron-beam lithography is used in the scanning mode. For exposing a larger mask or a bigger wafer, the combination of the *x*-*y* stage and the e-beam deflection mechanisms is used, as the e-beam cannot be deflected beyond a point. More about electron-beam lithography is described in Chapter 11.

6.10 ION-BEAM LITHOGRAPHY

The resolution of the pattern can be further increased using ion-beam lithography. The wavelength of an ion is even shorter than that of an electron. In addition, the ion-beam shows less scattering and backscattering effects in the resist as compared to the electron beam, due to heavy ion mass. The ion-beam can also be used for direct writing on the wafer. The Polymethyl Methacrylate (PMMA) is used as the resist for the ion-beam lithography. Ion-beam lithography takes less time for exposure than e-beam lithography. Unfortunately, the ion-beam is the most difficult to deflect and focus. Therefore, the ion-beam system is highly complicated and costlier as compared to the e-beam system. Ion-beam lithography is most suitable for MOS transistors in the ULSI generation of the IC. Ion-beam lithography is still in the developing stage.

Summary

The role of the photoresist (PR) is to protect the film below from the film etchant solution. The unprotected film reacts with the etchant and is removed from the wafer. Once the film etching is complete, the PR is removed completely from the wafer, and the protected film patterns remain on the wafer.

Usually, PR coating is done by two techniques, namely, spray coating and spin coating. The spin-coating technique is more popular than the spray-coating technique. Normally, the thickness of the coated PR is around 1 micrometre. After the PR is coated, the wafer is pre-baked for hardening. After PR exposure, the wafer is developed, rinsed, washed, dried, and post-baked. Post-bake promotes adhesion to the wafer and increases resistance to the etchant. Then, the film is etched, the PR is stripped, and the wafer is processed for second mask lithography; the patterns of the second mask are aligned with the patterns created by the first mask on the wafer.

Both positive and negative PRs contain organic polymer material, photosensitive material (dye), and solvent, but they are complimentary in nature. Negative PR swells in water and loses its resolution significantly, so it is not used for high-density VLSI and ULSI applications. For higher optical resolution, shorter wavelength and higher numerical aperture is required.

The lithography exposure techniques are divided into two categories, namely, the optical technique and the non-optical technique. In the optical technique, the PR is exposed to optical light in the UV range. The optical lithography technique is further divided into categories, namely, shadow and projection printing. In the non-optical technique, X-ray and electron beam techniques are used, where different kinds of polymers are used and these polymers are called the resists.

The X-ray wavelength is around 10^3 times shorter than the UV light wavelength. Therefore, the pattern resolution of the X-ray is expected to be 10^3 times higher than that of the UV light, but the resolution degrades due to the scattering and the secondary generation of the X-ray inside the resist. In addition, the mask is fragile and mask engineering is needed. In spite of these drawbacks, the electron-beam technique can resolve in the sub-micrometre range of patterns and is used in the VLSI and ULSI applications.

The electron wavelength is around 10^4 orders shorter than the wavelength of the ultraviolet light at 15000 volts. Hence, the theoretical resolution of the electron beam lithography is around 10^4 orders higher than that of the UV lithography. The resolution is further increased due to lower numerical aperture. Generally, the numerical aperture of the electron beam is around 0.01, i.e., ~50,000 times smaller than the 100 × optical lens. Furthermore, the electron beam lithography system has high depth of focus because of lower NA. Unfortunately, the electron-beam lithography suffers from backscattering and proximity effects during exposure. In spite of these drawbacks, the electron-beam technique can resolve in the sub-micrometre range of patterns and is used extensively in VLSI and ULSI applications. The electron beam lithography is most suitable for chrome mask fabrication and for direct writing on the wafer. The DWW technique is most suitable for IC design validation and for custom-made ICs, where large volumes of ICs are not required.

References

- David J Elliott, *Integrated Circuit Fabrication Technology*, McGraw-Hill, 1982
- J D Plummer, M Deal and P B Griffin, *Silicon Fundamental Technology: Fundamentals, Practice and Modeling*, Prentice Hall, 2000
- R C Jaeger, *Introduction to Microelectronic Fabrication: Volume 5 of Modular Series on Solid State Devices*, Prentice Hall, Second Edition, 2001
- S M Sze, *VLSI Technology*, Second Edition, McGraw-Hill, 1988
- S K Gandhi, *VLSI Fabrication Principles*, Second Edition, Wiley, 1994
- D Nagchoudhuri, *Principles of Microelectronics Technology*, Wheeler, 1998
- S A Campbell, *The Science and Engineering of Microelectronic Fabrication*, Oxford University Press, 1996

Multiple-Choice Questions

6.1 During lithography, to which light is the wafer exposed to?
(a) Green (b) Red (c) Ultraviolet

6.2 In the first lithography, the scribe track of the mask is aligned with
(a) wafer cut (b) previously printed die(c) mask alignment mark

6.3 After the first lithography, the mask alignment mark is aligned with
(a) wafer cut (b) scribe track (c) mask alignment mark

6.4 Resolution of the negative photoresist is better than that of the positive photoresist.
(a) True (b) False (c) It is the same

6.5 Which photoresist swells in water?
(a) Negative photoresist (b) Positive photoresist

6.6 To increase mask life, which exposure technique is better?
(a) Contact printing (b) Proximity printing (c) Projection printing

6.7 Which exposure has better resolution?
(a) Contact technique (b) Proximity technique (c) Projection technique

6.8 Electron beam exposure is used
(a) to expose the emulsion of the photographic plate
(b) to expose the photoresist
(c) none of them

6.9 Positive photoresist developer is made of
(a) alkaline solution (b) acidic solution (c) petrochemical solution

6.10 Generally, what is the photoresist thickness after wafer coating?
(a) 1–2 microns (b) 5–6 microns (c) 10–11 microns

6.11 Generally, what is the range of the pre-bake temperature?
(a) 40–50°C (b) 80–90°C (c) 120–140°C

6.12 Generally what is the range of the post-bake temperature?
(a) 40–50°C (b) 80–90°C (c) 120–140°C

Descriptive Problems

6.1 If the gap between the wafer and the mask is kept at 40 μ, and the resist is exposed to an X-ray of 4.4 Å wavelength, calculate the theoretical resolution.

6.2 If optical constant is not considered then compare the resolution between a 100X optical system and an e-beam at 15 kV where, the optical numerical aperture is 1.5, the wavelength is 436, and the electron charge is 1.60×10^{-19} coulombs, mass is 9.10×10^{-31} kilograms and the numerical aperture is 0.01.

6.3 Compare the depth of focus of a 100X optical lens and an e-beam at 15 kV.

6.4 State the reasons why X-ray lithography is not popular.

6.5 Justify that e-beam lithography plays a vital role in VLSI fabrication.

CHAPTER

7

Etching

7.1 INTRODUCTION

Thin localised diffused areas inside the silicon and stacks of films on the silicon wafer, is the basic structure of a MOS transistor. These localised diffused areas and films are made by lithography process followed by the etching process. The lithography process has been covered in the previous chapter, and in this chapter, details of the etching process are described. Etching is one of the most critical process steps in IC fabrication. Generally, the etching process is done to remove the film from selected areas. The film removed from the localized areas is used for diffusion, contact, hole in dielectric, and trench isolation. On the other hand, the retained film patterns are used for MOS gate, capacitor, electrical isolation, mask for diffusion, and other applications.

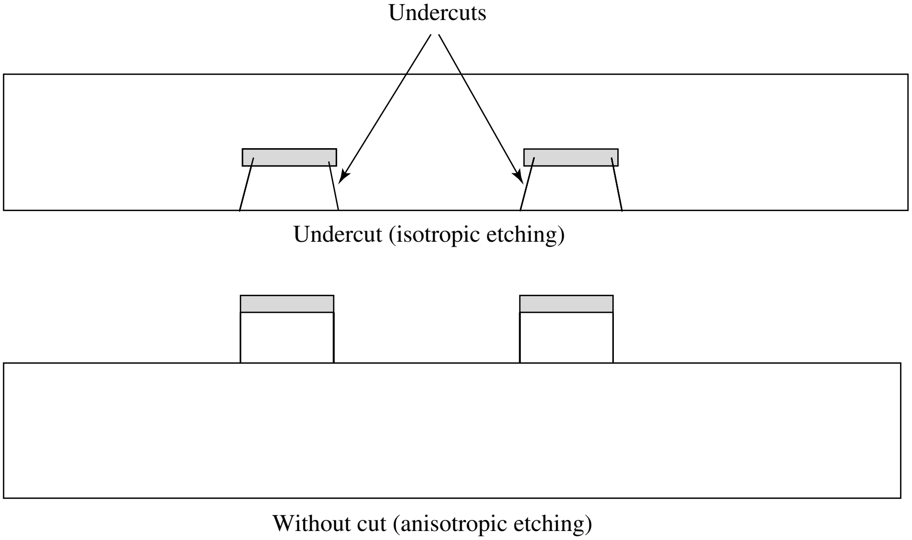

Undercuts
Undercut (isotropic etching)
Without cut (anisotropic etching)

Fig. 7.1 Etching process

There are many types of films required for IC fabrication. These films are either deposited or grown on the wafer. Thus, different etching techniques are used for etching different types of films. In the ideal condition, a particular film should be etched leaving the other films. In addition, the film should not have undercut below the sides of the PR pattern as shown in Fig. 7.1; otherwise the MOS dimensions will be lost. In the ideal case, the film should be etched without the undercut as shown in Fig. 7.1. Unlike the lithography process, if etching is not done satisfactorily, the chances of wafer recovery are very remote; hence, it is carried out with great attention and precautions.

7.2 ETCHING TECHNIQUES

Film etching techniques can be divided into two categories: **wet etching** technique and **plasma (dry) etching** technique. In the wet etching technique, a liquid chemical (or solution) is used to etch the film; therefore, this etching technique is called the wet etching technique. Wet etching is isotropic in nature which means that it etches in the horizontal direction as well as in the vertical direction that results in undercut below the PR, as shown in Fig. 7.1. This undercut is undesirable in IC fabrication, especially for high density ICs. The undercut also leads to the wastage of silicon. Many a times, the undercut is around 0.7% of vertical etching. Apart from this serious limitation, the wet etching process is also a function of the dimensions of the pattern size. In addition, it is difficult to etch the patterns below a certain limit. For example, wet etching is not recommended below ~1.5 micrometre geometry of MOS fabrication. The etching is faster in bigger etching windows (patterns) than the smaller etching windows (patterns).

The plasma (dry) etching is done in vacuum by ions or reactive atoms in a high electric field. As no liquid is used in the plasma etching film, therefore, this process of film etching is called **dry etching**. The film is etched anisotropically (vertically) with insignificant undercut. For this reason, dry etching is preferred for today's high-density VLSI and ULSI applications.

7.2.1 Wet Etching Mechanism

Wet etching mechanism is an old process of film (or material) etching. The chemical etching chemistry is well developed and well understood. In the chemical etching process, the top surface of the film is first oxidised using a chemical or a solution, and then the oxide layer is dissolved in the solution. The etching solution used is called the **etchant**. Many a times, the pH value of the etchant is maintained by adding an appropriate chemical into it, in order to maintain a constant etch rate of the film. This type of etchant is called the **buffer etchant**.

7.2.2 Etching Parameters

There are many etching issues related to wet etching. These issues are mainly selectivity, etch rate, undercut, uniformity, damage, and contamination. In Chapter 3, we have seen that hydrofluoric acid (HF) is used for silicon oxide etching. It is desired that the HF does not etch any film other than silicon oxide. In other words, the HF should neither etch the masking PR nor the silicon oxide material below. But in reality, this never happens and almost all the materials are etched more or less by the etchant. The ratio of the rate of etching of different materials in an etchant is called the **etching selectivity** or in short, the **selectivity**. The etch rate of the film should neither be too fast nor too slow. In case of fast etching rate, the control of the etching rate is difficult to attain and it may lead to many serious problems. On the other hand, a slow etching rate may take longer etch time and it may also lead to other types of problems. Therefore, the film etch rate is optimised to a moderate etch rate. The film should be etched completely over the entire wafer. In reality, this situation never occurs, because of two main reasons. Firstly, the etch rate is higher in bigger areas (windows) than in the smaller areas, and secondly, because of the non-uniformity of film thickness over the wafer. Therefore, to ensure complete film etching, extra etching time is always allotted. This extra etching time is called **over etching**.

7.2.3 Wet Etching Parameters

There are four main etching parameters that are involved in the wet etching process. These etching process parameters are

1. Concentration of the etchant
2. Etching time
3. Temperature of the etchant
4. Agitation of the etchant

Concentration of the Etchant

The etch rate of the film is highly dependent on the concentration of the etchant. It has been found that etching is nonlinear with respect to the etchant concentration, as shown in Fig. 7.2(a). Hence, the etchant concentration has to be optimised prior to the etching process. To maintain the etchant concentration (pH value), buffer etchant is used during film etching.

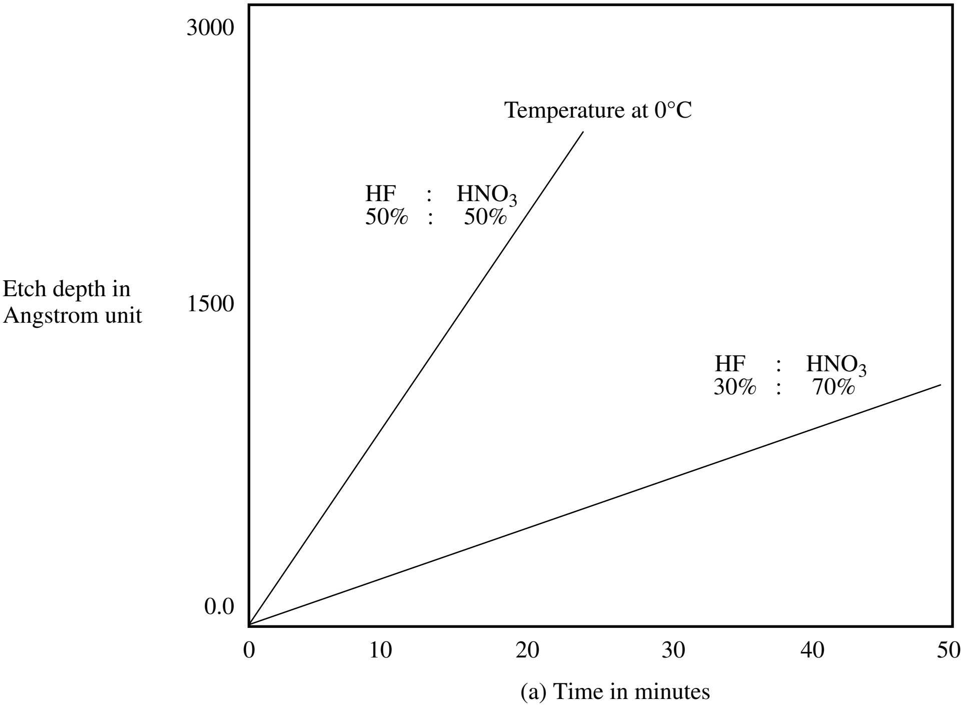


Fig. 7.2(a) Typical etching rate of boron-doped oxide with etchant concentration

Temperature

The etch rate is highly nonlinear with respect to temperature and it has significant effect on the etch rate, as shown in Fig. 7.2(b). Therefore, wet etching is carried out at a fixed etchant solution temperature. Generally, etching is done at room temperature, and the etchant temperature is maintained meticulously using a constant temperature bath.

Agitation of the Etchant

The main purpose of etchant agitation is to create turbulence in the etchant. The turbulence helps reacted products push away from the film surface and bring fresh etchant at the film surface. The agitation has immense impact on the film etch rate. It has been found that the etch rate is highly nonlinear, as shown in Fig. 7.3.

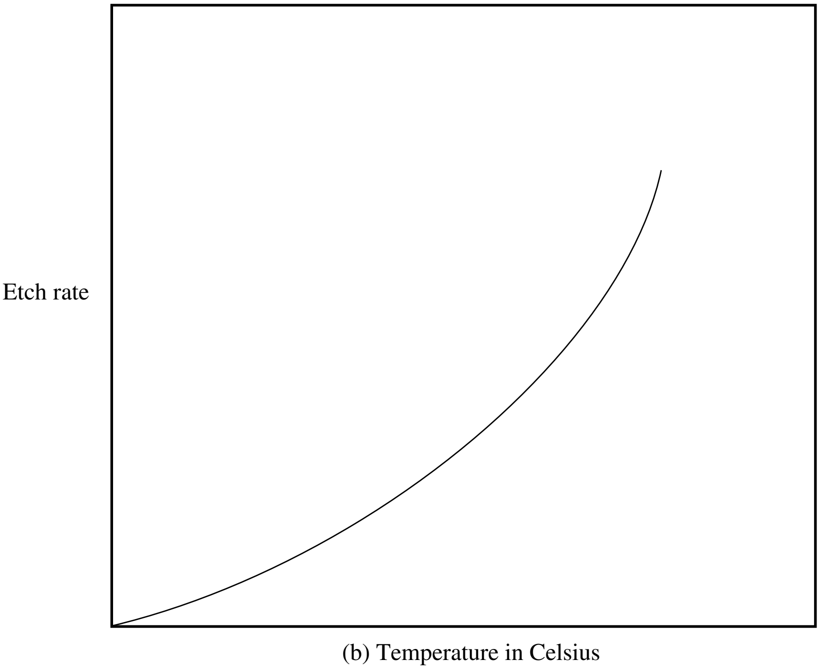


Fig. 7.2(b) Typical etching rate of oxide with temperature

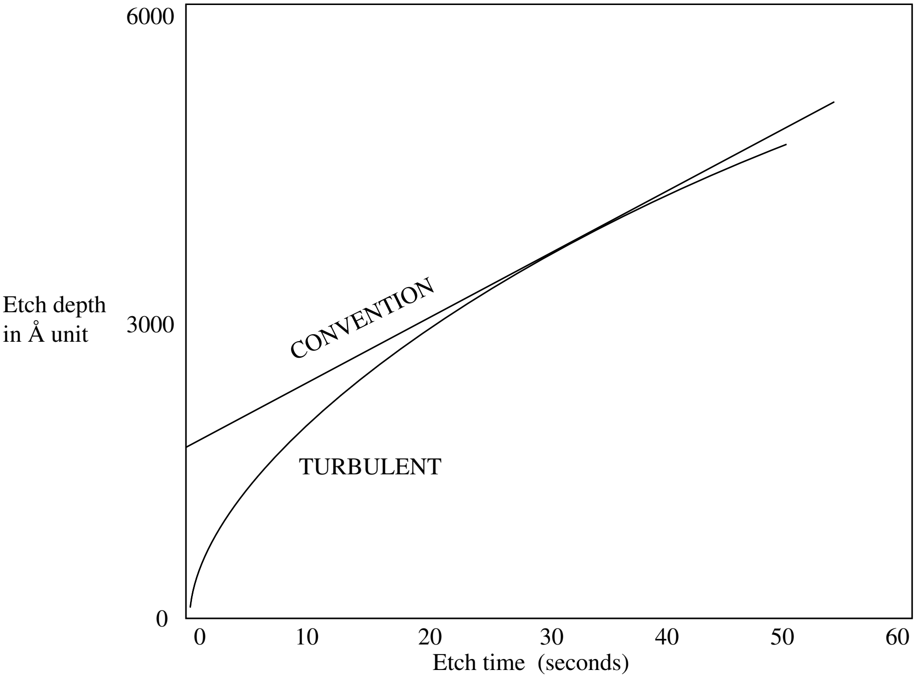


Fig. 7.3 Typical etching rate with etchant agitation

Generally, the etchant is agitated by four ways: thermal convection, mechanical spray, ultrasonic and bubble. In thermal convection, movement of the etchant and the reacted product is due to the thermal gradient generated by heating the etchant from the bottom. The movement of the etchant is slow in the thermal convection process and it leads to slow etch rate; hence, undercut is significant. In the mechanical agitation process, the etchant movement is generated by mechanical force. Generally, the mechanical agitation is created using a mechanical stirrer. Mechanical agitation of the chemical may also be created by ultrasonic technique. In this technique, the wafer is dipped in the chemical container and the whole container is vibrated using an ultrasonic vibrator. The film etch rate using mechanical agitation is found to be higher by 25%; and film undercut is much less than in thermal etching. Another way of creating chemical agitation is by **spray etching**. In this process, a chemical is sprayed on the film using a spray machine. The movement of the chemical on the film is created by the spray etchant force. A uniform etching with less undercut is observed by this spray technique. In the bubble agitation technique, the chemical agitation is created by passing nitrogen gas into the etchant solution that leads to turbulence in the etchant. This technique is expensive due to the nitrogen gas consumption, and inferior to the mechanical and spray etching techniques in view of the undercut. Hence, the bubble technique is hardly used in IC processing.

Etch Time

The etch rate is found to be linear with the etch time, as shown in Fig. 7.4. This property of etching is exploited in the wet etching process. As the etch rate of the film is linear with time, the film etching can be controlled by the etch duration. For example, if the thickness of the film is 1000 Å and the etch rate of the film is 200 Å per minute then a time of 5 minutes will be required to etch out the film completely from the wafer. But it has to be ensured that all the other etching parameters, such as concentration, agitation, and temperature remain constant during the entire etching process.

It is important to mention here that the film etch rate depends on whether the film is undoped or doped. Furthermore, the etch rate also depends on the type of dopant element present in the film. In addition, the concentration of the dopant in the film also dictates the etch rate. Apart from the film being doped or undoped, the etch rate also depends on the quality of the film. For instance, the etch rate of a thermally grown silicon dioxide film is slower than that of the deposited silicon dioxide film. This is

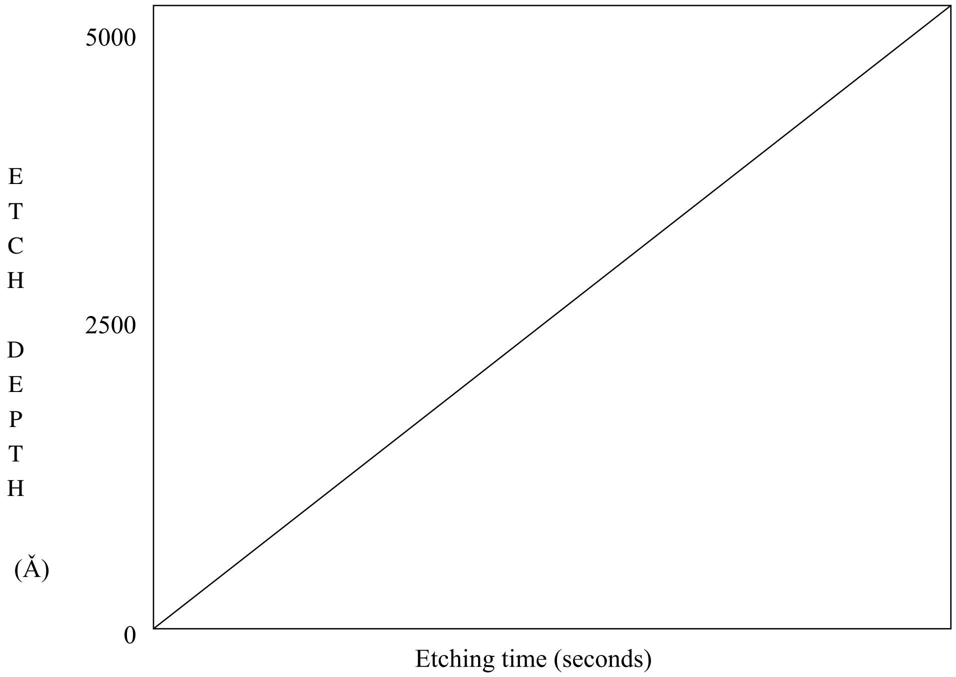


Fig. 7.4 Typical etching rate with etching time

because the deposited silicon dioxide is less dense and inferior to the thermally grown silicon dioxide. Furthermore, the etch rate is highly subject to film damage.

7.3 WET ETCHING OF COMMON FILMS

7.3.1 Silicon Dioxide (SiO_2) Etching

It has been mentioned previously that the etch rate of the grown silicon dioxide is different than that of the deposited silicon dioxide. Generally, whether the silicon dioxide is grown or deposited, is etched by the HF acid at room temperature. The etching selectivity between the PR and the oxide is very high; therefore, the PR works as an excellent mask for oxide etching in the HF. Many a times, ammonium fluoride is added in the HF to maintain the pH value for better etch control. Table 7.1 below shows the typical propositions of the oxide etching solution and the approximate etch rate of the grown oxide and the deposited oxide.

Table 7.1 Typical proposition of oxide etching solution and etch rate

HF	Ammonium fluoride	Water	Etch rate	
			Grown oxide	Deposited oxide
30 ml	110 g	200 ml	1000 Å/min	3000 Å/min
50 ml	30 g	1000 ml	150 Å/min	500 Å/min

Example 7.1

Give the importance of overetching.
Silicon is oxidised and a 1 ± 0.2 μm thick dioxide is grown on the silicon wafer. If the oxide etch rate is 1000 Å per minutes in the HF then determine the total time including the overetch time.

Answer 1 μm = 10,000 Å
The etch rate of silicon dioxide in the HF is 1000 Å.
Hence, to etch 10,000 Å thick oxide, it will take 10 minutes of time.
Furthermore, to etch 0.02 μm due to the non-uniformity of oxide thickness, it will take 2 minutes of extra time.
So the total etching time will be 12 minutes.
The 1 μm oxide will also be etched, and the minimum thickness will be 8000 Å.

It is important to note that the total oxide is etched in 10 minutes, but to be on the safe side, an extra 2 minute is given. In this case, it will further reduce the oxide thickness to 8000 Å. So, during IC processing, one has to take care of these facts.

It is important to mention that the HF reacts with the glassware; therefore, etching is carried out in special types of polymer containers. Furthermore, the HF is highly reactive; hence, it should be handled with great care.

7.3.2 Silicon Nitride (Si_3N_4) Etching

The silicon nitride etching is done using phosphoric acid at ~180ºC. Unfortunately, the PR cannot withstand this etching temperature; hence, it is not used as an etching mask. As the etching selectivity between silicon nitride and silicon oxide is excellent in case of phosphoric acid, hence, an oxide mask is used for silicon nitride etching. This needs extra process steps of silicon oxide deposition, lithography and silicon oxide etching. Generally, phosphoric acid is used for the global nitride film etching from the wafer. The etch rate of silicon nitride in phosphoric acid is around 60 Å/min. Silicon nitride film etching is carried out in a reflux system to save the phosphoric acid, as shown in Fig. 7.5. In the reflow system, wafers are kept in a cylindrical shaped glass container, and thereafter, phosphoric

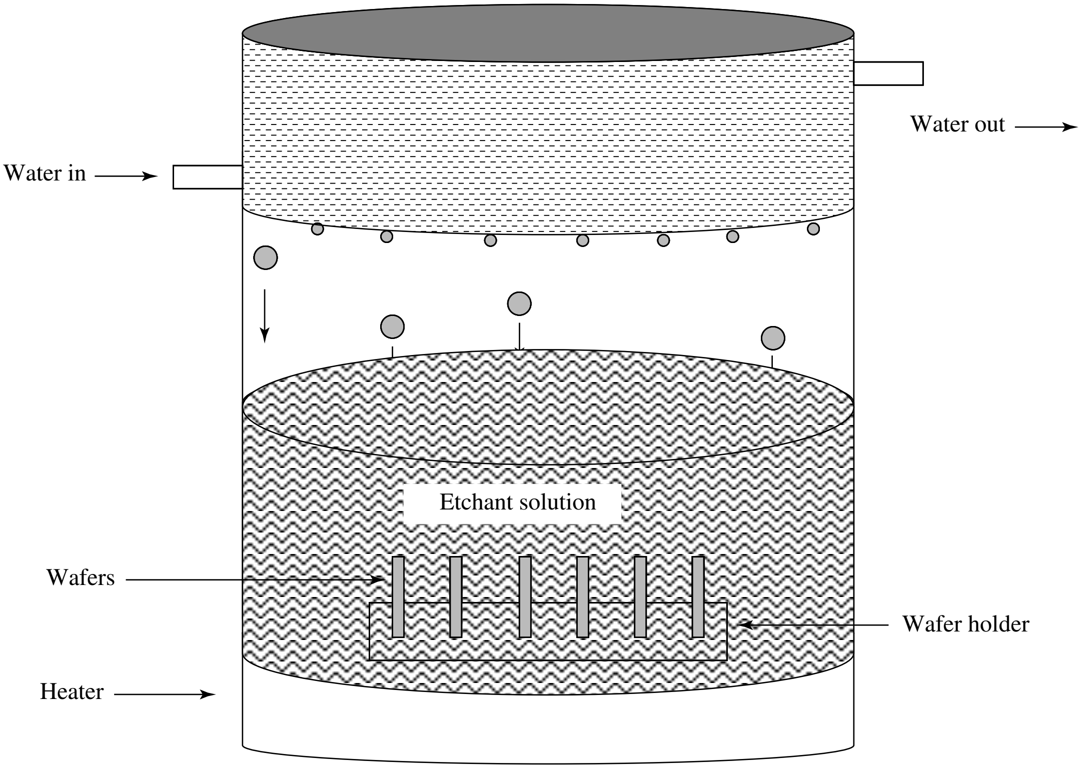


Fig. 7.5 Reflow system

acid is poured in the container till the wafers are submerged completely. Thereafter, a water circulating jacket is placed on top of the container. Then, the container is heated from the bottom. Once the phosphoric acid is heated, nitride etching takes place. When the phosphoric acid is hot, it evaporates and its vapours are produced. These phosphoric acid vapours come into contact with the water-cooled jacket, and condense and fall back in the container in the form of droplets. This saves the phosphoric acid and also helps in keeping a check on environmental pollution.

Silicon nitride can be etched in phosphoric acid mixed with fluoroboric acid at ~105°C temperature. Its etch rate is around 100 Å/min. A special type of PR is used as a mask. In this case, the PR has to be post-baked at 140°C to 160°C prior to etching. Unfortunately, a slight change in phosphoric acid and fluoroboric acid ratio affects the selectivity drastically.

7.3.3 Polysilicon Etching

Commonly, a mixture of nitric acid and diluted HF chemicals is used for polysilicon etching. Acetic acid can also be added to maintain the pH value of the etchant. The etch rate of polysilicon is very high; this is because the polysilicon film is made of polysilicon grains. For high selectivity, the ratio of the chemicals has to be optimised carefully.

7.3.4 Aluminium Film

Generally, the aluminium etchant solution is a mixture of phosphoric acid and nitric acid. Nitric acid converts the surface of the aluminium into an aluminium oxide layer, and then phosphoric acid dissolves the aluminium oxide. For a moderate etch rate, the phosphoric and nitric acid solution is diluted in water. To maintain the pH value of the solution, acetic acid is added into the etchant. One of the popular aluminium etchant compositions is: phosphoric acid (80 ml): nitric acid (5 ml): acetic acid (5 ml): water (10 ml). The etch rate of the aluminium film is around 2000 Å/min at 25ºC. Etching of the aluminium line less than 1.5 micrometre is not advisable in wet etching. The selectivity of the aluminium film and the PR is very high; hence, the PR is generally used as a mask.

7.3.5 Doped Silicon

Etching also depends on whether the silicon is doped or undoped, and the type of dopant. The etch rates of the phosphorus and the boron doped silicon are shown in Figs. 7.6 and 7.7 respectively. The different etch rates of these doped silicons becomes a serious issue, because it may alter the dopant diffusion profile significantly, especially when the MOS transistors is scaling down (see Chapter 8).

7.4 PLASMA ETCHING (DRY ETCHING)

The wet etching process is simple, inexpensive, and has good selectivity. In addition, wet etching is a batch process that leads to a very high wafer throughput. Unfortunately, wet etching severely suffers from undercut (lateral etching). The undercut limits the use of wet etching at lower dimensions. Therefore, to avoid the undercut, the plasma (dry) etching technique is inducted in IC processing. In the plasma technique, the film is etched either by physically knocking out the atoms of the film material using high energised ions, or by chemical reaction by the reactive chemical species. These ions or the reactive chemical species are produced in an electric discharge called the **plasma**. Plasma etching is carried out in a low pressure chamber having high electric field. Plasma etching is also commonly known as **dry etching**. The principle of plasma film etching and deposition is covered below.

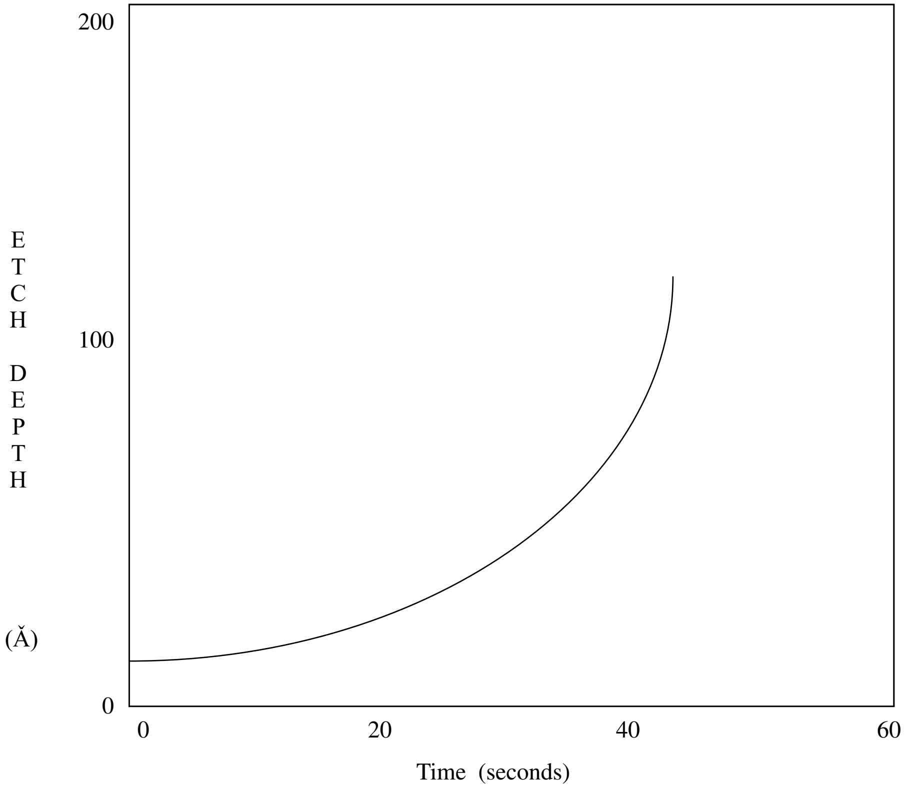


Fig. 7.6 Typical etching rate with phosphorus-doped silicon

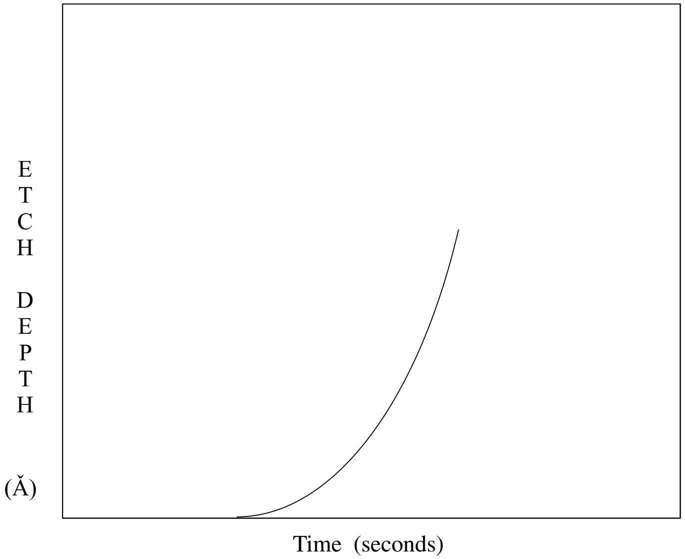


Fig. 7.7 Typical etching rate with boron-doped silicon

7.4.1 Plasma Fundamentals

Generally, the air discharge is called the **plasma**, and the discharge region of the air is called the **plasma region**. The plasma region contains a large numbers of ions, electrons and neutral gas atoms. For MOS fabrication, the plasma is created in between the two parallel metal electrodes, which are separated at a distance in a partially evacuated chamber by applying high voltage across the electrodes as shown in Fig. 7.8. The mobility of the electrons is higher than that of the ions due to their light weight. Thus, the electrons are collected faster at the anode electrode than the ions are collected at the cathode electrode. This results in a depletion of electrons with respect to the ions in the plasma region. Thus, the plasma region possesses a positive potential with respect to both the electrodes. To sustain the plasma, the lost electrons are supplemented by newly generated electrons in the plasma because of the electric field. The condition that is required to sustain the plasma (discharge) between the electrodes is called the **plasma condition**.

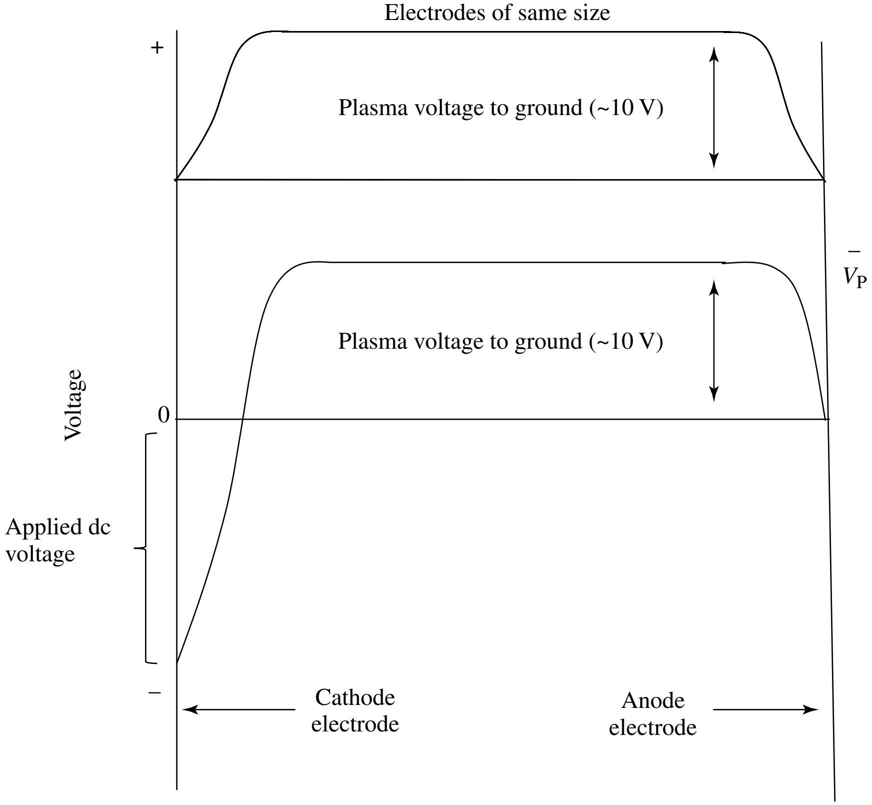


Fig. 7.8 The dc voltage distribution when cathode electrode is at high dc voltage w.r.t. anode electrode

In the plasma etching process steps, the wafer is placed on the cathode electrode and then the chamber is evacuated initially to a high vacuum which is lesser than 10^{-6} torr. Thereafter, the chamber pressure is increased to the range of 10^{-2} to 10^{-3} torr by introducing an inert (argon) gas. Once the chamber pressure is stabilised, high potential in the range of kilovolts is applied between the electrodes. Generally, a few thermally generated electrons and ions are present in the chamber. These electrons and ions are attracted towards their opposite directions due to the high electrical field. When electrons travel with great speed towards the anode electrode, they collide with neutral argon atoms and produce significant electrons and ions by ionising the atoms. Usually, the temperature of the electrons in the plasma ranges from 10^4 to 10^5 K, but the temperature of the plasma region remains in the range of 50 to 100°C. This phenomenon can be explained through an example: at 1 torr chamber pressure, the plasma contains 10^9 to 10^{12} cm^{-3} of electrons, which is 10^4 to 10^7 times less than the neutral gas atoms concentration. Therefore, the total plasma region temperature comes down in the range of 50 to 100°C. This temperature is considered to be low in the context of IC fabrication and does not change the dopant distribution.

7.4.2 Plasma Generation Techniques

Plasma can be generated by high dc or RF voltage sources in a partially evacuated chamber, and they have their own merits and demerits. In addition, the mechanisms and system configurations of these two modes of plasma generation are described below.

dc Discharge

The physics of discharge (plasma) produced by the high voltage dc of equal size electrodes can be explained through Fig. 7.9. Let us consider that one end (say left side) of the electrode of the discharge tube is connected to negative (cathode) voltage and the other end of the electrode is connected to positive (anode) voltage. The tube is partially filled with argon gas. When a high voltage is applied to both the electrodes, the thermally generated ions and electrons present in the tube are accelerated towards the cathode and the anode respectively. The accelerated ions collide with the cathode electrode with great force, and produces ions, electrons and photons. Similarly, the electrons strike the anode with great speed and generate ions, electrons and photons. As the ions and the electrons strike the cathode and the anode with great speed, material of both the electrodes sputter (etch out). The material of both the electrodes etch out almost equally. Apart from the sputtering of the metal of the cathode and the anode electrodes, photon emission also takes place and that makes the electrodes region to glow. Hence, these are called the **cathode glow** and

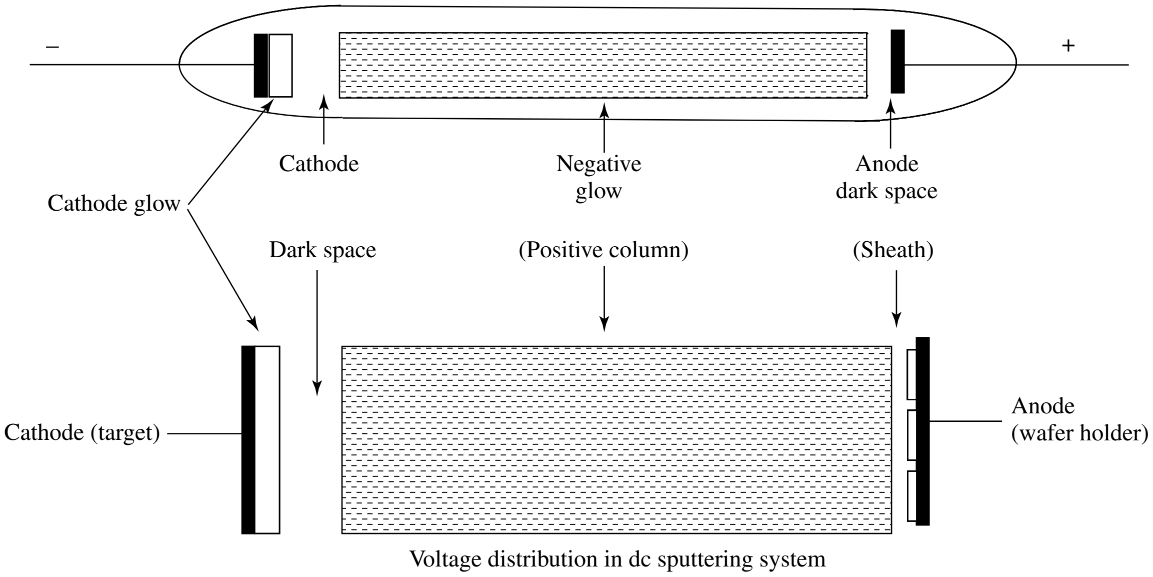


Fig. 7.9 Glow discharge in discharged tube

the **anode glow** regions. The cathode glow region is electrically conducting. Electrons generated near the cathode electrode move fast towards the anode electrode. These newly generated electrons, just near the cathode electrode, do not possess sufficient kinetic energy; hence, no photons are produced, and a dark region is formed. This region of discharge is called the **cathode dark space** or the **cathode sheath**. This cathode sheath is also called **Crookes dark space**. The cathode dark space possesses very less number of ions and electrons; therefore, it is almost electrically non-conducting. Similarly, a dark space is also formed just near the anode electrode and this dark space is called the **Faraday dark space**. The electrons, after leaving the cathode dark space, are accelerated towards the anode electrode due to high voltage of the anode, and gain enough kinetic energy to ionise the neutral argon gas atoms and generate large number of ions, electrons and photons. This region of discharge is called the **negative glow**. The negative glow region has the properties of constant positive potential (plasma voltage) due to the depletion of electrons, and electrical conductivity with low resistance. As the electrons keep colliding with the neutral gas atoms in the negative glow region, they constantly lose their kinetic energy and finally cannot ionise the neutral gas atoms, as they approach the anode electrode. Hence, again a dark region is formed. This dark region is called the **anode dark space** or the **anode sheath**. This sheath is also called the **Faraday dark space**. This anode dark space has low electrical conductivity similar to the cathode dark space. As a result, the negative glow is bound by two non-conducting regions.

Radio Frequency (RF) Discharge

When the ac frequency is in a lower range, then the polarity of the electrodes varies very slowly and the discharge pattern is similar to the dc discharge pattern. This is because the rate of change of ac polarity is slower than the mobility of the electrons and the ions; hence, the electrons and ions have enough time to respond to the ac cycle. Therefore, the pattern of the plasma discharge will be the same as the dc discharge pattern except that the Crookes dark space and the Faraday dark space will interchange their positions during the ac cycle. When the ac frequency is increased in the range of RF, the electrons and ions cannot travel from one electrode to another, hence, they oscillate between the electrodes in the negative glow region and a steady pattern is observed, as shown in Fig. 7.9.

The RF plasma has many advantages over the dc plasma. In the RF frequency, the electrons gain sufficient kinetic energy to ionise a large number of neutral gas atoms in its oscillation region and these newly produced electrons and ions further produce electrons and ions in the oscillation region and that in turn produce a large numbers of electrons and ions. This phenomenon is called **cascading effect** and it compensates for the electrons lost in constituting the plasma current and thus, sustains the plasma condition. This is the most essential requirement for material etching (or film deposition by the sputtering method as is explained in Chapter 10). The RF current can pass through both the non-conducting regions at the cathode electrode and the anode electrode (even if covered with dielectric material) and sustain the plasma. This is not possible in case of dc plasma, as dc current cannot pass through the dielectric material. Hence, the dc plasma is limited to the etching of the conducting materials.

Plasma Voltage Distribution

In the negative glow region, the mobility of the electrons is much higher than that of the ions, so the rate of collection of electrons by the anode electrode is much higher than the rate of collection of ions by the cathode electrode. Furthermore, the negative glow is almost isolated by the two Crook and Faraday non-conducting dark spaces as explained. Hence, the negative glow develops a positive potential V_p of around 10 volts across the two dark regions, as shown in Fig. 7.8. The negative glow being positive with respect to both the electrodes, the ions and the electrons are attracted and collected at their respective electrodes resulting in a loss of electrons and ions from the negative glow. As the negative glow loses the electrons and ions, it is not possible to sustain the plasma condition. To sustain the plasma, the lost electrons and ions must be replenished by ionising fresh neutral gas atoms using an external potential. For this reason, a high dc voltage is applied to the electrodes so that a large number of electrons and ions are generated at the electrodes to sustain the plasma condition. Furthermore, the voltage difference between the negative glows with respect to both the electrodes is the same; hence, the etching (sputter) rate of both the electrodes is almost the same. As the ions are heavier than the electrons, the etch rate of the cathode electrode is relatively higher than the anode electrode. Therefore, the wafer is always kept on the cathode electrode for film etching. To enhance the film etch rate, the cathode electrode is kept at significantly higher potential than that of the anode electrode. To further increase the etching rate, the cathode is made much smaller than the anode electrode. This electrode configuration and their electrical potentials are described below.

7.4.3 RF Plasma Etching System

Generally, film etching is carried out in the RF source because of certain advantages over the dc source, especially for non-conducting materials on the electrodes, as mentioned above. The RF plasma etching system consists of vacuum units, etching chamber, electrodes and electronic gadgets. The schematic of the RF powered plasma etching system is shown in Fig. 7.10 and the details of the vacuum system are covered in Chapter 10. Generally,

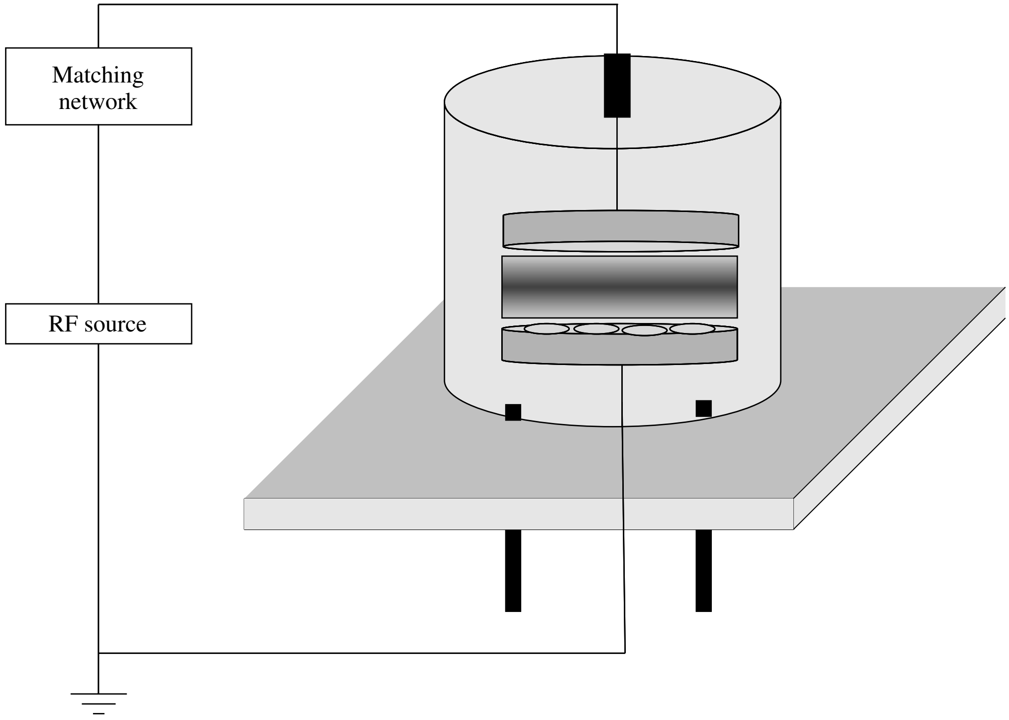


Fig. 7.10 Schematic of RF powered plasma etching system

plasma etching is carried out between 1 mtorr to 5 torr of pressure. Prior to the etching process, the chamber is evacuated to less than 10^{-6} torr to remove maximum possible air and moisture from it. This initial vacuum is called the **base vacuum**. Thereafter, regulated argon gas is introduced through the gas mass flow which controls the flow of gas in the chamber. The etching chamber pressure is controlled by a valve called the **throttle valve** which is fitted inside the vacuum system. The gas feed and the throttle valve are so adjusted that the plasma etching pressure remains constant.

Two separate plane electrodes are fixed in the centre of the etching chamber. The lower electrode is tied with a high negative potential to make it the cathode electrode where the wafers are kept, and the other electrode, i.e., the anode electrode is tied up to the ground. Generally, RF etching is done at 13.56 MHz frequency. This RF frequency is allotted to the **Industrial, Scientific and Medical (ISM)** applications by the Federal Communication Commission. To match the electrical impedance between the RF source and the plasma (cathode and anode capacitance), a matching network is connected between the cathode and the RF source for better electric power transfer as shown in Fig. 7.10. The film etching material is monitored by the mass spectrometer analyser, and the etching process is controlled electronically. It is essential to mention that these etching monitoring and precise controls are not possible in case of wet etching. Another noteworthy advantage of plasma (dry) etching with respect to wet etching is the sequential (etching of one film after another) etching of films without breaking the vacuum.

The previous sections described the etch rate of both the electrodes considering both electrodes to be of equal size. To increase the cathode etch rate significantly, the cathode electrode is made much smaller than the anode electrode. The voltage ratio between these two electrodes of unequal size can be expressed as

$$\frac{V_1}{V_2} = \left(\frac{A_2}{A_1}\right)^m$$

where the ratio V_1/V_2 is the ratio of the electrical potential of the cathode electrode with respect to the anode electrode, and the ratio A_2 /A_1 is the ratio of the anode to the cathode electrode size. The letter m has been experimentally calculated and it comes out between 1 and 2; however, its theoretical value is around 4. The potential distributions of the RF plasma of equal and unequal electrode configurations are shown in Fig. 7.11. To further increase the potential of the cathode electrode, the anode electrode is tied with the rest of the etching system and connected to the ground.

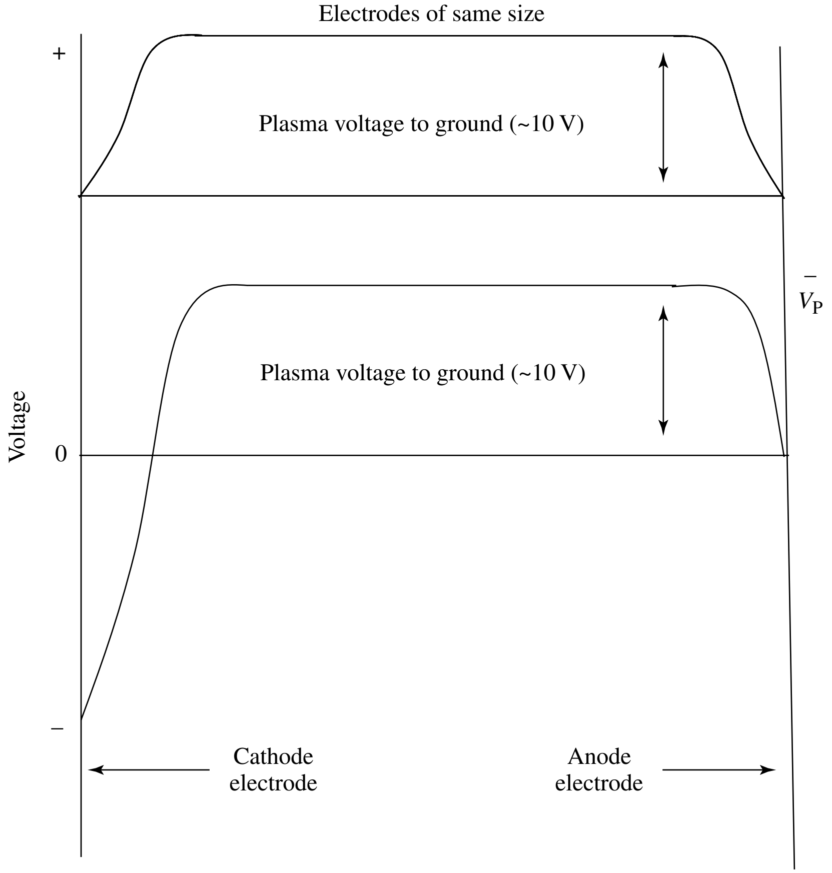


Fig. 7.11 RF voltage distribution when cathode electrode is smaller than the anode electrode

7.5 PLASMA ETCHING MECHANISMS AND ETCHING MODES

Plasma etching can be categorised into two modes according to their fundamental etching mechanism. These etching modes are called the **physical (sputter) etching mode** and the **plasma assisted etching mode**. Many a times, both physical and plasma assisted etching modes are used simultaneously to get better selectivity. This mode of etching is called the **reactive plasma enhanced etching mode** or the **ion-enhanced etching mode**.

7.5.1 Physical (Sputter) Etching Mode

In the physical etching by plasma mode, the etching is carried out at low range of pressure from 10^{-2} to 10^{-1} torr in the presence of argon gas. The pressure used during the etching of the film is called the **etching pressure**. The film containing the wafers is kept on the cathode electrode. The plasma is created in between the electrodes by applying dc or RF power in the range of kilovolts. The dc or RF power at which etching is carried out is called the **etching power**. In the plasma region, a large number of ions and electrons are present. These ions and electrons are attracted by the respective powered electrodes. The cathode electrode is made much smaller in size than the anode electrode. Argon ions present in the plasma region gain high kinetic energy while they are travelling towards the cathode electrode. These highly accelerated argon ions strike the film with great force and the atoms of the film get physically knocked out. For this reason, this process of film etching is called the **physical (sputter) etching technique**. In the physical etching process, the argon ions predominately strike perpendicular to the wafers; hence, the film is etched vertically (anisotropically) with an insignificant undercut, as shown in Fig. 7.12. In physical etching, the etch rate of the film is the same irrespective of the type of film material; therefore, this mode of etching suffers from very poor selectivity.

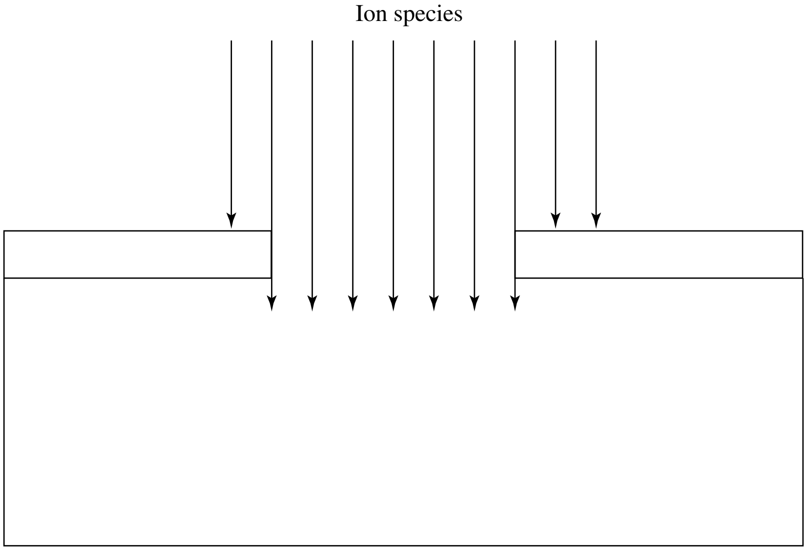


Fig. 7.12 Ion etching (physical)

7.5.2 Plasma Assisted Etching Mode

The plasma assisted etching technique can be further classified into two modes: the **Reactive Etching (RE) mode** where the film is etched purely by a gaseous form of the reactive species, and the **plasma enhanced etching mode** where etching is carried out by the reactive as well as the physical etching modes simultaneously.

7.5.3 Reactive Etching (RE) Mode or Reactive Ion Etching (RIE) Mode

The reactive etching (RE) mode is also called the **Reactive Ion Etching (RIE)** mode. In the RIE process, etching is generally carried out between 10^3 to 10^1 torr range of pressure. A regulated molecular gas is introduced through the mass flow controller in the range of 5 to 100 sccm. In the RIE mode, the molecular gas is dissociated into a reactive species in the plasma. This reactive species is called the **radical**. The most common molecular gases are halide based such as CF_4 molecules that dissociate (break) into free fluorine (F) radicals and neutral CF_3 molecules in the plasma. The fluorine radical species atom possesses seven electrons, which is one electron less than in normal fluorine atom. In the absence of one electron, the radical fluorine atom becomes highly reactive and reacts with the selective material (film) instantaneously. The selective etching of RIE is reasonably good for many materials. For example, the fluorine radical reacts with silicon and produces SiF_4, but it hardly reacts with the PR. The chemical reactions of CF_4 with silicon can be written as

$$e^- + CF_4 \rightarrow CF_3 + F + e^-$$
$$4F + Si \rightarrow SiF_4$$

The gaseous forms of CF_3 and SiF_4 are exhausted from the etching chamber by vacuum system. In the reactive ion etching (RIE) process, radicals move in all directions; hence, film etching takes place in both the vertical as well as the lateral direction, and that leads to undercut, as shown in Fig. 7.13.

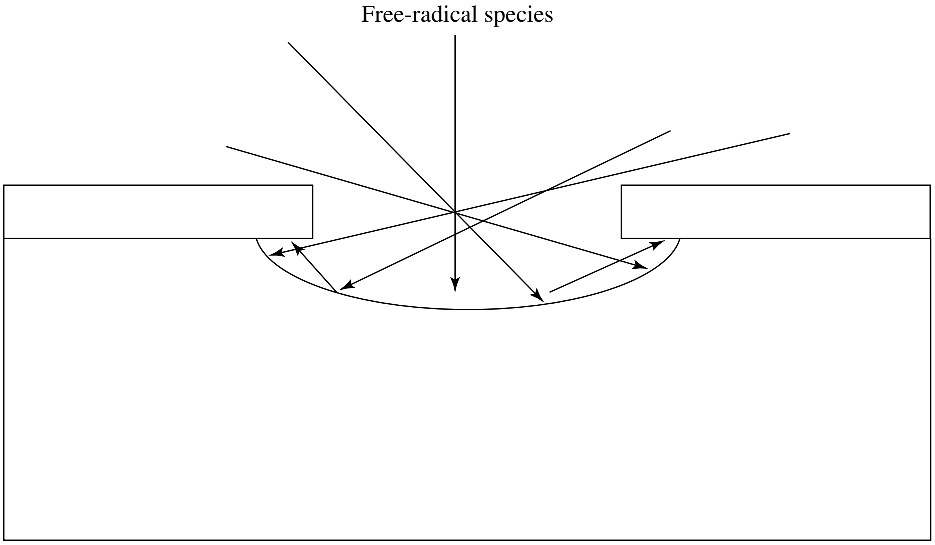


Fig. 7.13 Reactive Ion Etching (RIE)

7.5.4 Ion-Enhanced Etching Mode or Reactive Plasma Assisted Etching Mode

System configuration of the ion-enhanced etching is similar to the reactive ion etching system. In the ion-enhanced etching mode, the ions move vertically to the wafer and physically etch the film, and at the same time, the RIE etches with high selectivity. With the combination of these etching modes, one can optimise for less undercut with good selectivity. Furthermore, the etch rate of the film significantly increases. All three modes of film etching can be demonstrated separately through a single experiment, as shown in Fig. 7.14. From this graph, it is observed that the etch rate in the reactive etching (RIE) mode in the presence of XeF_2 molecular gas is much low; but it increases drastically, when ion-enhanced etching along with the RIE mode of etching is carried out. The etch rate again decreases significantly when molecular XeF_2 gas is stopped and etching is done purely using the physical etching mode. For these reasons, the ion-enhanced etching (reactive plasma assisted etching) mode is preferred in high density IC fabrication nowadays.

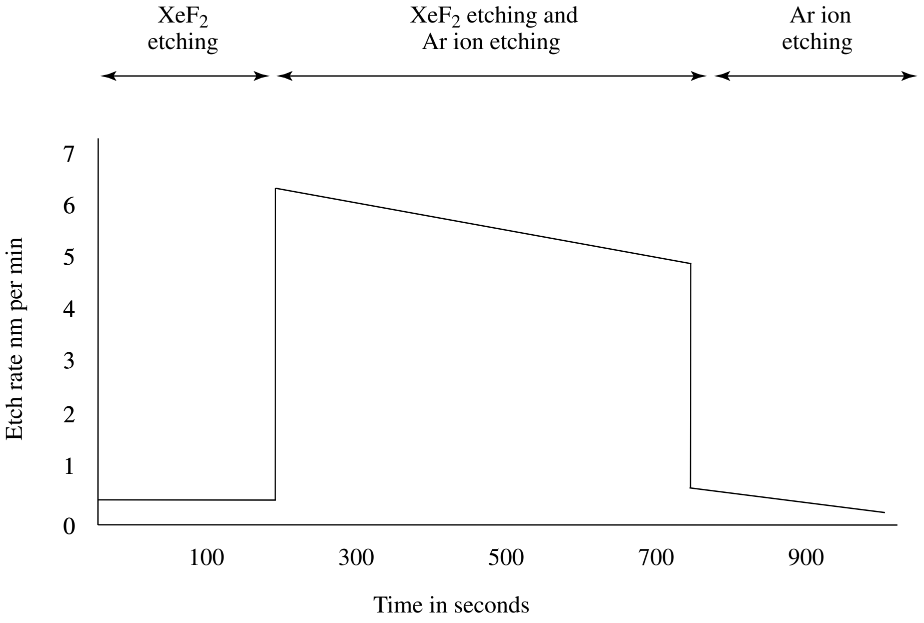


Fig. 7.14 Typical etch rate of silicon by RIE, RIE and physical etching

7.6 PLASMA ETCHING PARAMETERS

Plasma etching depends on many operating parameters such as gas composition, etching power, wafer spacing or loading, gas flow rate, and chamber pressure. Hence there are a large number of etch parameters that have to be optimised prior to material etching in the plasma etching process. In plasma etching, the wafer throughput is less as compared to wet etching. It also has poor etching uniformity and selectivity than the wet etching process.

Summary

Etching is one of the most critical process steps in IC fabrication. Generally, etching is used to remove the film from the selected areas and these etched areas are used for diffusion, contact, hole in dielectric, and trench isolation. On the other hand, the unetched film is used for the MOS gate, the capacitor, for electrical isolation, mask for diffusion, and other applications. The selectivity of etching should be high so that only a particular film is etched and other films are not etched. In addition, the film should not undercut below the sides of the PR pattern. Generally, overetch is given to take care of non-uniform film thickness on the wafer.

Film etching techniques can be divided into two categories: wet etching and plasma (dry) etching. Wet etching has good selectivity, but undercut is significant. Many a times, the pH value of the etchant is maintained by adding an appropriate chemical to maintain a constant etch rate of the film. This type of etchant is called the buffer etchant. Plasma (dry) etching is done in vacuum by ions or reactive atoms under high electric field. The film is etched anisotropically (vertically) with insignificant undercut, but it has poor selectivity. The selectivity is increased by choosing a proper precursor and combination of the etching modes. For this reason, dry etching is preferred for today's high-density VLSI and ULSI applications.

Plasma etching can be categorised into two types of etching according to their fundamental etching mechanism. These etching modes are called the physical (sputter) etching mode where the film atoms are physically knocked out by argon ions, but it has poor selectivity. In the plasma assisted etching mode, the film material is removed selectively by the reactive ions produced in the plasma. Many a times, both physical and plasma assisted etching modes are used simultaneously to get better selectivity and etch rate.

References

- J D Plummer, M Deal and P B Griffin; *Silicon Fundamental Technology: Fundamentals, Practice and Modeling*, Prentice Hall, 2000
- R C Jaeger; *Introduction to Microelectronic Fabrication: Volume 5 of Modular Series on Solid State Devices*, Prentice Hall, Second Edition, 2001
- S M Sze; *VLSI Technology*, Second Edition, McGraw-Hill, 1988
- S K Gandhi; *VLSI Fabrication Principles*, Second Edition, Wiley, 1994
- D Nagchoudhuri; *Principles of Microelectronic Technology*, Wheeler, 1998
- S A Campbell; *The Science and Engineering of Microelectronic Fabrication*, Oxford University Press, 1996

Multiple-Choice Questions

7.1 Does the buffer etchant solution
(a) maintain the PH value (b) lower the PH value
(c) has nothing to do with the PH value?

7.2 The etch rate is linear to which parameter?
(a) Temperature of the etchant (b) concentration of the etchant
(c) agitation of the etchant

7.3 The wet etching is
(a) isotropic (b) anisotropic.

7.4 Dry etching is
(a) isotropic (b) anisotropic (c) isotropic and anisotropic both

7.5 The silicon wafer is
(a) hydrophobic (b) hydrophilic

7.6 Silicon dioxide is
(a) hydrophobic (b) hydrophilic

7.7 Etching of the boron doped silicon is
(a) faster (b) slower (c) neither faster nor slower

7.8 Generally, the chamber pressure kept during dry etching is
(a) 10^{-2} torr (b) 10^{-4} torr (c) 10^{-6} torr

7.9 Plasma potential in dry etching is around
(a) 10 volts (b) 50 volts (c) 1000 volts

7.10 Which RF frequency is used for the plasma etching system?
(a) 1.2 MHz (b) 13.56 MHz (c) 15.56 MHz

7.11 Which gas is used for physical dry etching?
(a) Argon (b) Neon (c) Oxygen

7.12 Which gas can be used for reactive ion etching?
(a) CF_4 (b) Hydrogen (c) Nitrogen

Descriptive Problems

7.1 Why is RIE faster than anisotropic etching?

7.2 What are the basic differences between plasma etching and wet etching?

7.3 Explain RCA wafer cleaning procedure with reasons.

7.4 The etch rate is faster in bigger geometry patterns than small geometry patterns. Why?

7.5 What is the difference between dc and RF plasma etching?

CHAPTER

8 Diffusion

8.1 INTRODUCTION

The electrical conductivity of a semiconductor changes significantly when it is doped with (introduced to) selective elements. This property of a semiconductor is exploited to realise many semiconductor devices. The element that is doped in the semiconductor and alters the semiconductor conductivity is called the **dopant** and the process by which the element is introduced into the semiconductor is called **diffusion.** Mostly, the third and the fourth group elements of the periodic table, namely, boron, phosphorus and arsenic are diffused (doped) into the silicon semiconductor for MOS transistor fabrication. Henceforth, the intentionally doped element (P, As, B, etc.) that is used to change the silicon conductivity will be referred to as the **dopant,** and the unwanted elements present in the silicon wafer prior to MOS fabrication and/or introduced during MOS fabrication, for example, Fe, Au, etc., are referred to as **impurities.**

To get the desired MOS transistor characteristic, the type of dopant (element) and its dose (concentration) play an important role. For instance, a high dose of the dopant is required to make the source, the drain, the gate, the interconnection, the capacitor, and the contacts; whereas, a low dose is required for the V_T adjustment, the channel stop, the well formation, etc. The objective of doping has been described in the process flow in Chapter 3.

Generally, to fabricate the MOS, two types of dopant diffusion techniques are used, namely, thermal diffusion technique and ion-implantation technique. The mechanisms of these two diffusion techniques are quite different; therefore, these diffusion techniques are described in two different chapters. In this chapter, the thermal diffusion technique is described wherein thermal diffusion models, dopant diffusion parameters, and diffused layer evaluation are covered; and in the subsequent chapter, the ion-implantation technique is described.

8.2 DIFFUSION EQUIPMENT AND PROCESS

There are four thermal diffusion techniques used for doping the silicon wafer. These are doped-solid-oxide source diffusion technique, gas-source diffusion technique, planar solid source diffusion technique, and liquid source diffusion technique. In the doped-solid-oxide diffusion technique, the silicon wafer is coated with the liquid form of doped-solid-oxide sources. The process of doped-solid-oxide coating over the silicon wafer is same as PR coating, as described in Chapter 6. Then, the silicon wafer is heated at an elevated temperature. The dopant comes out from the doped-solid-oxide film and diffuses into the silicon wafer. This technique is no longer used in high-density IC fabrication due to many practical limitations. In the gas-source diffusion technique, the silicon wafer is doped with the dopant gas at an elevated temperature. In the planar solid-diffusion technique, the planar solid source, which is in the shape of a wafer is placed beside the silicon wafer and heated at an elevated temperature. The dopant vapours come out of the planar solid source and diffuse into the silicon wafer. In the liquid-source diffusion technique, the vapour from the dopant liquid is transported by a carrier gas into an elevated temperature furnace where the silicon wafer is kept. At the elevated temperature, the dopant diffuses into the silicon wafer.

There are many types of dopant sources that are used for MOS transistor fabrication. For *N*-type doping, phosphorus source is preferred, where the phosphorus source may be in the form of liquid phosphorus oxy-chloride ($POCl_3$), gaseous phosphine (PH_3), or phosphorus nitride solid planar; and, for arsenic doping, only silane gas source is used. For *P*-type doping, liquid boron tri-chloride (BBr_3), gaseous diborane (B_2H_6), or planar source boron nitride sources are commonly used. Generally, phosphorus and boron dopants are used for deep junctions such as the well, the source, and the drain; whereas, arsenic is used in place of phosphorus for shallow junctions, channel stops and V_T adjustment.

The construction of the diffusion furnace is almost the same as that of the oxidation furnace, as depicted in Fig. 8.1. The diffusion furnace is brought to a high temperature in the range of 900ºC to 1100ºC and then, wafers are loaded inside the furnace. Generally, boron is diffused at 960ºC and phosphorus is diffused at 1050ºC. Loading of the wafer is done in the same way as in the case of thermal oxidation (see Chapter 4). At the time of dopant diffusion, 50 cc lit/min of oxygen and 1 lit/min of N_2 are passed in the furnace. After the completion of the diffusion process, O_2 gas is stopped and the wafer is heated in

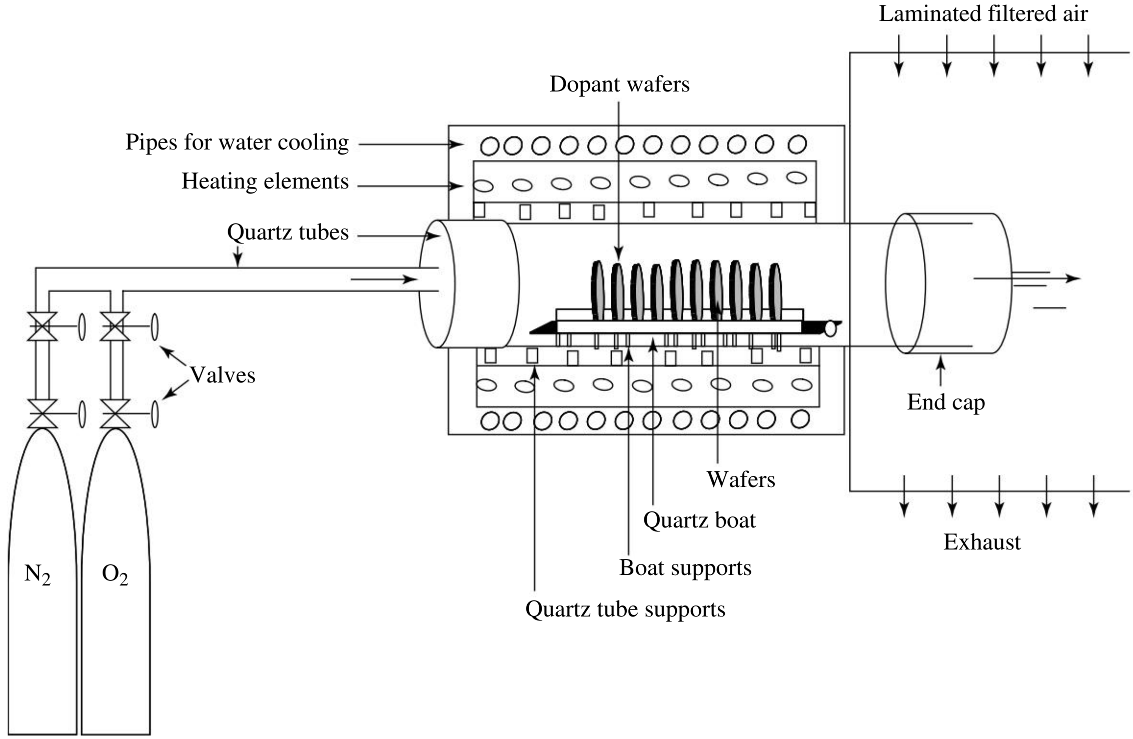


Fig. 8.1 Thermal diffusion furnace

N_2 gas for annealing. It is important to mention that at the time of diffusion, a thin oxide layer is formed on the wafer due to the presence of oxygen that comes from different sources as an impurity. This thin oxide is heavily reached with the dopant and cannot be conveniently etched by the wet chemical. To reduce (dilute) the dopant concentration in the thin oxide, the silicon wafer is oxidised so that the silicon dioxide volume increases and the dopant per unit volume reduces; then, it becomes easier to etch by the wet etching process. This process is called **dopant dilution**.

8.3 DIFFUSION MODELS

The electrical characteristics of the MOS are mainly dictated by the dopant concentration inside the silicon in terms of junction depth and junction profile. The vertical depth of the dopant inside the silicon is called the **junction depth** and denoted by either J_n or $x_{i;}$ and, the dopant distribution inside the silicon is called the **doping profile**. Slight change in the doping distribution and the junction depth from the required specifications may change the MOS electrical characteristics drastically. Therefore, the diffusion process parameters (recipe) are optimised carefully prior to IC processing. To save time and cost, the junction depth and the doping profiles are optimized through unit process (recipe) simulation.

The thermal diffusion can be explained by two diffusion models, the first being the **Fick's diffusion model**, which is based on classical mechanics and, the second being the **atomic diffusion model,** which is based on the atomic interaction between the ionised dopant atoms and the electrically charged point defects present in the silicon wafer.

8.3.1 Fick's Diffusion Model

Fick's diffusion model is based on two Fick's diffusion laws which were initially used to predict the heat diffusion in the metal. The Fick's diffusion model is applied to dopant diffusion and gives an almost accurate result.

Fick's First Law

According to Fick's first law, **the dopant diffusion (flux) is proportional to the dopant concentration gradient**, and can be expressed as

$$F = -D\frac{\partial C}{\partial x} \tag{1}$$

where F is the flux (atoms cm^{-2} sec^{-1}), C is the dopant concentration per unit volume (atoms cm^{-3}), $\frac{\partial C}{\partial x}$ is the dopant concentration gradient, and D represents the diffusion constant (cm^2 sec^{-1}). The diffusion constant D is also referred to as the **diffusivity**. The negative sign in Eq. (1) indicates that the dopant is diffusing from a higher concentration to a lower concentration.

Fick's Second Law

The second law of Fick is based on the continuity theory of diffusion, which says, **the rate of change of dopant density in a volume element is equal to the divergence of the dopant flux density**. Assuming that there is no creation or annihilation of the dopant atoms in the silicon, the diffusion continuity equation for one dimensional case can be written as

$$\frac{\Delta C}{\Delta t} = \frac{\Delta F}{\Delta x} = \frac{F_{\text{in}} - F_{\text{out}}}{\Delta x} \tag{2}$$

or,

$$\frac{\delta C}{\delta t} = \frac{\delta}{\delta x}\left[D\frac{\delta C}{\delta x}\right] \tag{3}$$

Equation (3) can be written as

$$\frac{\delta C}{\delta t} = \left[D \frac{\delta^2 C}{\delta x^2} \right] \tag{4}$$

where, D is taken as the constant and Eq. (4) is referred to as Fick's second law of diffusion.

To fabricate the MOS transistor, the diffusion process is carried out in two steps. In the first step, a controlled amount of the dopant is introduced (diffused) at the surface of the silicon wafer. This process of diffusion is called **pre-deposition.** Pre-deposition diffusion process is carried out in the undiluted dopant gaseous source at an elevated temperature for a stipulated period of time. In the second step, a controlled amount of the dopant introduced by the pre-deposited dopant is pushed inside the wafer to obtain the required junction depth and dopant profile. This part of the process is called **drive-in.** The drive-in process is done at an elevated temperature for a stipulated period of time in the presence of nitrogen gas.

8.3.2 Pre-deposition

Let us consider that the diffusion is carried out with no depletion of the gaseous dopant source at the bare silicon wafer and then, the analytical solution of Fick's law under these boundary conditions can be obtained as follows:

$$C(0, t) = C_s$$

where C is the concentration of the dopant, C_s is the surface (at $x = 0$) concentration of the dopant and it is independent of time.

The second boundary condition can be expressed as

$$C(\infty, t) = 0$$

This boundary condition states that at a large distance from the wafer surface ($x = 0$), there are no dopant atoms present at any point of time in the wafer. In the real situation, the silicon wafers are uniformly doped with a low concentration of the dopant during the crystal growth, but in deriving Fick's model, it is assumed that there is no dopant present in the silicon wafer prior to diffusion.

The analytical solution of Fick's diffusion law (Eq. 4) can be written as

$$C(x, t) = C_s \, erfc\left[\frac{x}{2\sqrt{Dt}} \right] \tag{5}$$

where C_s is the dopant surface concentration and it is considered to be a constant; and, erfc is the complementary error function. The normalised complementary error function of the dopant distribution profile is shown in Fig. 8.2, where, $\sqrt{Dt}$ is referred to as the diffusion length or the junction depth and generally denoted by *Jn* or x_i, where, the *x*-axis and the *y*-axis represent the junction depth and the normalised dopant concentration (dopant profile) in the log scale respectively.

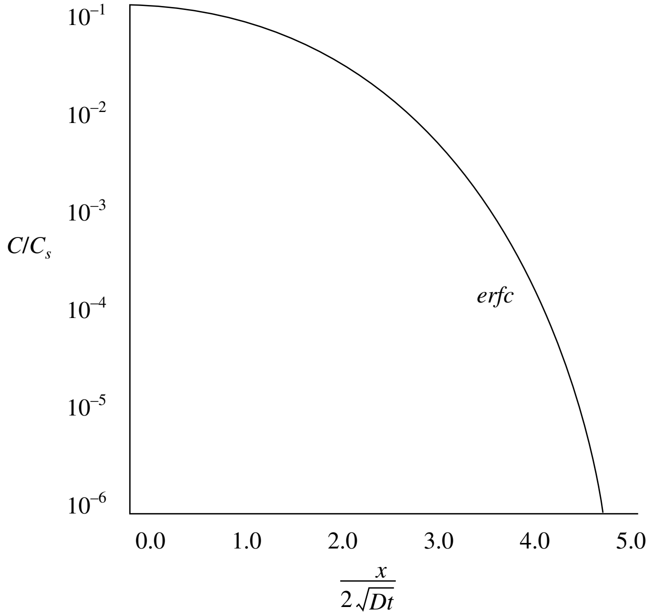


Fig. 8.2 Normalised complementary error function: Pre-deposition case

The total dopant quantity Q (dose) introduced by the pre-deposition process can be approximated as

$$Q = \int_0^t C(x, t)dx$$

On substituting the above into Eq. (5), we can get the approximate value of the dopant quantity as

$$Q = \frac{2}{\sqrt{\pi}} C_s \sqrt{Dt} \approx 1.1 C_s \sqrt{Dt} \tag{6}$$

The above equation can be approximated as a triangle, as shown in Fig. 8.3; where, the diffusion profile is linearly plotted and C_s and ($\sqrt{Dt}$) are the height and the base respectively of the triangle; hence, Q can be approximately estimated as

$$Q = (C_s)\ x\ (\sqrt{Dt}) \tag{7}$$

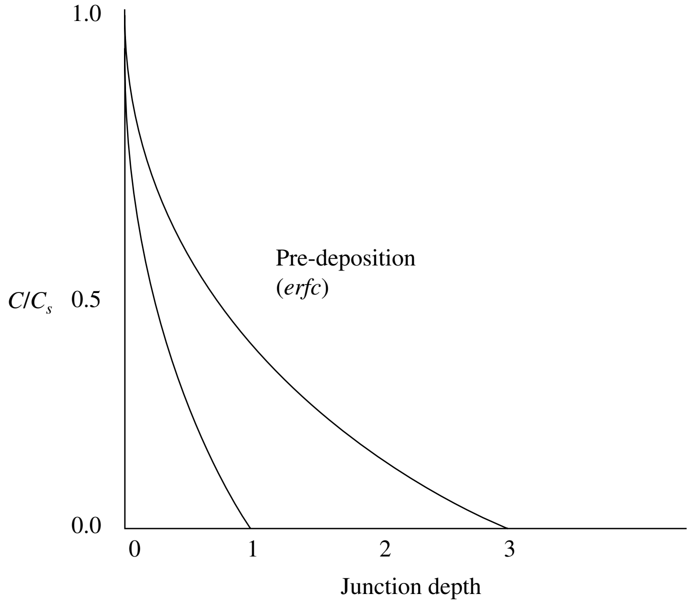


Fig. 8.3 Triangle represent ion of doping

8.3.3 Drive-in

The main aim of the drive-in process is the redistribution of the dopant (pre-deposited dopant) to obtain the required doping profile and junction depth. Let us consider that a total quantity of dopant atoms (Q) is introduced in the pre-deposition process at the surface of the wafer and then, the wafer is subjected to an elevated temperature. The Fick's second-law boundary condition states that **the amount of the dopant deep inside the silicon at any point of time is zero**; this can be mathematically written as

$$C(\infty, t) = 0$$

This boundary condition is the same as the pre-deposition boundary condition, and can be expressed as

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

Here, Q is the total quantity of the dopant present on the wafer surface. Then, the analytical solution of Eq. (4) becomes

$$C(x,t) = \frac{Q}{\sqrt{\pi Dt}}\exp\left(-\frac{x^2}{4Dt}\right) \tag{8}$$

or,

$$C(x, t) = C(0, t)\exp\left(-\frac{x^2}{4Dt}\right) \tag{9}$$

The above equation is a Gaussian distribution in nature, as shown in Fig. 8.4. One can observe from the figure that the peak concentration C (0, t) decreases with l/√t and it falls to 1/e (dopant concentration) at a distance $x = 2\sqrt{Dt}$ (i.e. the diffusion length) from the wafer surface. The normalized diffusion profiles of error-function and Gaussian distribution have similar nature and are shown in Fig. 8.5. The dopant distribution after pre-deposition and drive-in is shown in Fig. 8.5 for comparison.

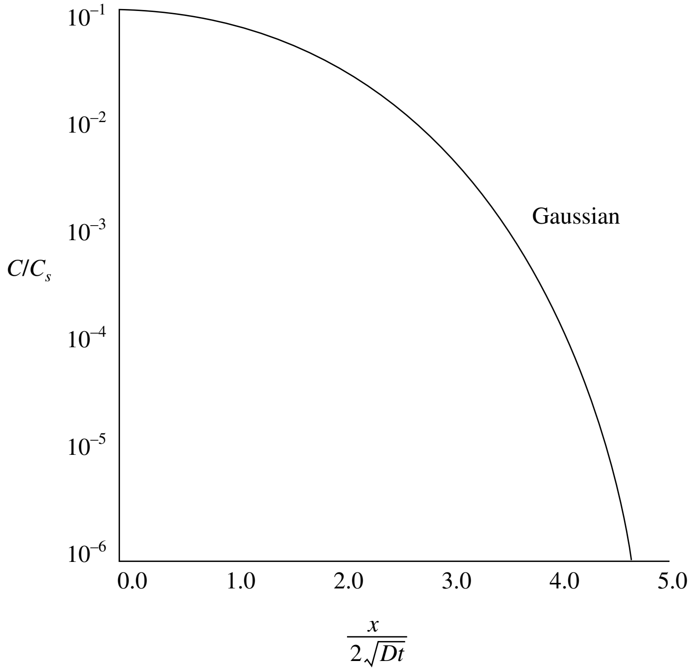


Fig. 8.4 Normalised Gaussian distribution: Drive-in case

Example 8.1

Boron pre-deposition is done at 950°C on the silicon wafer for 90 minutes, where the boron concentration in the silicon is 10^{19}/cm^3. The diffusion constant of the boron in the silicon is 2×10^{-14} cm^2/s. Find out the junction depth, the total quantity of boron in the silicon wafer, and plot the graph of the boron profile in the silicon.

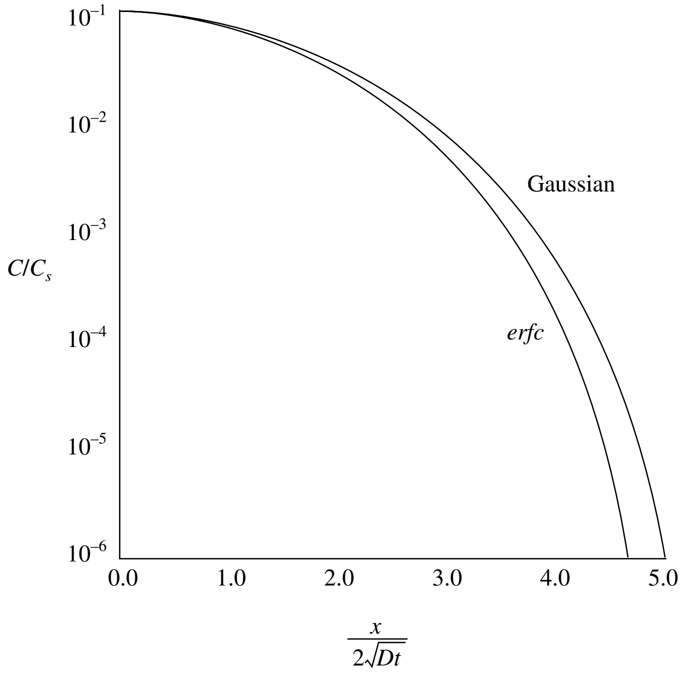


Fig. 8.5 Normalised pre-deposition and drive-in

Answer

Junction depth can be approximated as

$$\sqrt{Dt} = \sqrt{2 \times 10^{-14} \times 9000} = 1.34 \times 10^{-5}$$

From Eq. 8.6, the total quantity of boron in atoms/cm^2 is

$$\begin{aligned} Q(t) &\approx 1.13 C_s \sqrt{Dt} \\ &= 1.13 \times 10^{19} \times 1.34 \times 10^{-5} \\ &= 1.51 \times 10^{14} \text{ atoms/cm}^2 \end{aligned}$$

The concentration of boron at the surface of the silicon wafer is 10^{19}/cm^3 and, at the junction, it is zero. Hence, one can get the approximate boron profile by a straight line joining 10^{19}/cm^3 as shown in Fig. 8.3.

8.3.4 Atomic Diffusion Model

In the atomic diffusion mechanism, the diffusivity is calculated in terms of the interactions between the ionised dopant atoms and the charged point defects present in the wafer.

At an elevated temperature, some of the silicon atoms acquire sufficient kinetic energy and leave their lattice sites, and create neutral point defects in the silicon lattice (see Chapter 2). These neutral point defects become electrically active when they capture charge or charges. For example, if a neutral vacancy point defect captures one electron then it becomes a singly negatively charged point defect, as indicated below:

$$V + e \Leftrightarrow V^-$$

Similarly, if a neutral interstitial point defect acquires an electron it becomes negatively charged as

$$I + e \Leftrightarrow I^-$$

These point defects can even capture two or more electrons. The total number of negative point defects present in the silicon can be denoted by the neutral vacancy V^0, the acceptor charge V^-, and the doubly charged acceptor V^{2-}. The total number of positive point defects in the silicon can be denoted by the vacancy donor V^0, singly charged donor V^+ and doubly charged donor V^{2+}. It is found experimentally that the doubly charged point defects are less in number; hence, their interactions with the ionised dopant atoms are almost negligible in the diffusion process.

It is experimentally found that the diffusivity depends on the concentration of the point defects and the dopant concentration. To define the dopant-concentration level, the intrinsic carrier concentration is taken as a reference. If the dopant concentration n is less than the intrinsic carrier concentration n_i at the diffusion temperature then the dopant concentration is called **low dopant concentration** and the diffusion process is referred to as **intrinsic diffusivity.** If the dopant concentration n, on the other hand, is more than the intrinsic carrier concentration n_i at the diffusion temperature then the dopant concentration is called **high dopant concentration** and the diffusion process is referred to as **extrinsic diffusivity**.

The relationship of diffusivities between the acceptor atoms and the intrinsic carrier concentration at a particular temperature is expressed as

$$\frac{D}{D_i} = \frac{n}{n_i} \tag{10}$$

where D, D_i, stand for extrinsic and intrinsic diffusivities respectively; n and n_i are the dopant concentration of the acceptor atoms, and the intrinsic carrier concentration respectively, as shown in Fig. 8.6.

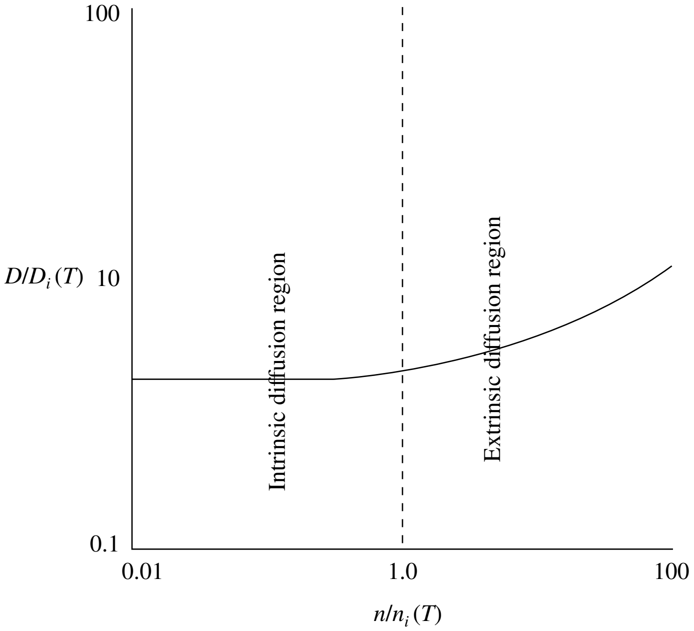


Fig. 8.6 Intrinsic carrier concentrations with temperature

The phosphorus ions interact with the negatively charged vacancy and, the phosphorus dopant diffusivity can be written as

$$D = D^0 + D^-\left(\frac{n}{n_i}\right) + D^{2-}\left(\frac{n}{n_i}\right)^2 \tag{11}$$

where D^0 is the neutral diffusivity and D^- and D^{2-} are the intrinsic dopant diffusivity associated with the singly negative and doubly negative point defects respectively.

The boron atoms interact with the positively charged vacancy point defects and, for the boron donor dopant, the diffusivity can be written as

$$D = D^0 + D^+\left(\frac{p}{n_i}\right) + D^{2+}\left(\frac{p}{n_i}\right)^2 \tag{12}$$

where D^0 is the neutral diffusivity and D^+ and D^{2+} are the intrinsic dopant diffusivity associated with the singly positive and doubly positive point defects respectively.

Equations (11) and (12) given above can be written in the generalised form as

$$D = D^0 + \sum_{r=1}^{m}(D^{-r})\left[\frac{n}{n_i}\right]^r + \sum_{r=1}^{m}(D^{+r})\left[\frac{n_i}{n}\right]^r \tag{13}$$

In addition, the diffusivity varies exponentially with respect to the diffusion temperature and, it is expressed as

$$D = D_0 \exp\left(\frac{-E_a}{kT}\right) \tag{14}$$

The generation of the intrinsic carrier concentration, electrons, or holes, at diffusion temperature T is expressed as

$$n = p = n_i = 3.9 \times 10^{16} T^{3/2} \exp\left(\frac{-E_G}{2kT}\right) \text{cm}^{-3} \tag{15}$$

In the equations (14) and (15) given above, the term D_0 is the diffusion constant (in units of cm^2/s), E_a is the activation energy (in eV), K is the Boltzmann constant, n is the electron concentration (in cm^{-3}), p is the hole concentration (in cm^{-3}), n_i is the intrinsic carrier concentration (in cm^{-3}), E_G is the band gap (in eV), and T is the diffusion process temperature in K.

8.4 MODIFICATION OF FICK'S LAW

The Fick's diffusion model is simple, fast, and easy to calculate. Unfortunately, Fick's model fails to predict many diffusion cases, especially those with high dopant concentration, electrical field, and a few IC process-related cases. To modify Fick's law based on these failure diffusion cases, the results are found experimentally and are then incorporated in the Fick's model.

8.4.1 Electrical-Field-Related Diffusivity

Electric-field-related diffusion on account of dopant diffusion of the electric field and atomic interaction is explained below.

8.4.2 Dopant Diffusion on Account of Electric Field

It has been observed experimentally that dopant diffusivity enhances due to the electric field. This phenomenon becomes prominent, as the dopant concentration increases. At diffusion temperature, the dopant atoms dissociate into ionised atoms and free electrons. Electron mobility is much faster in silicon than that of the ionised atoms; hence, the electrons go deep into the wafer leaving behind the ionised atoms. This separation of ions and electrons builds an electrical field (built-in potential) that pulls the ionised ions inside the wafer. The modified Fick's model due to electrical field can be written as

$$F_{\text{total}} = F + F_{enh} = hD\,\frac{\partial_n}{\partial_x} \tag{16}$$

where F_{enh} is the diffusivity enhancement flux due to the built-in electric field, n is the net concentration of the dopant, and h is the diffusion enhancement factor, and is represented by

$$h = \frac{n}{\sqrt{(n^2 + 4n_i^2)}} \tag{17}$$

It has been found that the enhancement factor can have a maximum value of 2.

8.4.3 Diffusion on Account of Atomic Interaction

Boron Diffusion

It has been established from experiments that boron atoms interact with the singly positively charged vacancy point defects V^+, and that the diffusion of boron is linear in the initial phase of diffusion and then falls rapidly at the end of the boron profile. A typical boron profile is shown in Fig. 8.7. Many a times, this property of boron diffusion is exploited for making shallow-junction transistor applications.

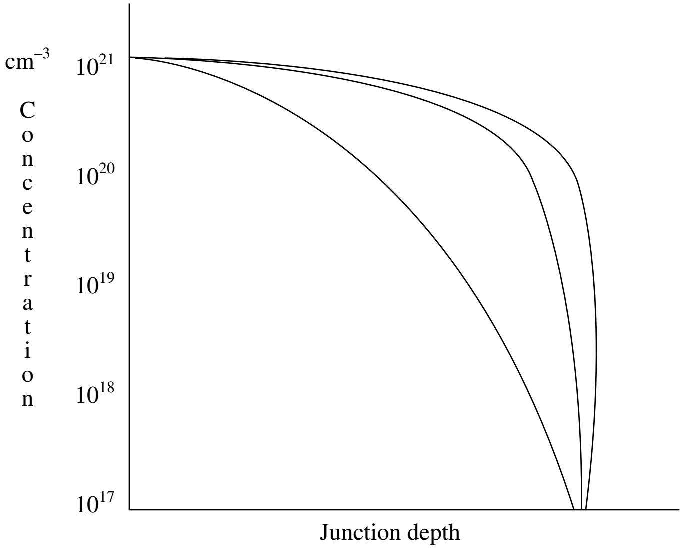


Fig. 8.7 Atomic model of boron-diffusion profile

Phosphorus Diffusion

Phosphorus diffusion has a unique characteristic, as shown in Fig. 8.8. It has been found that phosphorus at low (intrinsic) concentration (Fig. 8.8a), diffuses due to the interaction of the phosphorus ions with the doubly negatively charged vacancy V^{2-}; and that forms the P^+V^{2-} pairs. The diffused phosphorus profile at this low concentration fits well with the *erfc* function (Fick's law). But when the concentration of phosphorus approaches 10^{20} per cm^{-3}, the phosphorus diffusivity decreases fast and reaches a minimum point as shown in Fig. 8.8b. This minimum point of diffusion is called **kink.** At this point, the phosphorus-vacancy pair P^+V^{2-} starts dissociating into P^+ and V^-. This dissociation of P^+ and V^- releases electrons and a large number of singly charged acceptors V^- as shown in Fig. 8.8c. This manifests rapid phosphorus diffusion that goes deep into the silicon wafer and leads to **tail** formation as shown in Fig. 8.8d. This property of phosphorus diffusivity is exploited for deep n-wells for CMOS fabrication.

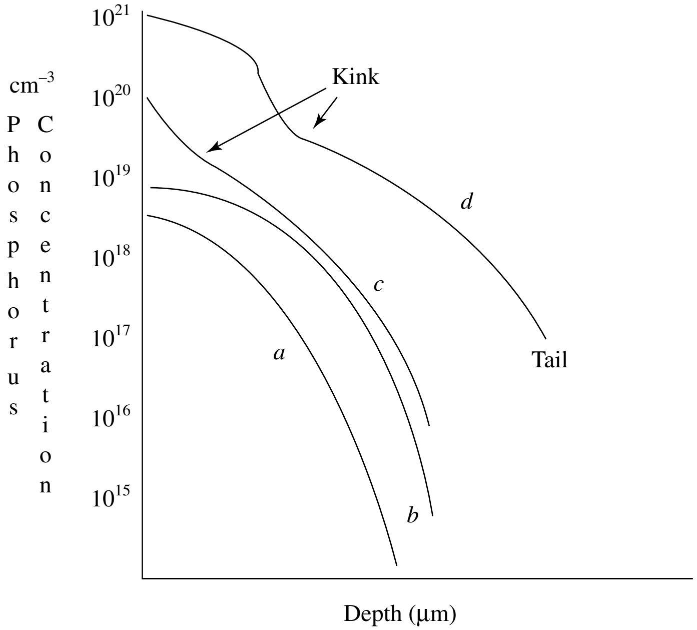


Fig. 8.8 Atomic model of boron-diffusion profile

8.4.4 Emitter-push Effect

The emitter-push effect is manifested because of the abnormal behaviour of phosphorus diffusion at the tail region. In the *n-p-n* transistor, a lightly doped boron base is made under the heavily doped phosphorus emitter. It is found that the boron atoms of the base of the transistor get influenced by the phosphorus dissociation that produces the V^- defects. This causes an increase in the diffusivity of boron and it moves around 0.6 micrometre into the emitter region. This effect is called the **emitter-push effect** and is illustrated in Fig. 8.9.

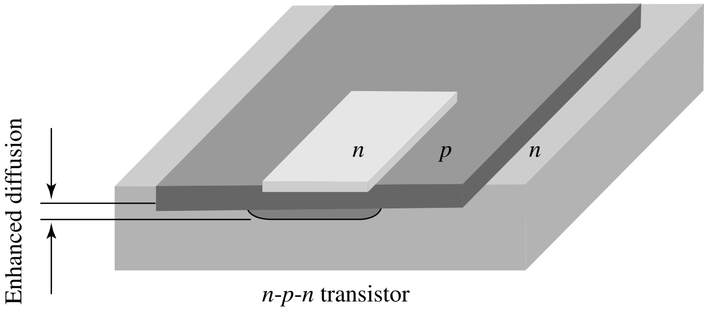


Fig. 8.9 Emitter push

8.4.5 Concentration Dependent Diffusivity

It has been mentioned in the previous section that doping profile depends on the dopant concentrations significantly. The concentration dependent diffusivity D can be expressed as

$$D = D_s \left[\frac{C}{C_s} \right]^{\gamma} \tag{18}$$

where C_s is the dopant concentration at the wafer surface (where the maximum doping concentration is present), C is the total dopant concentration in the silicon wafer, D_s is the diffusion coefficient at the wafer surface, and γ is a concentration dependent parameter. These computations of concentration dependent diffusion are based on the constant-source diffusion (C_s) and the constant diffusivity (D_s). To demonstrate the dopant concentration diffusion, typical discrete values of γ are taken in the form of positive integers, and are shown in Fig. 8.10. To compare the diffusivity with constant diffusivity (Fick's model), the value of γ is taken as zero and is shown in Fig 8.10. The typical plots for a higher constant-surface concentration diffusion profile ($\gamma \geq 0$) are found to be steeper (box type), as shown in this figure.

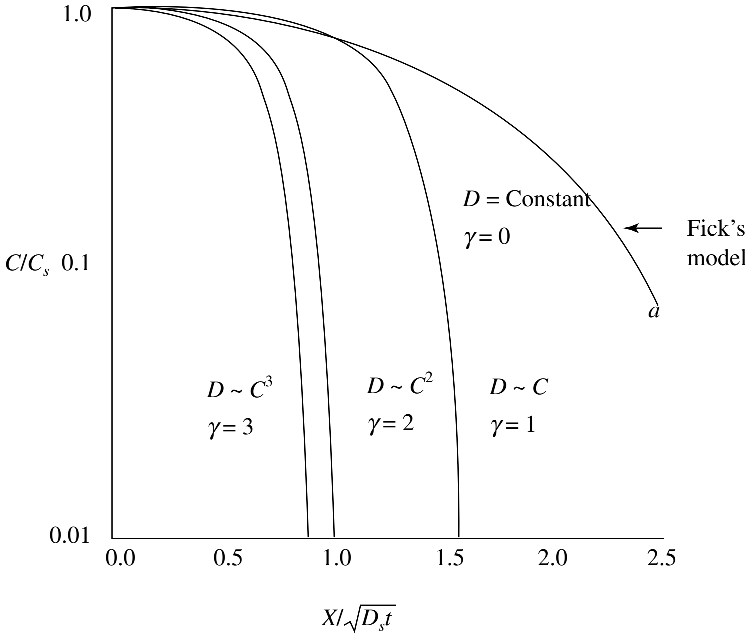


Fig. 8.10 Constant diffusivity when γ value is zero

8.4.6 Diffusion on Account of Other Related Processes

The dopant profile changes as a function of the high-temperature processes, and that leads to serious issues, especially in the case of high-density MOS transistors used in recent VLSI and ULSI applications. The changes to the doping profile on account of high-temperature processes are mentioned below.

8.4.7 Oxidation Related Diffusion

In Chapter 4 (oxidation), it has been explained that the dopant redistribution at the interface of silicon and silicon dioxide is due to high temperature oxidation. The redistribution of dopants at the interface is called the **segregation coefficient** and is generally denoted by *k* or *m*. The dopant redistribution takes place because of the different solubilities of the dopants in different materials. Generally, dopant segregation coefficient is defined as

$$K = \frac{\text{Equilibrium concentration of dopant in Si}}{\text{Equilibrium concentration of dopant in SiO}_2}$$

The boron atoms are more soluble in the oxide than in the silicon; therefore, boron atoms diffuse into the oxide and they deplete at the oxide/silicon interface. On the other hand, phosphorus atoms are rejected by the oxide; hence, they are pushed below the oxide/silicon interface. This segregation phenomenon alters the dopant profile at the interface significantly, where a MOS transistor channel is generally formed. In addition, boron in oxidation may change the oxide quality. The change in the dopant profile has a direct impact on the electrical parameters of the MOS transistor, especially the threshold voltage. The loss or gain of the dopant due to segregation has to be adjusted by the threshold adjustment process, as discussed in Chapter 3 and Chapter 9.

8.4.8 Interfacial Dopant Segregation Related Diffusivity

It has been found that the dopant piles up at the oxide/silicon interface when the junction is shallow. This phenomenon is different from the phenomenon of segregation. The piled dopant is lost when the oxide/silicon interface is etched out. This is a serious matter when the dopant dose is low for shallow junction transistors. Hence, the loss of the dopant is adjusted by the threshold adjustment process.

8.5 OXIDATION EFFECTS ON DIFFUSION

8.5.1 Oxidation Enhanced Diffusion (OED) and Enhanced or Retarded Diffusion

It has been found that during oxidation, the diffusivity of P, B and As increases; whereas, the diffusivity of antimony decreases. This phenomenon is called **Oxidation Enhanced Diffusion (OED)**. The enhancement and retardation of the dopant diffusion occurs because of the generation of point defects during the oxidation process. This diffusion enhancement and retardation cannot be predicted by Fick’s model; hence, it is modified in the case of OED. Sometimes, this effect is exploited in cases such as for deep-well formation in the CMOS fabrication process.

8.5.2 Measurements of Diffused Layer

In this section, the measurements of junction depth, sheet resistivity, and diffusion profile are described.

Junction-Depth Measurement

The dopant junction depth (J_n or x_i) dictates the electrical parameters of a MOS transistor. Hence, it is essential to know the junction-depth measurement. The junction depth is measured in many ways. The commonly used ways are the bevel, the groove, and the Secondary-Ion-Mass-Spectroscopy (SIMS) techniques. Out of these three techniques, the bevel and the groove techniques are simple, but they are less accurate than the secondary-ion-mass-spectroscopy technique. Generally, the bevel and the groove measurements are used for a deep junction; whereas the secondary-ion-mass-spectroscopy is suitable for shallow junction-depth measurement used in the VLSI and ULSI applications.

In the bevel technique, one end of the diffused wafer is bevelled at around 1° angle and thereafter, the wafer is stained by a chemical. A simple chemical stainer made of 100 cc of HF added with a few drops of nitric acid is used. During staining, the wafer is kept under a strong illuminated light. After chemical staining, different types of silicon show different kinds of colours. For instance, the *N*-doped silicon becomes darker than the *P*-doped silicon. Then, a half-silvered mirror is placed on the bevelled surface and monochromatic light is illuminated from the top of the mirror, as shown in Fig. 8.11. Part of the monochromatic light is reflected from the silver mirror and the remaining part of the light is reflected from the bevelled silicon surface. These two reflected lights interfere with each other and interference fringes are produced. The mirror is adjusted in such a way that the fringes become perpendicular to the length of the levelled surface. The *N*-doped and *P*-doped regions (junction depth) are measured in terms of the fringe width; more precisely in terms of the wavelength of light. To increase the accuracy of measurement, the silicon is bevelled at an angle of less than 1°.

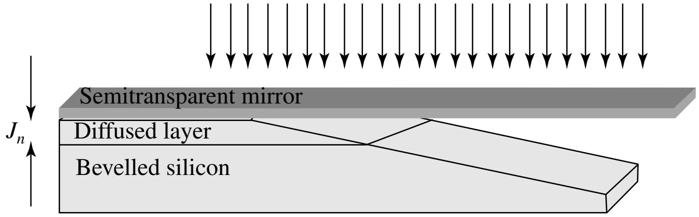


Fig. 8.11 Junction-depth measurement by bevel technique

A shallow junction with a high degree of accuracy of up to 0.1 μm can be measured by the groove technique. In this technique, a groove is made in the dopant-diffused silicon wafer and then it is delineated by the staining chemicals in strong light, as explained in the bevelled staining technique. To measure the junction depth, the dimensions of the stained grooves are measured in combination with the translation stage and the microscope arrangement. A typical schematic of the junction depth by groove technique is shown in Fig. 8.12. The groove can be made either using a cylinder rod or using a spherical ball.

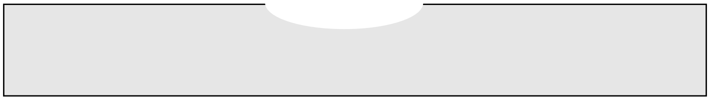


Fig. 8.12 Junction-depth measurement by groove technique

If R_{ball} is the radius of the spherical ball, then the junction depth can be expressed as

$$x_j \approx \frac{a^2 - b^2}{2R_{\text{ball}}} \tag{19}$$

where x_j is the junction depth, a is the radius of the silicon groove, and b is the radius of the dark p-type silicon (in case of boron diffusion).

Sheet Resistivity

Film resistance is also called sheet resistivity and it is a very important parameter to calculate the delay in any IC circuit. Basically, sheet resistivity is used to know the resistance of a particular material. As the thickness of the material used in the fabrication of an IC is of the order of a few Augustan to a few microns (except in case of silicon wafer), the sheet resistivity has different values as compared to that of the bulk of that material. Thereafter, the resistance of the material is found out from the sheet resistivity. Once the resistance of the material is obtained, the delay in IC circuit is calculated.

The resistance of a bulk material is expressed as

$$R = \rho \frac{L}{A} \Omega \tag{20}$$

where ρ is the resistivity, L is the length, and A is the area of the material.

If the depth of the diffused layer in the silicon, i.e., if junction depth is J_n then (width X thickness)

$$R = \frac{\rho}{j_n} \frac{L}{W} \Omega$$

If we consider $L=W$, i.e., a square whose length and width are equal; then the resistance per square can be expressed as

$$R = \frac{\rho}{j_n}\frac{L}{W} = R_s \tag{21}$$

where $R_s = \frac{\rho}{j_n}$ (Ω/□) and is called the sheet resistivity (resistance). Many a times, ρ is written as ρ_s in the above equation.

Sheet resistivity is measured using the four-probe technique. In the four-probe technique, four equally spaced needles (probes) are placed in a straight line. Generally, the distance between the probes is around 1 mm. These four probes are attached to a single probe called the **probe head.** During sheet measurement, the probe head is brought down till the probes touch the surface of the diffused wafer. Thereafter, a constant current is passed from the two outer probes and the voltage drop is measured across the two inner probes, as shown in Fig. 8.13. If the diameter of the wafer is larger than the probe spacing, then sheet resistivity can be approximated as

$$R_s = \frac{\pi}{In}\frac{V}{I} \quad \text{or} \quad R_s = 4.53\frac{V}{I} \tag{22}$$

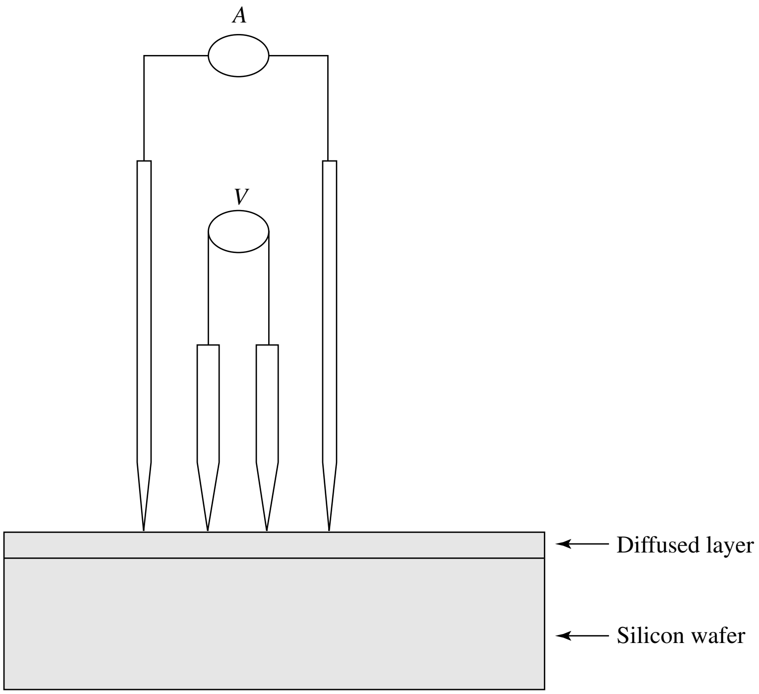


Fig. 8.13 Four-probe method to resistivity

where R_s is the sheet resistivity, V is the voltage across the inner probes, and I is the current passing through the outer probes through the doped layer.

The sheet resistivity measurement is simplified by passing a constant current of 4.53 μA through the outer probes; then the voltage that is displayed on the meter directly represents the sheet resistivity R_s.

The sheet resistivity measurement is also done on the chip; where, the sheet resistive patterns (structures) are fabricated along with IC fabrication, as shown in Fig. 8.14. These sheet resistive structures are called **Van der Paul test structures**. Generally, the length L and width W of the diffused layer (structures) are made into the wafer in a definite ratio. A constant current is passed from the two extreme contacts through the diffused layer, and voltage is measured at the two inner contacts points, as shown in Fig. 8.14.

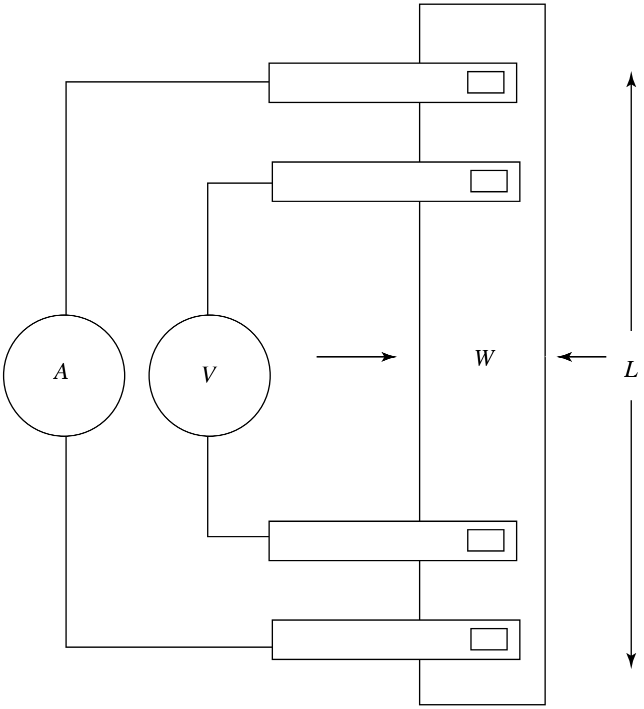


Fig. 8.14 Van der Paul test structure for resistivity measurement

Spreading Resistance Technique for Dopant Profile Measurements

One of the methods of dopant profile measurements is the spreading resistance technique. The dopant profile is obtained by measuring the resistance (not sheet resistance) between the two probes. Generally, these two probes are kept 25 micrometres apart. The typical spreading resistance measurement set-up is sketched in Fig. 8.15. The spreading resistance probes are placed on the shallow bevelled angle of the silicon wafer. Usually, the silicon wafer is bevelled using a fine diamond paste. Thereafter, two probes are placed on the wedge of the silicon and resistance is measured in the increment of 2 to 10 micrometres. Thereafter, dopant profile is obtained by the standard resistance–dopant concentration graph. This standard resistance–dopant concentration graph is called the **Irvin curve** and is shown in Fig. 8.16.

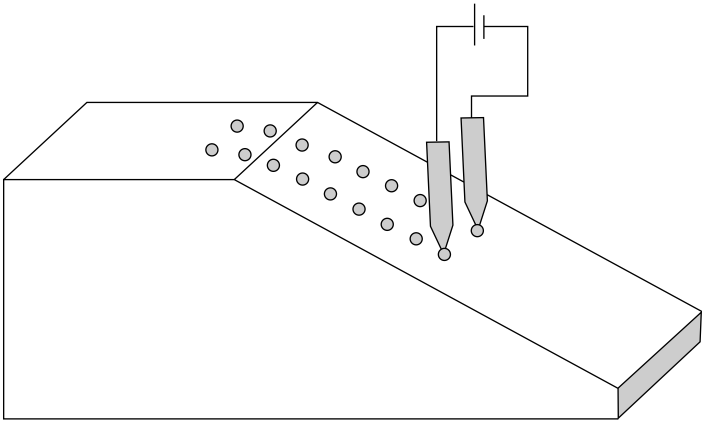


Fig. 8.15 Spreading resistance technique for dopant-profile measurements

C-V Measurement

The *C-V* measurement technique has been described in detail in Chapter 4 (oxidation). From the *C-V* curve, diffusion profile can be obtained. Unfortunately, diffusion profile measurement by this technique is not accurate. This is because the *C-V* curve depends on many structural as well as electrical parameters. In the *C-V* curve, the normalised capacitance in terms of voltage can be written as

$$\frac{C}{C_o} = \left(\sqrt{1+\frac{V_G}{V_o}}\right)^{-1} = \left(\sqrt{1+\frac{2V_G C_o^2}{qN_A\varepsilon_s\varepsilon_s}}\right)^{-1} \tag{23}$$

where C is the total capacitance, C_o is the accumulation capacitance, V_G is the gate voltage, V_0 is $qN_A\varepsilon_0\varepsilon_s/2C^2{}_0$, N_A is the substrate doping, ε_s is the semiconductor dielectric constant, and ε_0 is the permittivity of free space.

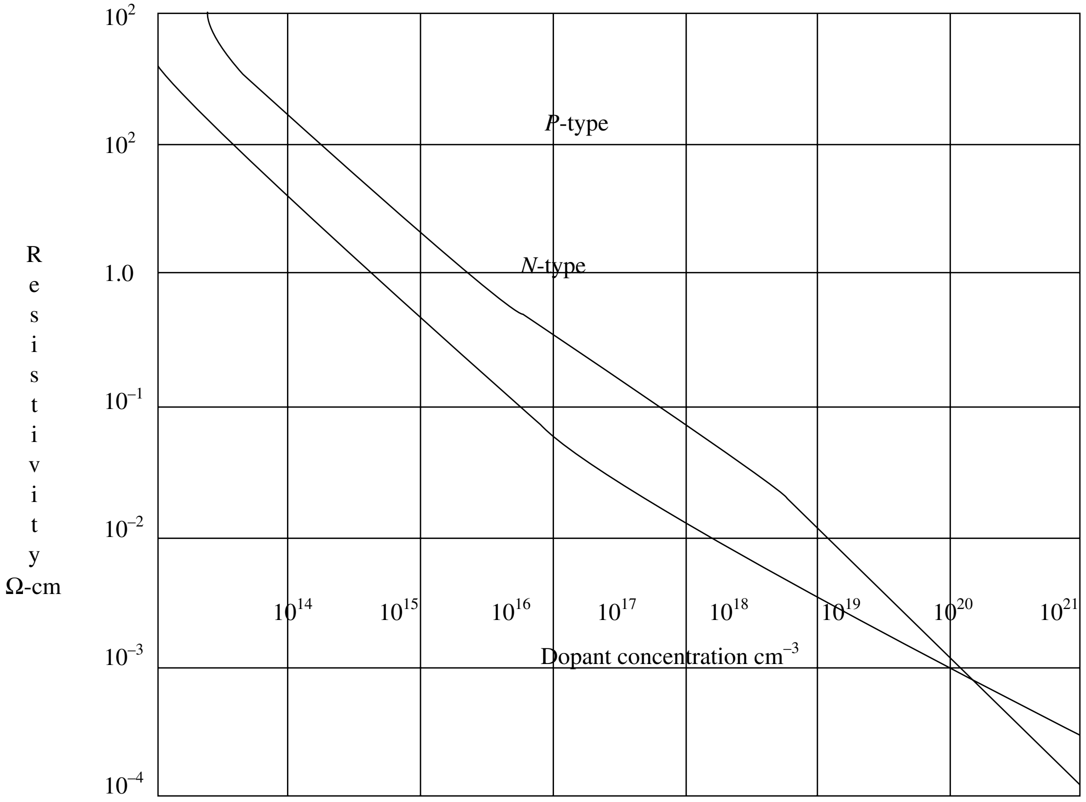


Fig. 8.16 Resistivity v/s dopant concentration (typical Irvin's curves)

If normalised capacitance is differentiated by the gate voltage then one gets

$$\frac{d\left(\dfrac{C}{C_o}\right)}{dV_G} = -\frac{1}{2}\left(\frac{C}{C_o}\right)\frac{1}{V_o} = -\left(\frac{C}{C_o}\right)^3 \frac{\varepsilon_{\text{ox}}^2 \varepsilon_o}{\varepsilon_s q N_A t_{\text{ox}}^2} \tag{24}$$

For a known value of t_{ox}^2, the profile of the dopant concentration N_A can be obtained at a particular point.

Secondary-Ion-Mass-Spectroscopy (SIMS) Technique

The diffusion profile and junction depth are accurately measured using secondary-ion-mass-spectroscopy. In this technique, the surface of the wafer is sputtered by the inert-gas ions and the mass of the sputtered elements is analysed by the sputtered depth. This technique is highly sensitive and most suitable for the dopant profile measurements for lightly doped shallow doping and junction depth. Presently, this technique is used for VLSI and ULSI fabrication.

Summary

The electrical conductivity of semiconductors changes significantly, when they are doped (introduced to) with selective elements. The element that is intentionally doped in the semiconductor and alters the semiconductor conductivity is called the dopant and the process by which the element is diffused into the semiconductors is called diffusion. Mostly, the third and fourth group elements of the periodic table, namely, phosphorus, arsenic, and boron are diffused (doped) into the silicon semiconductor for MOS transistor fabrication. A high dose of the dopant is required to make the source, the drain, the gate, the interconnection, the capacitor, and the contacts; whereas, a low dose is required for the V_T adjustment, the channel stop, the well formation, etc.

Generally, to fabricate the MOS, two types of dopant diffusion techniques are used, namely, thermal diffusion technique and ion-implantation technique. The thermal diffusion technique can be explained by two Fick's laws: one based on the classical diffusion theory and the other, the atomic diffusion model, based on the atomic interaction between the ionised dopant atoms and the electrically charged point defects present in the silicon.

The Fick's diffusion model is simple, fast, and easy to calculate. Unfortunately, Fick's model fails to predict many diffusion cases, especially those with high dopant concentration, electrical field and a few IC process-related cases. In these cases, the Fick's diffusion model is modified by incorporating experimentally found results.

At the diffusion temperature, the dopant atoms dissociate into ionised atoms and free electrons. The electron mobility is much faster in silicon than that of the ionised atoms; hence the electrons go deep into the wafer, which builds an electrical field (built-in potential) and pulls the ionised ions inside the wafer. Thus dopant diffusivity increases due to the manifestion of the electric field.

It has been established from experiments that boron atoms interact with the singly positively charged vacancy point defects V^+, and that the diffusion of boron is linear in the initial phase of diffusion and then falls rapidly at the end.

It has also been established from experiments that phosphorus atoms interact with the singly negatively charged vacancy point defects V^- and at low concentrations, follow the erfc function; but when phosphorus approaches 10^{20} per cm^{-3}, the diffusivity decreases fast up to a minimum point called the kink. At this point, the phosphorus-vacancy pair P^+V^{2-} starts dissociating into P^+, and singly charged acceptors V^- that leads to tail formation. One of the evidences of the dissociation of P^+ and V^- is the emitter-push. In addition, at higher constant–surface concentration, the diffusion profiles of boron and phosphorus are different. Boron and phosphorus concentrations alter at the oxide/silicon interface because of their different segregation coefficients when the wafer is oxidised. The diffusivity also increases due to oxidation and it is called Oxidation-Enhanced Diffusion (OED).

It is essential to measure the junction depth, the sheet resistivity, and the diffusion profile of the diffused layer. The junction depth is measured by the bevel and the groove techniques, and more accurately by the Secondary-Ion-Mass-Spectroscopy (SIMS) technique. For VLSI and ULSI applications, secondary-ion-mass-spectroscopy is suitable for the junction-depth and dopant-profile measurements. The dopant profile can also be measured by the spreading resistance technique. The sheet resistivity is measured by the four-probe technique.

References

- J D Plummer, M Deal and P B Griffin, *Silicon Fundamental Technology: Fundamentals, Practice and Modeling*, Prentice Hall, 2000
- R C Jaeger; *Introduction to Microelectronic Fabrication: Volume 5 of Modular Series on Solid State Devices*, Prentice Hall, Second Edition, 2001
- S M Sze; *VLSI Technology*, Second Edition, McGraw-Hill, 1988
- S K Gandhi; *VLSI Fabrication Principles*, Second Edition, Wiley, 1994
- D Nagchoudhuri; *Principles of Microelectronic Technology*, Wheeler, 1998
- S A Campbell; *The Science and Engineering of Microelectronic Fabrication*, Oxford University Press, 1996

Multiple-Choice Questions

8.1 Why is silicon dioxide used for dopant diffusion?
(a) To prevent diffusion
(b) To promote diffusion
(c) Does not take part in the diffusion process

8.2 Why is oxygen used at the time of dopant diffusion?
(a) To dilute the dopant in the oxide
(b) To dilute the dopant in the silicon
(c) Neither for dopant dilution in the oxide nor for dopant dilution in the silicon

8.3 Pre-diffusion is carried out
- (a) to obtain the junction depth
- (b) to introduce a controlled amount of the dopant
- (c) for redistribution of the dopant

8.4 Drive-in is carried out
- (a) to obtain the junction depth
- (b) to introduce a controlled amount of the dopant
- (c) gettering of the dopant

8.5 Around what temperature is boron diffused?
(a) 950°C (b) 1050°C (c) 1200°C

8.6 Around what temperature is phosphorus diffused?
(a) 950°C (b) 1050°C (c) 1200°C

8.7 When does the silicon wafer become intrinsic?
- (a) When the dopant atoms are more than n_i
- (b) When the dopant atoms are less than n_i
- (c) Not dependent on n_i

8.8 The electric film enhances diffusivity of the dopant by a factor of approximately
(a) 10 times (b) 5 times (c) 2 times

8.9 In what way does the emitter-push alter the base dopant?
- (a) It decreases the base dopant junction depth.
- (b) Does not alter the base dopant junction depth.
- (c) It increases the base dopant junction depth.

8.10 What is the unit of sheet resistivity?
(a) Ω per cm (b) Ω per square (c) Ω

Descriptive Problems

8.1 If 1×10^{19} atoms per cm^{-3} boron is doped into a silicon wafer and the wafer is kept in a high temperature furnace. Find the temperature at which the semiconductor becomes intrinsic.

8.2 Calculate the value of the diffusion enhancement factor (h), when the wafer has a net dopant concentration of 1×10^{20}, subject to a temperature of 1000°C.

8.3 Plot the normalised dependent diffusivity D against the diffusion temperatures of 1000°C, 1050°C, 1100°C, and 1050°C.

8.4 Prove that the junction depth is approximately equal to $\dfrac{a^2 - b^2}{2R_{\text{ball}}}$, where, R_{ball} is the radius of the spherical ball used for silicon grooving, and a and b are the radius of the silicon groove and the radius of the dark p-type silicon respectively.

CHAPTER 9

Ion-Implantation

9.1 INTRODUCTION

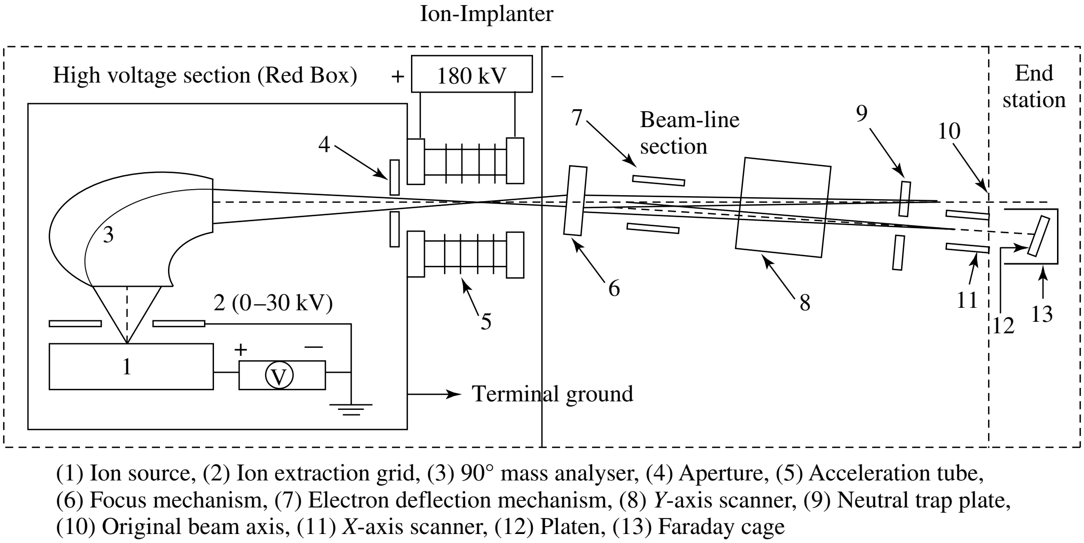
The ion-implanter was invented by William Shockley in 1954, but its importance was realised after decades of semiconductor device fabrication. The ion-implanter is used for introducing the dopant ions inside the semiconductor wafer by force. The ion-implantation technique is more versatile and advantageous than the thermal diffusion technique (Chapter 8) in many respects, especially in the recent high-density VLSI and ULSI applications. The ion-implantation technique is used for doping a specific type of species in which doping is not possible by thermal diffusion. Ion-implantation is used for many functions of which some are worth mentioning, like field channel stop, V_T adjustment of transistors, enhancement and depletion of MOS fabrication, gettering, buried layer, and formation of buried insulator layer. Furthermore, the requirements of precise doping control and shallow junction depth are most desirable for the submicron technology, and these are only possible by the ion-implantation technique. Therefore, the ion-implantation technique is indispensable in IC fabrication nowadays. Unfortunately, ion-implantation damages the wafer surface and requires an additional process of heat treatment to recover the damaged surface. In addition, ion-implantation is costly and elaborate equipment that needs constant maintenance; furthermore, it is a single wafer process so the wafer throughput is less.

9.2 ION-IMPLANTATION EQUIPMENT

The ion-implanter can be divided into three distinct sections, namely, high voltage (red box), beam line, and end-station sections, as shown in

Fig. 9.1. The red box houses the gas-delivery systems, the ion source, the ion-extraction grid, the mass analyser, and the ion accelerator. The dopant gas is introduced from the gas-delivery system into the ion source. In the ion source, the dopant gas is ionised (plasma) by filament heating and high electric field at low pressure. For this reason, the ion source is also called the **ionisation chamber** or the **plasma chamber**. The dopant ions are extracted out from the ionisation chamber by a negatively biased ion extraction grid. The extracted dopant ions then enter into the high-power mass analyser. The mass analyser separates different dopant species and impurities in the space domain. The desired dopant ions are further separated by a narrow slit filtered from the rest of the ionised atoms. This narrow slit is called the **resolution aperture** or simply the **aperture**. The filtered dopant ions then enter into the high-voltage acceleration unit. In this acceleration unit, the dopant ions are accelerated to the desired velocity by adjusting the high-voltage acceleration unit.

Generally, arsine (AsH_3), phosphine (PH_3), and boron difluoride (BF_2) gases are used for arsenic, phosphorus, and boron implantation respectively. The arsine and phosphine gases are toxic in nature; therefore, in many laboratories, less toxic solid dopant sources are used. These solid state sources release vapours of the dopant after it is heated in the crucible. Generally, the solid sources of phosphorus and arsenic are in the form of compounds. The high voltage section, i.e., the red box is kept at a low pressure of around 10^{-3} torr to avoid interaction with the residual gas and the oxidation of ion-implantation components, especially, the ion-source filament.



(1) Ion source, (2) Ion extraction grid, (3) 90° mass analyser, (4) Aperture, (5) Acceleration tube, (6) Focus mechanism, (7) Electron deflection mechanism, (8) *Y*-axis scanner, (9) Neutral trap plate, (10) Original beam axis, (11) *X*-axis scanner, (12) Platen, (13) Faraday cage

Fig. 9.1 Ion-implanter

After leaving the red box, the dopant ions enter into the beam-line section of the ion-implanter. The beam-line section uses a high voltage to accelerate the dopant ions. The high-velocity ions then pass through another aperture (slit), and are made to scan the entire wafer by electrostatically deflecting the *X-Y* plates. This beam-line section is evacuated to less than 10^{-6} torr using high-vacuum systems to minimise the collision of ions with the residual gases, so that they do not deviate from their original path; it also reduces the probability of ion and electron recombination to form neutral atoms. The beam-line section is slightly tilted from its original beam axis, so that the neutral atoms do not strike the wafer, as shown in Fig. 9.1.

The dopant ions then enter into the end station section of the ion-implanter. At the end of the end station, a wafer is held almost perpendicular to the ion beam by a special type of jig called the **platen**. The end station is kept at 10^{-6} torr or higher vacuum, to avoid the loss of kinetic energy of the ions, beam direction and electrical charge neutralisation due to recombination. Outside the end station, the wafer delivery (feeding) system is attached (not shown in the figure). The wafer-delivery system consists of two vacuum load-lock systems. The upper load-lock system feeds the wafer into the platen for implantation and the lower load-lock system collects the wafer after the implantation. These load-lock systems are used to avoid breaking the vacuum of the implanter for loading and unloading of the wafer so that it allows an uninterrupted implantation process. The quantity of the implanted ions (ion dose) is measured precisely by the electronic charge integration circuit.

9.3 ION-IMPLANTATION PARAMETERS

Ion-implantation has two main process parameters, namely, implantation energy and implantation dose.

9.3.1 Ion-implantation Energy Parameters

The implantation energy decides the penetration depth of the ions inside the silicon wafer. The higher the implantation energy, the deeper the dopant ions in the silicon; for example, the ion can be implanted inside the silicon wafer from the shallow junction depth of around 100 Å to 1 μm (10,000 Å) by adjusting the implantation energy of the acceleration voltage from 3 keV to 400 keV. For a very shallow junction depth, the implantation can be done in the de-acceleration mode.

In the ionisation chamber, most of the ions are singly charged, but a small number of doubly (~10%) and triply charged (~1%) ions are also present in the plasma. For the same acceleration voltage, the doubly charged ions are accelerated twice ($E = 2qV$) than the singly charged ions. Similarly, for the same acceleration voltage, the triply charged ions are accelerated thrice ($E = 3qV$) than the singly charged ions. In other words, to attain the same penetration depth, the singly charged ions need 400 keV, whereas the doubly charged ions need 200 keV of energy. As for their presence in the plasma, the number of doubly and triply charged ions is significantly less. The duration of implantation of the doubly and triply charged ions is significant as compared to that of the singly charged ions. Therefore, usually the higher order charges are not implanted.

9.3.2 Implantation Dose Parameter

The amount of dopant atoms introduced in the silicon wafer is called the **dose**. The implanted ions do not go deep, but remain just below the wafer surface as a sheet; therefore, ion-implantation dose is expressed in terms of area (cm^{-2}). The implantation dose is a function of the ion beam current and the duration of implantation. A higher beam current is obtained by the higher extraction voltage of the red box. The ion-implanter with a high beam current of around 10 mA is available, but at the high-beam current, the wafer gets heated up. In addition, the previously implanted ions get knocked out of the wafer by the ions implanted subsequently. To avoid these issues, the platen is cooled. Another way of increasing the ion dose is to increase the duration of implantation. The ion dose is measured electronically by the charge integrated circuit, as shown in Fig. 9.2. When the ions go inside the wafer, they get neutralised by capturing the electrons from the wafer, which is connected to the charge integrated circuit through the platen. The charge integrator counts the number of electrons per unit time during the implantation. The implantation ions (Q_{imp}) can be expressed in terms of the current and the time as

$$Q_{\text{imp}} = \int I\, dt$$

where I is the current (number of electrons per unit time) in the time duration dt.

The ion dose is measured in terms of the ion charger Q, and the area as

$$D = Q_{\text{imp}} / mqA \text{ ions /cm}^2$$

where A is the area of implantation, q is the electron's charge, and m is the state of the ion charge (singly or higher charges), the charge associated with an ion is mq and m stands for singly ionised or doubly ionised charge. For singly and doubly charges, m is 1 and 2 respectively. The implantation dose for the known charge state (m) with time, can be expressed as

$$D = \frac{1}{Am} \int \frac{I}{q} dt$$

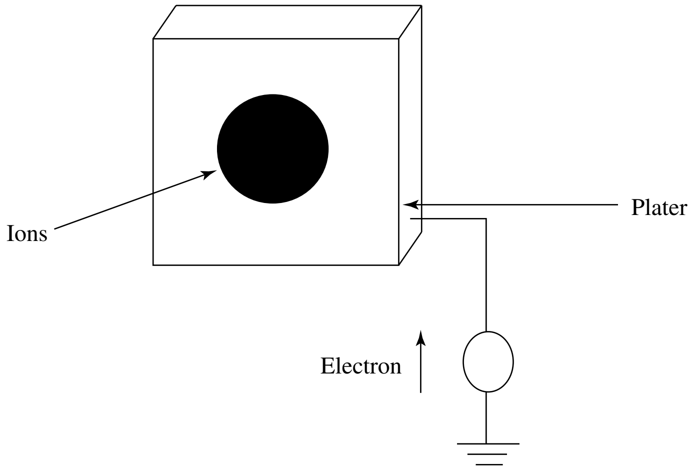


Fig. 9.2 Charge integrator

9.4 PARAMETERS AFFECTING THE DOSE AND UNIFORMITY

There are many parameters that may lead to wrong information about the ion dose. The charge integrator reads only the ion charge and not the neutral dopant atoms. If the neutral atoms are implanted along with the charged ions into the wafer then the charge integration information will be incorrect. These neutral atoms are generally created in the beam line section of the implanter. To prevent the implantation of neutral atoms, the end station of the implanter is tilted by 7–10 degrees from the ion beam axis, as shown in Fig. 9.1. The second cause of inaccuracy in dose monitoring is the knocking out of electrons from the wafer during implantation. This phenomenon is called **secondary electron emission**. Secondary electron emission occurs when high-velocity ions strike the wafer. These electrons are collected by the positively charged metal cage called the **Faraday cup** and added to the dose. The third parameter leading to erroneous dose measurement is the sputter material that either deposits on the wafer or gets implanted in the wafer. This sputter material is generated when the ion beam strikes on the aperture and the other parts of the implanter. To minimise the sputtering effect, low-sputtering materials such as carbon or silicon are used for the aperture and other vulnerable parts of the implanter. The fourth parameter of inaccurate dose measurement is the sputtering (knocking out) of the previously doped atoms from the wafer, by the successive ion strikes. This phenomenon can be minimised by heating the wafer during implantation. By heating, the previously doped atoms move inside the wafer and the probability of sputtering of the dopant ions is reduced.

9.5 ADVANTAGES OF ION-IMPLANTATION

There are many advantages of the ion-implantation technique over the thermal diffusion technique such as more accuracy, high reproducibility, high doping range, greater range of process temperature, ability of doping through thin film, greater profile control, short-channel doping, and the implantation of pure dopant and molecular dopant. These advantages are described below.

9.5.1 Dose Accuracy

An accurate ion dose is measured during implantation. There is no mechanism by which diffusing flux can be measured during thermal diffusion. Furthermore, it has been found experimentally that a dose accuracy of 1–5% is possible by the implantation technique as against the dose accuracy of 5–10% by the thermal diffusion technique.

9.5.2 Reproducibility

The reproducibility of dose is very high in the ion-implantation technique. There is no dose variation across the wafer. Furthermore, the ion-implanted dose is measured electronically during the implantation process and is very precise. Therefore, no dose variation occurs from one wafer to the other which means that the **run-to-run reproducibility** is high. In thermal diffusion, the dose is heavily dependent on the pre-optimised diffusion recipe; hence, it varies in run to run (one batch of diffusion to the next batch of diffusion) wafer diffusion. In addition, there is doping variation across the wafer.

9.5.3 Doping Concentration

One of the biggest advantages of ion-implantation is the doping range. Ion-implantation can be carried out from a minimum dose of ~10^{11} atoms /cm^2; whereas in the thermal diffusion technique, it is not possible for a dose less than 10^{13} atoms/cm^3. On the other hand, higher doping level is also limited by the solid solubility (>10^{16} atoms /cm^3) in the thermal diffusion technique, but there is no such limitation in the ion-implantation technique.

9.5.4 Process Temperature

The ion-implantation technique can be carried out at room temperature. This excellent property of the ion-implantation technique gives flexibility to use the photoresist as an implantation mask. On the contrary, thermal diffusion is a high temperature process; therefore, the PR cannot be used, and in place of the PR, silicon dioxide or some other type of dielectric film is needed. In addition, the previously diffused dopant profile also changes.

9.5.5 Doping through Thin Film

The ions can be implanted through a thin film. Usually, the implantation is done through a thin oxide (SiO_2); whereas in the thermal diffusion process, the dopant atoms cannot penetrate through the dielectric film and it is usually used as a thermal diffusion mask. Furthermore, depending on the thickness of the thin film and the implantation energy, the implantation depth can be adjusted inside the wafer. In addition, the thin film protects the silicon wafer from environmental contamination and also from the channelling effect. The channelling effect is explained in Section 9.9.

9.5.6 Profile Control and Junction Depth

Different dopant profiles in the silicon can be obtained with different ion-implantation energies, as shown in Fig. 9.3. A desired profile depth can be obtained by selecting a particular implantation energy. The implantation depth can be adjusted anywhere from 100 Å (at ~1 keV) to 10 µm (at ~1 MeV) in the silicon wafer. This quality of ion-implantation is exploited for the fabrication of buried insulator and buried conductor layers. The buried insulator layer of silicon dioxide or silicon nitride is created by the implantation of oxygen or nitrogen ions. When the wafer is heated in a furnace, the implanted oxygen/nitrogen reacts with the silicon and forms a silicon dioxide or a silicon nitride film. The buried insulator formation on the silicon wafer is used for **Silicon on Insulator** (SOI). Similarly, the buried conductor layer can also be formed by implanting the boron or phosphorus ions in the wafer at a very high implantation energy. These buried layers are used for many applications in the IC fabrication. In contrast, in thermal diffusion, the peak concentration of the dopant is always the highest at the surface of the wafer and it decreases with depth of the wafer, as described in Chapter 8 (diffusion). Therefore, the tailormade dopant profile and the buried layers cannot be formed by the thermal diffusion process. The buried layers can be made by the thermal diffusion technique, but they need extra processing steps and equipment.

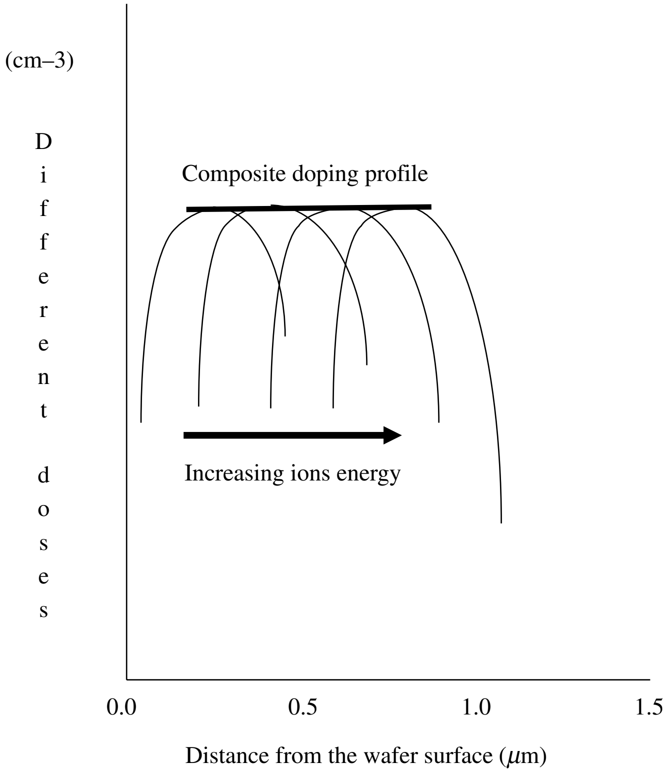


Fig. 9.3 Typical implant profile at different implantation energy (keV)

9.5.7 Short-channel Device

The ion-implantation technique is also used to check the short-channel effect. As the density of the transistors is increasing day by day, the source and the drain are coming closer and closer. When the transistor is biased, the drain depletion extends towards the source and influences the channel length, and that results in a high current. This phenomenon is called the **short-channel effect**. The short-channel effect is prominent in the deep junction depth transistor and less prominent in the shallow junction depth transistor. To reduce the short-channel effect, the source and the drain are implanted two times with different dose and energies. In the first implantation, a light dose of around 5×10^{13} atoms cm^{-2} at low implantation energy is implanted in the source and the drain. This makes lightly doped shallow in the wafer, and thereafter, a second implantation with high dose and high energy is done for the source and the drain formation. The lightly doped implanted dopant extends towards the gate more than the highly doped source and drain. This extended lightly doped source and drain is called the **tip** or the **extension**, or the **Lightly Doped Drain (LDD)** region. The LDD structure increases the resistance that reduces the current, and checks the short channel. This LDD structure is also helpful in reducing hot electron generation in between the source and the drain regions. The short-channel and hot electron generation becomes prominent when the transistors are scaled down to submicron geometry (Chapter 11). The process sequence of LDD formation is described in Chapter 3.

9.5.8 Pure Dopant

The highly pure dopant ions can be implanted by the ion-implantation technique. The mass analyser separates the specific dopant species from the other dopant and impurities; for example, BF_2 gas dissociates into B^{++}, B^+, and F^+ elements, the BF^+ and ${BF_2}^+$ molecules and the boron isotopes (mass 10 and mass 11). These dissociated individual elements, molecules or isotopes can be implanted.

The separating dopant species in the space domain can be expressed as

$$RB = \sqrt{\frac{2VM}{mq}}$$

where R is the radius of curvature of the element or elements, B is the magnetic field, M is the mass of the ion, V is the accelerator voltage, m is the ionised charge, and q is the electron charge.

The trajectory of the ion can be adjusted by adjusting the magnetic field of the mass analyser so that only the desired species pass through the aperture (slit) and get implanted, and the rest of the undesired species are blocked.

9.5.9 Molecular Dopant

The molecular species (e.g. BF_3) can also be implanted in the wafer. The beam current of BF_3 implantation is many times higher than the beam current of single-charged boron (B^+). The high beam current of BF_3 implantation reduces the implantation time manifold. After the implantation, BF_3 dissociates into boron and fluorine elements inside the wafer. Fluorine rapidly goes out of the wafer leaving behind the boron ions. In addition, BF_3 implantation makes the wafer surface more amorphous, which also reduces the undesired channelling effect.

9.6 DISADVANTAGES OF ION-IMPLANTATION

The ion-implanter is a sophisticated and costly equipment, and it needs constant care as well as supporting services. The implanter consists of high- and low-voltage circuits and most of these circuits are interfaced through opto-couplers, and relays, and to control the implantation process, low-voltage logic circuits are also used. In addition, it consists of vacuum systems and many movable mechanical parts. Therefore, the implanter develops faults very frequently; thus, its up time (working hours) is less and it needs constant attention and frequent maintenance.

Ion-implantation damages the crystallinity of the silicon wafer. Highly energetic dopant atoms displace the silicon atoms from their lattice sites that makes the silicon surface almost amorphous. To bring back to the crystalline form and repair the damage, the wafer is subjected to high temperature in the nitrogen environment for annealing.

9.7 ION-IMPLANTATION MODEL

9.7.1 Ion Range Distribution

The dopant distribution of ion-implantation is almost like a Gaussian distribution where the maximum numbers of ions (peak concentration) stop at a specific depth (distance from the wafer surface). This concentration peak from the wafer surface is called the **projected range** and is denoted by R_p. The spread of dopant ions around the projected range is called the **standard deviation or straggle** and is denoted by ΔR_p. The typical implanted ion distribution is shown in Fig. 9.4.

The Gaussian distribution of the dopant ions can be expressed as

$$C(x) = Cp\exp\left(-\frac{(x-Rp)^2}{2\,\Delta\,R_p^2}\right) \tag{1}$$

where *C(x)* is the ion distribution inside the silicon, *Cp* is the peak concentration of ions at R_p (projected range), and ΔR_p is the standard deviation or straggle. The projected range (R_p) and the standard deviation (ΔR_p) are also called the **first-order moment**” and the **second-order moment**” respectively.

The total number of the ions implanted (dose) can be expressed as

$$Q = \int(x)dx \tag{2}$$

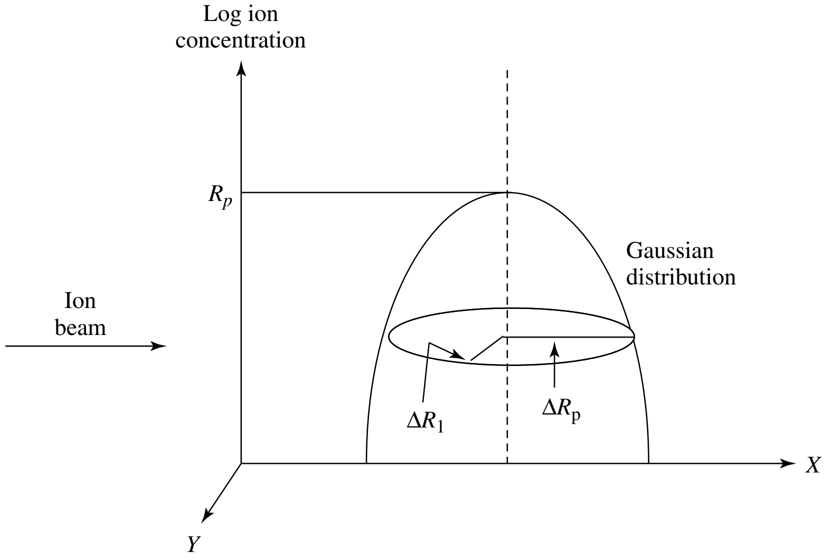


Fig. 9.4 Typical implanted ion distribution

where Q is the dose and x denotes the depth of the ions from the silicon surface of the wafer.

$$Q = \sqrt{2\pi}\ \Delta\ R_p C_p \tag{3}$$

where C_p is the peak concentration of the ions given by

$$Cp \approx \frac{0.4Q}{\Delta R_p} \tag{4}$$

Comparing with Equation 9.1 and the drive-in equation of thermal diffusion,

$$C(x,t) = C(0,t)\exp\left(-\frac{x^2}{4Dt}\right) \tag{5}$$

we find that the diffusion length $4Dt$ is $2\Delta R_p$ and the ion distribution is shifted along the depth on the x axis by R_p.

The ion distribution that takes place on the y axis is called the **lateral diffusion** and is expressed as

$$C(x,y) = C_{\text{lateral}}(x)\exp\left(\frac{y^2}{2\Delta R_\perp^2}\right) \tag{6}$$

where the perpendicular sign in the above equation depicts the standard deviation or straggle in the y-direction, which is perpendicular to the x-axis. It is found that the ion concentration in the y direction is considerably insignificant as compared to that in the thermal diffusion.

Example 9.1

A bare silicon wafer of 6″ (125 mm) diameter is implanted with boron at 1×10^{-6} A implantation current. Find the total number of boron ions at peak concentration in the wafer and the time of implantation, in the following cases.

(i) Dose of 1×10^{13} ions/cm^2 with 30 keV, which has a projected range of 0.1 μm and a projected straggle of 0.035 μm.

(ii) Dose of 5×10^{18} ions/cm^2 with 80 keV, which has a projected range 0.25 μm and a projected straggle 0.063 μm.

Answer

Case I

Dose $D_s = 1 \times 10^{13}$ ions/cm^2

Implanted energy = 30 eV

$$R_p = 0.1\ \mu\text{m}$$

$$\Delta R_p = 0.035\ \mu\text{m} = 0.035 \times 10^{-8}\ \text{cm}$$

The peak concentration of Gaussian distribution is

$$N_{\text{peak}} = \frac{D_s}{\Delta R_p\sqrt{2\pi}} = \frac{1\times10^{13}}{0.035\sqrt{6,28}} = 1.14\times10^{18}\ \text{cm}$$

The total number of boron ions implanted in the 12.5 cm diameter silicon wafer is

$$Q = 1\times10^{13}\times3.14\left(\frac{12.5}{2}\right)^2$$
$$= 1.22\times10^{15}\ \text{atoms}$$

Implantation time

$$t = \frac{qQ}{I} = \frac{(1.6\times10^{-19})Q}{1\times10^{-6}}$$
$$= \frac{(1.6 \text{ x } 10^{-19})\times1.22\times10^{15}}{1\times10^{-6}} = 1.95\times10^{2}\ \text{sec}$$
$$= \frac{195}{60} = 3.25\ \text{min}$$

Case 2

Dose D_s = 5 × 10^{18} ions/cm^2

Implanted energy = 80 eV

$$\Delta R_p = 0.063\ \mu\text{m}$$

The peak concentration of Gaussian distribution is

$$N_{\text{peak}} = \frac{D_s}{\Delta R_p \sqrt{2\pi}} = \frac{5 \times 10^{18}}{\sqrt{6.28} \times 0.063} = \frac{5 \times 10^{18}}{0.1578} = 3.166 \times 10^{23}\ \text{cm}$$

The total number of boron ions implanted in the 12.5 cm diameter silicon wafer is

$$Q = 5 \times 10^{18} \times 3.14 \left(\frac{12.5}{2}\right)^2 = 613.28 \times 10^{18} = 6.13 \times 10^{20}$$

Implantation time

$$t = \frac{qQ}{I} = \frac{(1.6 \times 10^{-19}) \times 6.13 \times 10^{20}}{1 \times 10^{-6}} = 9.80 \times 10^7$$

$$= \frac{9.8 \times 10^7}{60} = 0.1633 \times 10^7\ \text{min}$$

$$= 2.721 \times 10^4\ \text{hrs}$$

Case 1 is used for channel stop implantation and Case 2 is used for source and drain implantation.

The channel stop and threshold voltage implanted current can be taken as an example, but for the source and drain implantation, a much higher implantation current is needed to reduce the time drastically.

Example 9.2

Find the threshold voltage when 10^{15} atoms/cm^3 boron is doped in 10^{13} atoms/cm^3 phosphorus intrinsic silicon wafer at 300° K temperature and the oxide thickness is 800 Å.

Answer For *P*-type,

$$2\phi_f = \frac{E_i - E}{q}$$

$$p_{po} = N_A = 10^{15}\ \text{atoms/cm}^3$$

$$n_i = 1.45 \times 10^{10}\ \text{atoms/cm}^3$$

$$2\phi_f = 0.0259 = 0.288\,\text{V}$$

Therefore, $\phi_f = 0.144\,\text{V}$

$$V_T = 2\phi_f + \frac{\sqrt{2\varepsilon_s\varepsilon_0 q N_A 2\phi_f}}{C_{ox}}$$

$$= 0.288 + \frac{\sqrt{2\times 11.7\times 8.85\times 10^{-14}\times 1.9\times 10^{-19}\times 10^{15}\times 0.288}}{C_{ox}}$$

$$= 0.637\ \text{V}$$

In reality, the distribution of the implanted ions is not a true form of the Gaussian distribution. The lighter ions like the boron atoms collide with the heavier silicon atoms that result in the bounce back of many of them towards the wafer surface. This phenomenon is called **back scattering**. On the other hand, arsenic and antimony, which are heavier than the silicon atoms, penetrate deeper than the projected range and there is hardly any back scattering towards the wafer surface. In both cases, the ion distribution is skewed away from the ideal Gaussian distribution, as shown in Fig. 9.5. Apart from this skewness, it has also been found that some of the ions go very deep inside the wafer due to ion channelling, especially in the case of boron ions, and this

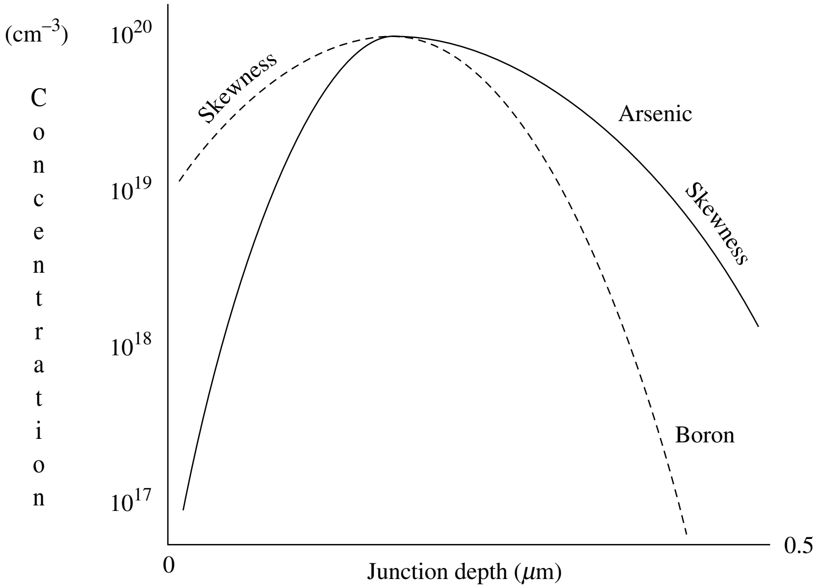


Fig. 9.5 Typical skewness of implanted boron and arsenic ions

phenomenon is called **tail formation**. The dopant concentration in the skew and the tail are calculated by the third and fourth order moments respectively. All the four normalised moments are expressed below.

The first-order moment is simply the **projected range** and is expressed as

$$R_P = \frac{1}{Q}\int_{-\infty}^{\infty} xC(x)dx \tag{7}$$

The second moment is the **standard deviation or straggle** that describes the spread of the ion distribution and is expressed as

$$\Delta R_P = \sqrt{\frac{1}{Q}\int_{-\infty}^{\infty} (x - R_P)^2 C(x)dx} \tag{8}$$

The third moment describes the **skewness** of the ion distribution and is expressed as

$$\gamma = \frac{\int_{-\infty}^{\infty} (x - R_P)^3 C(x)dx}{Q\Delta R_P^3} \tag{9}$$

and
The fourth order moment describes the **tail** (called **Kurtosis** in statistics) of the ion distribution and is expressed as

$$\beta = \frac{\int_{-\infty}^{\infty} (x - R_P)^4 C(x)dx}{Q\Delta R_P^4} \tag{10}$$

For accurate ion distribution calculation (simulation), all the four moments are taken into account. The ion distribution under the various values of the moments leads to the Pearson-*IV* distribution, which is available in tabular form.

9.8 ION STOPPING

The ion stopping in the silicon is due to two mechanisms, namely, nuclear stopping and electronic stopping. In nuclear stopping, there is a direct interaction between the ions and the silicon atoms. The nuclear ions are due to the Coulombic stop and the head-on collision nuclear stop; whereas, the electronic stoppage is due to the interaction between the ionised ions and the polarised dielectric field, and the interaction of the cloud of electrons of silicon atoms with the dopant ions.

The average rate of ion energy (E) loss with wafer depth x can be expressed as

$$\frac{dE}{dx} = S_n(E) + S_e(E) \tag{11}$$

where S_n *(E)* and S_e *(E)* denote the nuclear and the electronic stoppage mechanisms respectively.

The total distance (R) travelled by the ion can be written as

$$R = \int_0^R dx = \int_0^{E_0} \frac{dE}{S_n(E) + S_e(E)} \tag{12}$$

where R is the ion range and E_0 is the initial ion energy.

9.8.1 Nuclear Stopping

It has been found that the nuclear stopping power S_n *(E)* is a function of the dopant ionic mass and the silicon mass and this nuclear stopping power dominates over the electronic stopping power, where the ion mass is greater than the silicon mass.

The implanted ions are stopped by nuclear stopping by two mechanisms. In the first mechanism, the implanted ions interact with the nuclear electric field of the host atoms and lose the ion energy due to Coulomb interactions. In the second mechanism, the dopant ions directly collide with the silicon atoms.

In the first mechanism, the Coulomb interaction between the ions and the silicon atoms is expressed as

$$V(r) = \frac{q^2 Z_1 Z_2}{4\pi\varepsilon r} \tag{13}$$

where *V(r)*, Z_1, Z_2, ε and r are the coulomb potential, atomic numbers, dielectric constant, and distance between the ion and the nucleus respectively.

The above equation is modified on account of the ions steering away from their original path as

$$V(r) = \frac{q^2 Z_1 Z_2}{4\Pi r\varepsilon} \exp\left(\frac{r}{a}\right) \tag{14}$$

where a is the screening distance and is represented as

$$a = \frac{0.88a_0}{(Z_1^{2/3} + Z_1^{2/3})^{1/2}} \tag{15}$$

where a_0 is the Bohr radius.

The second nuclear stopping mechanism is by the head-on collision between the implanted ions and the silicon atoms, as shown in Fig. 9.6. The impact energy E_{trans} can be expressed using the classical theory as

$$E_{\text{trans}} = \frac{4m_1m_2}{(m_1 + m_2)^2}E \tag{16}$$

where m_1 and m_2 are the mass of the ion and silicon nucleus, and E_{trans} is the impact energy.

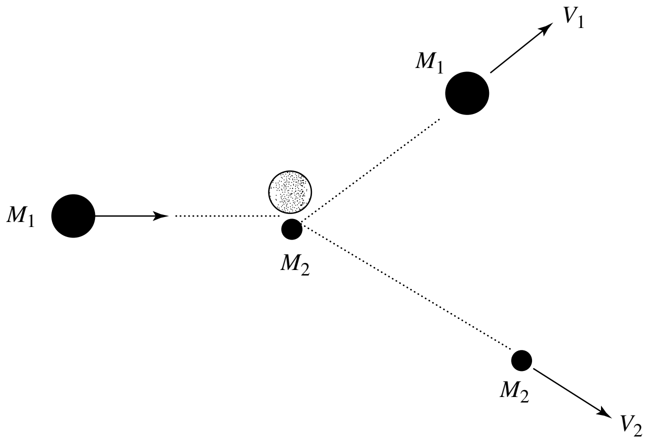


Fig. 9.6 Nuclear stoppage collision of hard spheres

The total nuclear stopping S_n *(E)* is expressed as

$$S_n(E) = 2.8 \times 10^{15}\frac{Z_1Z_2}{(Z_1^{2/3} + Z_2^{2/3})^{1/2}}\frac{m_1}{m_1 + m_2} \tag{17}$$

9.8.2 Electronic Stopping

The implanted ions are stopped by electronic stopping by two mechanisms, namely, nonlocal electronic stopping and local electronic stopping. In the non local electronic stopping, the implanted ions stop due to the dielectric polarisation. When an implanted ion passes at a high-speed through the dielectric media, it disturbs the equilibrium of the electric field of atoms in the dielectric. To minimise the electric field, the dielectric gets polarised, as shown in Figs. 9.7 and 9.8. The manifested polarisation field drags the ion backwards and restricts the ion from going deep inside the silicon wafer. In the local electronic stopping, the ion collides with the localised cloud of electrons of the silicon atoms, and in this process, the implanted ion loses its velocity. The electronic stopping is expressed as

$$S_e(E) = k\sqrt{E} \tag{18}$$

where k is the coefficient and its value is 0.2×10^{15} $\text{eV}^{1/2}$ cm^2 for silicon, and E is the ion's energy.

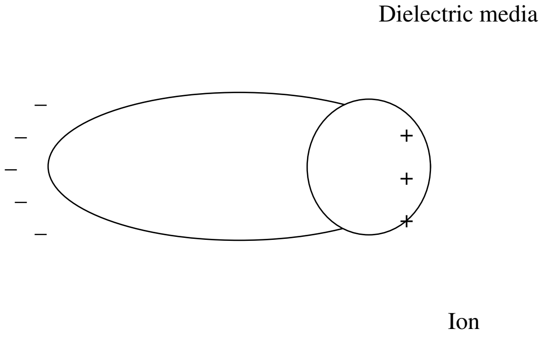


Fig. 9.7 Electronic stoppage mechanism

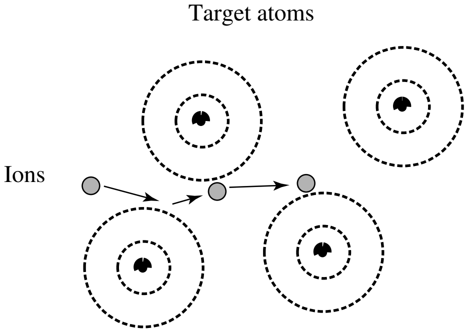


Fig. 9.8 Retarding electric field

It has been found that at higher energy of implanted ion, electronic stopping is significant and silicon crystallinity is also less damaged; whereas, at lower energy of the implanted ion, the nuclear stopping is significant and crystallinity of the silicon is also significantly damaged. The damage created by these two stoppings is shown in Figs. 9.9 and 9.10. At high energy, most of the ions, for example, the boron ions pass very quickly through the host nucleus and there is hardly any time for Coulombic interaction, as shown in the figure. The electronic stopping dominates till the velocity of the ion reduces significantly and, thereafter, nuclear stopping takes place. On the other hand, the heavy ions like arsenic are stopped by the nuclear stopping irrespective of their energies and these ions are mostly stopped near the wafer surface. A typical stopping profile of the boron ions (B), the arsenic ions (As) and the phosphorus ions (P) is shown in Fig. 9.11, where the electronic stopping crosses over the nuclear stopping at 10, 130 and 700 keV for boron, phosphorus, and arsenic respectively.

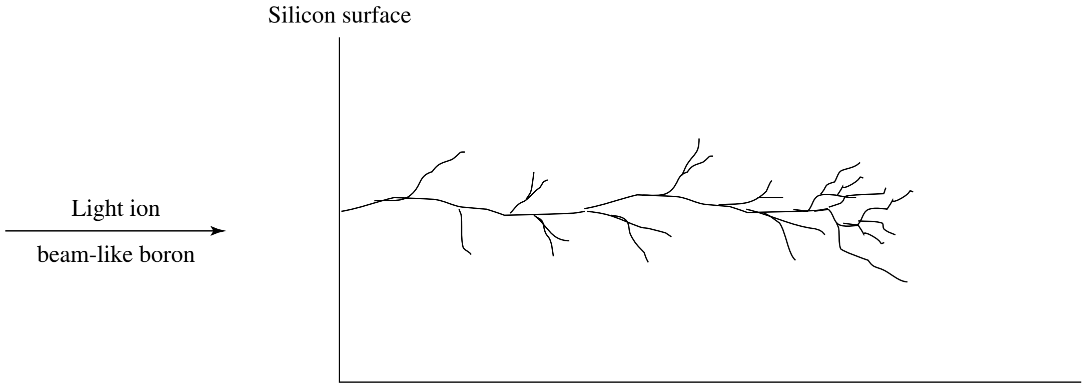


Fig. 9.9 Damage due to electric stopping from the surface of a silicon wafer

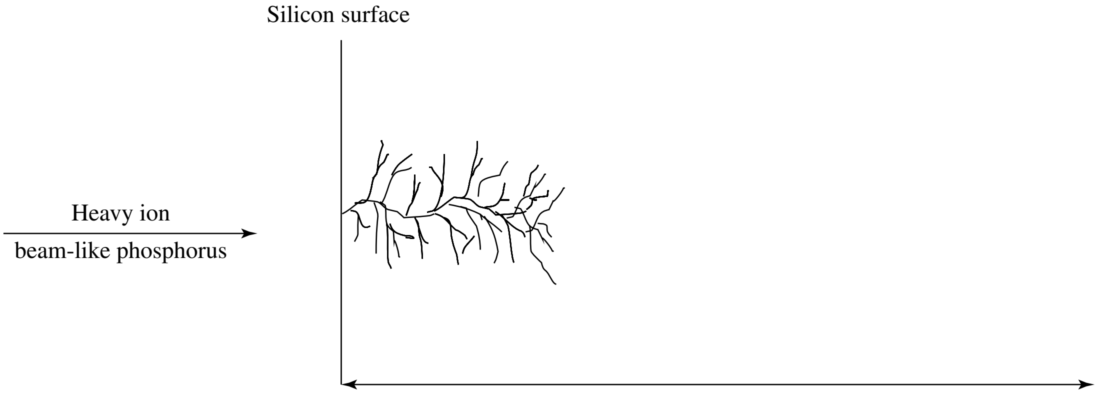


Fig. 9.10 Damage due to nuclear stopping from the surface of a silicon wafer

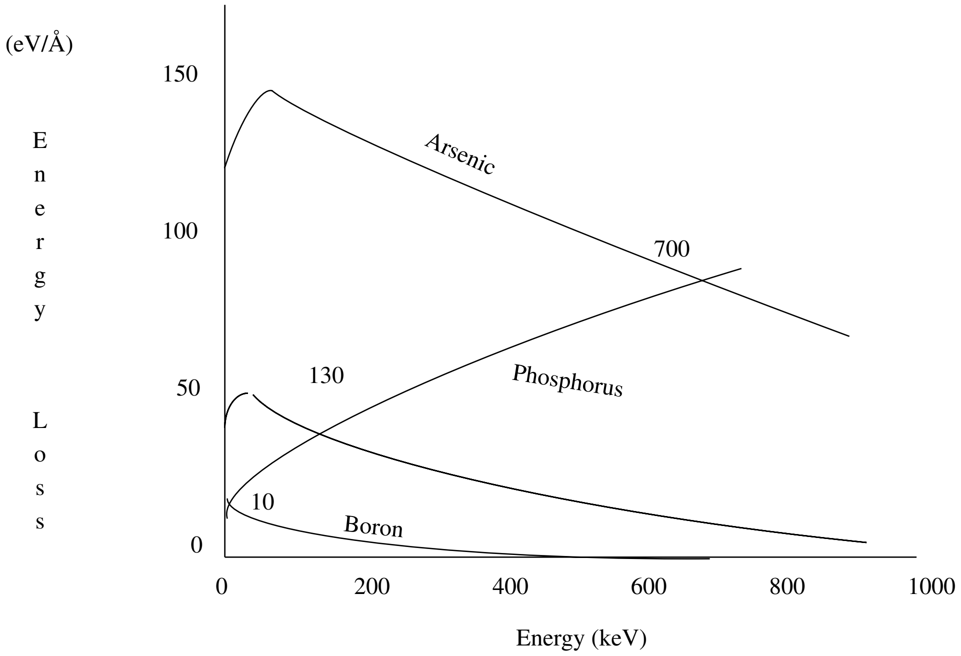


Fig. 9.11 The points of intersection show the approximate values of boron electronic stoppage correspond to phosphorus and arsenic nuclear stoppage energies

9.9 ION CHANNELLING

It has been mentioned in the previous section that many ions go deeper than ΔR_p in the wafer and form a tail region. The reason for this tail manifestation is the ion channelling effect. In the crystalline silicon, the lattices are separated by the lattice gap '*a*' in a regular fashion. Many of the implanted ions enter into these lattice gaps and travel deep inside the silicon. These ions are gradually stopped by electronic stopping. The length of the tail is a function of the implantation energy. For higher ion energy, the tail is larger and for a higher ion dose, the tail is shorter. At higher dose, the surface of the silicon wafer becomes amorphous and the amorphousness increases with the ion dose. The formation of this amorphous layer helps to reduce the tail formation. The best way to reduce the tail is to cover the wafer with a thin amorphous film, such as the silicon dioxide film. The ion channelling is dependent on the ion arriving angle ψ with respect to the lattice gap, as shown in Fig. 9.12. The maximum value of angle ψ at which the ions can enter in the channel is called the **critical angle**. It has been found that the critical angle is inversely proportional to the square root of the ion energy (i.e. $\psi \, \alpha \, 1/\sqrt{E_0}$). To increase the critical angle, the implanter platen which holds the wafer, is tilted at around 7 degrees from the ion beam. Figure 9.11 shows that at lower doses, channelling occurs significantly in spite of the platen being tilted at 7 degrees from the ion beam.

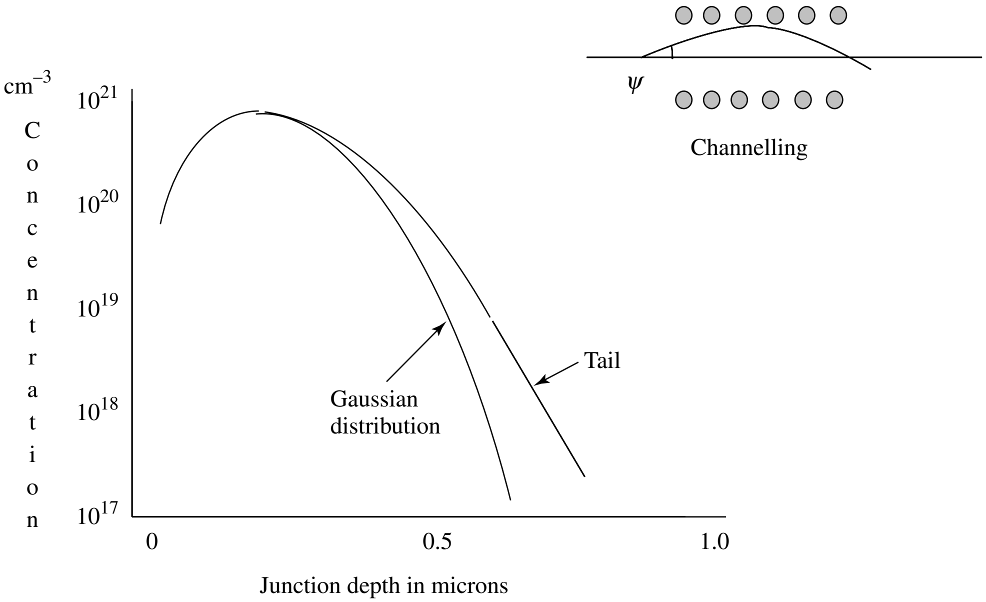


Fig. 9.12 Tail even wafer is tilted by 7 degrees

9.10 ANNEALING

Ion-implanttaion damages the crystallinity of the silicon wafer. The damaged crystalline silicon severely degrades the transistor carrier conductivity, mobility and lifetime. Therefore, the crystallinity of silicon has to be brought back to its original form. The process by which the silicon is brought back to its original crystalline form is called **annealing**. In the annealing process, the implanted wafer is heated in an inert gas/nitrogen gas/nitrogen gas mixed with hydrogen gas, for a certain period of time, at elevated temperature. The annealing process not only recrystallises the silicon, but it also puts the dopant atoms in the silicon substitutional sites. At least 90% of the total implanted ions should be put at the substitution sites of silicon to obtain good electrical characteristics of any transistor and it is called **activation**.

Annealing is done by three techniques, namely, furnace heating, light heating, and electron-beam or laser-beam heating technique. The annealing time and temperature is a function of the implanted ion dose. In the furnace heating (furnace annealing) technique, the wafers are heated in the furnace for around 30 minutes at about 950°C temperature in nitrogen gas. In the furnace annealing technique, annealing of a large number of wafers can be carried out in a single batch. The change in dopant profile and junction depth are the main drawbacks of furnace annealing, especially for the small geometry MOS.

In the light heating technique, the wafer is exposed to intense halogen light. The wafer gets heated up due to the absorption of light energy. Generally, annealing by intense light is done in a few minutes; therefore, this process of annealing is called **Rapid Thermal Annealing (RTA)**. In the RTA process, the annealing temperature ranges from 600°C to 1100°C. The redistribution of dopant atoms by the RTA process is almost insignificant. Unfortunately, RTA is a single wafer process; hence, the throughput is less. In addition, light exposure is not uniform over the entire wafer. The rapid thermal annealing (RTA) set up is shown in Fig. 9.13.

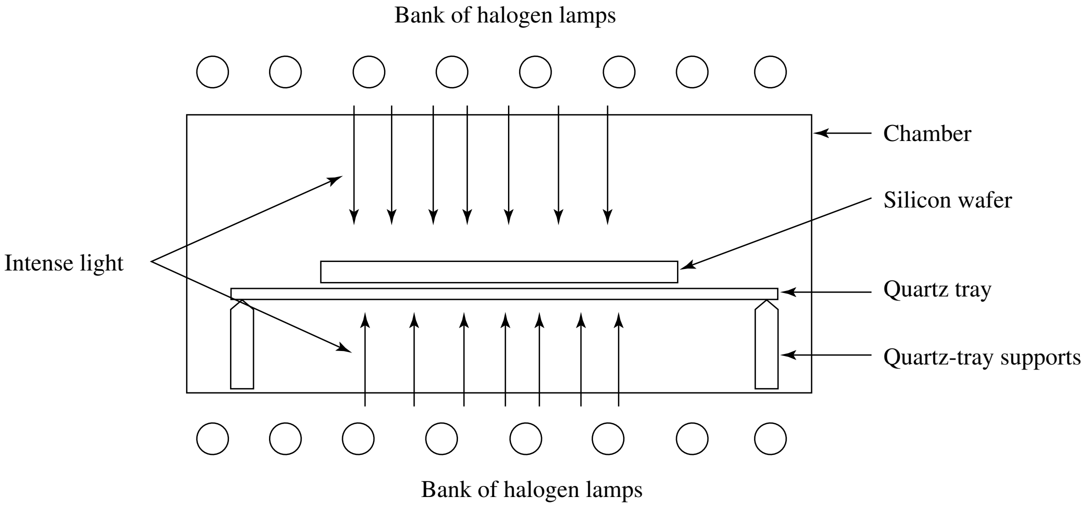


Fig. 9.13 Rapid thermal annealing

In the third annealing technique, a focused electron or laser beam is scanned over the wafer. The focussed beam of electron/laser heats the wafer. Annealing by the electron or laser beam techniques is depicted in Fig. 9.14. In the process of annealing by an electron or laser beam, the scanning beam is adjusted in such a way that the focussed beam is marginally overlapped in each scan so that the entire wafer is annealed. This annealing technique is a single wafer process and the systems are elaborate, costly, and complicated.

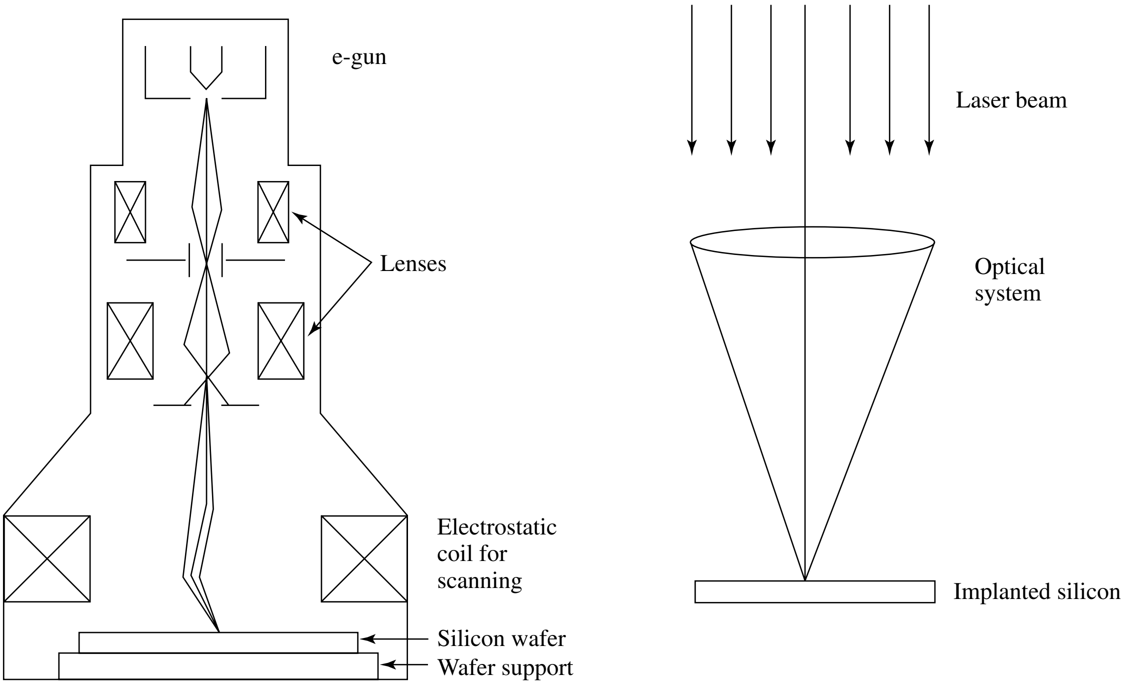


Fig. 9.14 Electron and laser beam annealing

The experimental result of boron and phosphorus implanted at room temperature and annealed for about 30 minutes is shown in Fig. 9.15. The boron and phosphorus are implanted at room temperature, so that no annealing takes place during ion-implantation. It has been found that the recrystallization of amorphous silicon for boron and phosphorus implants are different. In the case of boron implantation, the recrystallization temperature increases with the boron dose. As the boron implanted ions are stopped by electronic stopping; therefore there is less silicon damage. It has been found experimentally that the annealing temperature increases monotonically with the boron ion dose. In the case of phosphorus implantation, recrystallization of phosphorus is highly dependent on the phosphorus dose; where, the annealing temperature increases with the phosphorus dose monotonically and then suddenly falls at around 10^{15} cm^2 of the implant. It is mentioned that the phosphorus implanted ions are stopped by the nuclear stopping mechanism and that does the damage and makes the silicon surface partially amorphous. The amorphousness increases with the ion dose. The process of amorphisation by nuclear stopping is called **primary defects**. If the dose of phosphorus ions exceeds 10^{15} cm^{-2} then the crystalline silicon becomes almost amorphous. During heat treatment, the amorphous silicon starts to recrystallize at the amorphous silicon interface. The reason being that under the

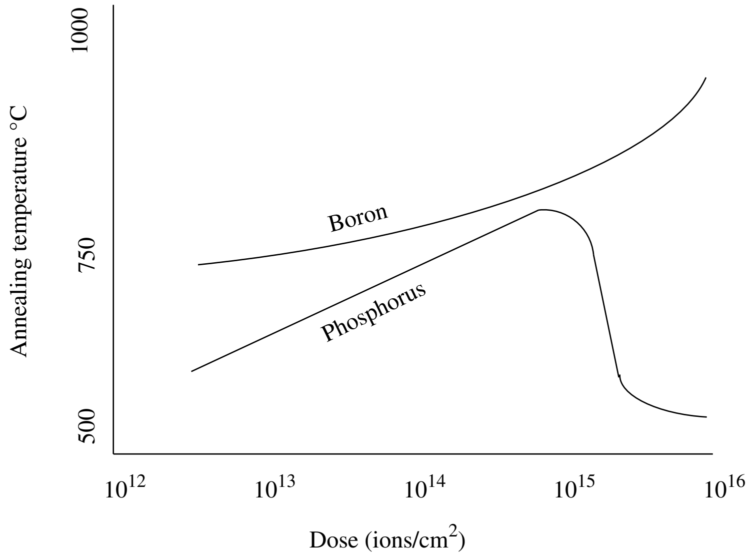


Fig. 9.15 90% activated boron and phosphorus versus annealing temperature

amorphous silicon is crystalline silicon and that initiates crystallisation of the amorphous silicon faster. One can say that crystalline silicon provides the seed to the amorphous silicon for recrystallisation. This process of recrystallisation is called the **solid-phase epitaxy** process. The solid-phase epitaxy is highly dependent on annealing temperature. It has been found that well-amorphised silicon takes very less annealing temperature than the partially amorphised silicon. Hence, a sudden fall in annealing temperature is observed when the implantation dose is more than 10^{15} atoms cm^{-2}.

Summary

The ion-implantation technique is more versatile and advantageous than the thermal diffusion technique in many respects, especially in the recent high density VLSI and ULSI applications. The ion-implantation technique is used for field channel stop, V_T adjustment of transistors, enhancement and depletion of MOS fabrication, gettering, buried conductor layer and buried insulator layer formation, accuracy, reproducibility, doping range, process temperature, doping through thin film, profile control, short-channel doping, pure dopant and molecular dopant, all of which are not possible by thermal diffusion. Furthermore, features such as precise doping control and shallow junction depth that are most desirable for the submicron technology, can only be obtained by ion-implantation. Therefore, ion-implantation is indispensable in the IC fabrication nowadays. Unfortunately, ion-implantation damages the wafer surface and requires an additional heat-treatment process to repair the damage. In addition, ion-implantation uses elaborate and costly equipment and needs constant maintenance; furthermore, ion-implantation is a single wafer process, so the wafer throughput is less.

Ion-implantation has two main process parameters, namely, implantation energy and implantation dose. The implantation energy decides the penetration depth of the ions inside the silicon wafer. Higher the implantation energy, deeper the penetration depth in the silicon. The amount of dopant atoms introduced in the silicon wafer is called the dose. Parameters affecting the dose are neutral atoms, secondary electron emission, and sputter material deposition and/or implantation.

The statistical distribution of the implanted dopant in the silicon is complicated. In principle, the implanted dopant distribution should be Gaussian distribution, but in reality, many lighter boron ions bounce back after collision with the heavier silicon atoms, and the reverse happens with the phosphorus and arsenic ions. In addition, many boron ions travel in between the silicon lattice before they stop and form the tail. These deviations (skewness) from the Gaussian distribution can be calculated by the third and fourth order moments.

The damages done to the crystallinity of the silicon wafer can be undone by annealing the wafer. In the annealing process, the wafer is heated in an inert gas or nitrogen gas or a mixture of nitrogen and hydrogen gases for a certain period of time at elevated temperature. Furthermore, at least 90% of the total implanted ions should occupy substitutional sites of silicon for good electrical characteristic of any transistor. There are three annealing techniques, namely, furnace heating, light heating, and electron beam or laser beam heating techniques. The annealing time and temperature is a function of the implanted ion dose and the type of dopant. Annealing by light heating is called Rapid Thermal Annealing (RTA) and is preferred for VLSI and ULSI fabrication.

References

- J D Plummer, M Deal and P B Griffin; "*Silicon Fundamental Technology: Fundamentals, Practice and Modeling*, Prentice Hall, 2000
- S Wolf and R N Tauber; *Silicon Processing for VLSI Era, Vol 1: Process, Technology*; Lattice Press, Second Edition, 2000
- S M Sze; *VLSI Technology*, Second Edition, McGraw-Hill, 1988

- S K Gandhi; *VLSI Fabrication Principles*, Second Edition, Wiley, 1994
- D Nagchoudhuri; *Principles of Microelectronic Technology*, Wheeler, 1998
- S A Campbell; *The Science and Engineering of Microelectronic Fabrication*, Oxford University Press, 1996

Multiple-Choice Questions

9.1 Generally, the MOS threshold voltage is adjusted in ion-implantation by
(a) increasing the substrate doping (b) decreasing the substrate doping
(c) it cannot be done

9.2 Can dopant ions implanted through a thin film?
(a) Yes (b) No (c) To some extent

9.3 The wafer is tilted during ion implantation at
(a) 30° (b) 70° (c) 7°

9.4 The dopant dose of ion-implantation is measured by
(a) current measurement (b) voltage measurement (c) resistance measurement

9.5 Is a lightly doped drain possible with ion implantation?
(a) Yes (b) No (c) To some extent

9.6 Projected range is a function of
(a) implantation energy (b) ion dose (c) implantation time

9.7 The penetration of boron ions in the silicon wafer is deeper than that of
(a) phosphorus (b) arsenic (c) antimony

9.8 Which dopant has more skewness towards the surface of silicon?
(a) Phosphorus (b) Boron (c) Arsenic

9.9 The ion-implanter beam line is made off from the main implanter axis to avoid the striking of which ions on the silicon wafer?
(a) Dopant ion (b) Neutral atom (c) Electrons

9.10 What is the minimum vacuum level required in the beam line of the ion implanter during implantation?
(a) More than 10^{-6} torr (b) 10^{-4} torr (c) 10^{-2} torr

9.11 At high dose which dopant goes deep into the silicon wafer?
(a) Phosphorus (b) Boron (c) Arsenic

Descriptive Problems

9.1 Why is the threshold voltage of the transistors adjusted? Also explain the threshold voltage mechanism by ion-implantation.

9.2 What is the main reason that the *p*-type transistor is not made prior to the implantation?

9.3 Why is the ion-implantation dose unit written as 10^{11} atoms/cm^2?

9.4 A bare silicon wafer of diameter 6″ (125 mm) is implanted with 1×10^{13} ions/cm^2 boron. Find the implantation time for 1×10^{-6} A, 2.5×10^{-6} A, 5×10^{-6} A, and 7.5×10^{-6} A implantation currents, and also draw the graph between the implantation time and the implantation current.

9.5 A bare silicon wafer of diameter 6″ (125 mm) is implanted with 1×10^{11} ions/cm^2, 1×10^{12} ions/cm^2, 1×10^{13} ions/cm^2, 1×10^{14} ions/cm^2, 1×10^{15} ions/cm^2, and 1×10^{16} ions/cm^2 boron. Find the implantation time and also draw the graph between the implantation time and the implantation current.

9.6 The peak concentration of 5×10^{12} ions/cm^2 and 5×10^{16} ions/cm^2 of phosphorus ions is obtained with 120 keV implanted in a bare silicon wafer of 6″ (125 mm) diameter. Find the projected straggles.

9.7 The projected straggle of 0.05 μm is obtained by the implantation of boron in 6″ (125 mm) diameter bare silicon wafer. Find the peak concentration.

9.8 Find the electronic stopping power of the phosphorus ions with implantation energy of 80 keV, 100 keV, and 120 keV, considering the stopping coefficient of 10^7(eV)$^{1/2}$ per centimetre.

CHAPTER

10 Thin Film Deposition

10.1 INTRODUCTION

In this chapter, the various types of film-deposition techniques are discussed. These films are deposited on the wafer for different applications in MOS fabrication. The films are either grown or deposited on the wafer. These films are used on many occasions during MOS transistor fabrication, for example, silicon dioxide deposition for shallow trench filling for MOS electrical isolation, silicon nitride film deposition for LOCOS, polysilicon deposition as the gate electrode, metal and silicon silicide deposition for electrical contacts, plugging holes for inter-level electrical connections and MOS interconnections, IC wiring, etc. These films are deposited globally on the wafer and later, they are patterned by the lithography and etching processes described in Chapters 6 and 7 respectively.

The deposited films should have the qualities of good adhesion, inertness to other IC materials, ease to deposit and pattern, and good composition; they should be free from contamination, there should be no pinholes or voids (empty spaces) and the stress should be minimum. In addition, metal, alloy, and silicide should have low resistance, whereas, dielectric films should have high density and breakdown.

Apart from the above-mentioned film properties, the deposited film should have uniform thickness over the entire wafer even if the surface of the silicon wafer is modulated with hills and valleys, as illustrated in Fig. 10.1(a). The film that is uniformly deposited without any thickness variation despite modulation is called a **conformal film**. Furthermore, when the wafer is conformed and the wafer surface is made almost plane, then this typical case is called **planarisation**.

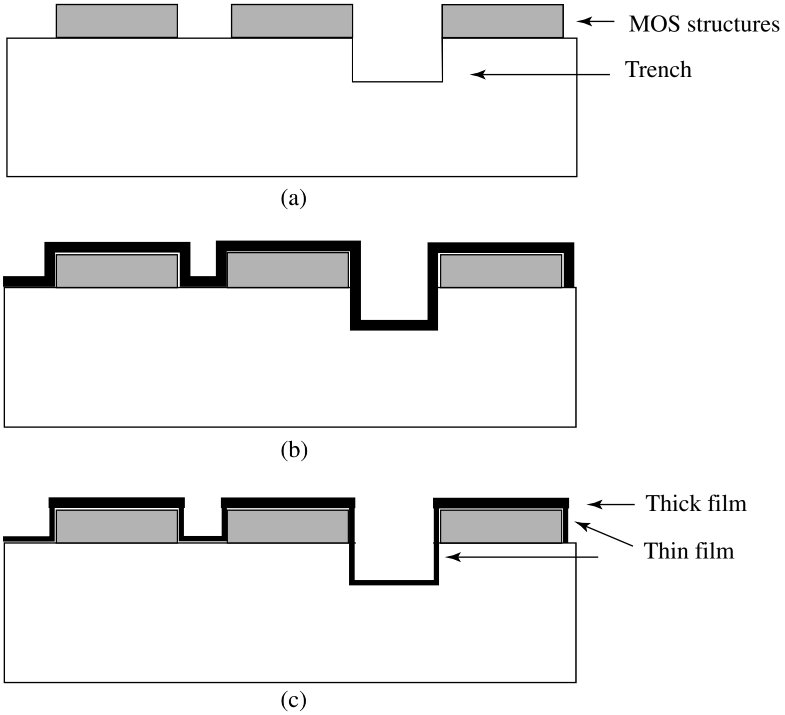


Fig. 10.1 (a) Hill sand valleys (surface modulation) of silicon wafer surface
(b) Conformal film over modulated surface
(c) Nonconformal film over modulated surface

A particular case of a conformal film where all vertical steps are covered with uniform film thickness is called **step coverage**. The conformal, step coverage and non-conformal films are depicted in Fig. 10.1(b) and Fig. 10.1(c).

10.2 FILM-DEPOSITION TECHNIQUES

The film-deposition techniques can be broadly classified into four categories: Spin-on-Glass (SOG) Deposition, Electrochemical Deposition, Physical Vapour Deposition (PVD), and Chemical Vapour Deposition (CVD).

10.2.1 Spin-on-Glass (SOG) Deposition

The film deposition by the Spin–On–Glass (SOG) technique has good conformal and planarisation properties. Apart from the planarisation, the SOG film is used as a dielectric material between the conducting layers for electrical insulation. There are three types of SOG sources available in the market; they are organic siloxane, inorganic silicate, and inorganic and hybrid (organic and inorganic) chemicals. The SOG film deposition process is very simple. The SOG liquid is spun over the wafer as it is done in the case of photoresist coating and then, the wafer is heated to around 400ºC temperature; then, the SOG becomes solid in the form of oxide film. Unfortunately, the residues of the SOG oxide degrade the film quality; therefore, the SOG film is hardly used for critical IC, VLSI and ULSI applications. More details of SOG are described in Chapter 11.

10.2.2 Electrochemical Deposition

The electrochemical film deposition is done by two techniques, namely, **electroplating** and **electroless**. Usually, these techniques are used for metal deposition. The electroplating deposition technique is also called **electrolyte deposition**. In this technique, two electrodes are dipped in the electrolyte solution and DC potential is applied to these electrodes which then produce metal ions. The generated metal ions move towards the cathode electrode where the wafer is attached and the ions get deposited on the wafer.

In the electroless deposition technique, the metal is deposited by the process of chemical reaction. The metal atoms are produced by the chemical reaction and these metal atoms deposit on the wafer. Generally, the electrochemical film deposition technique is not used in IC fabrication mainly because of inferior quality of the film.

10.2.3 Physical Vapour Deposition (PVD)

The Physical Vapour Deposition (PVD) technique is commonly used for metal, metal alloy and dielectric film deposition. The vapour of the metal atoms is produced from the bulk metal by the PVD technique and it gets deposited on the wafer. There are two physical vapour deposition techniques, namely, **thermal evaporation technique** and **sputtering technique** that are used for film deposition. Metal film deposition using these two techniques is called **metal coating** or **metallization**.

Thermal Evaporation Deposition Technique

In the thermal evaporation deposition technique, the bulk metal is heated till the vapour of the atoms is released from the bulk metal into an evacuated chamber called the **vacuum chamber**. As the film is thermally deposited, this deposition process is called **thermal evaporation deposition** and the metal evaporating system is called the **evaporation**

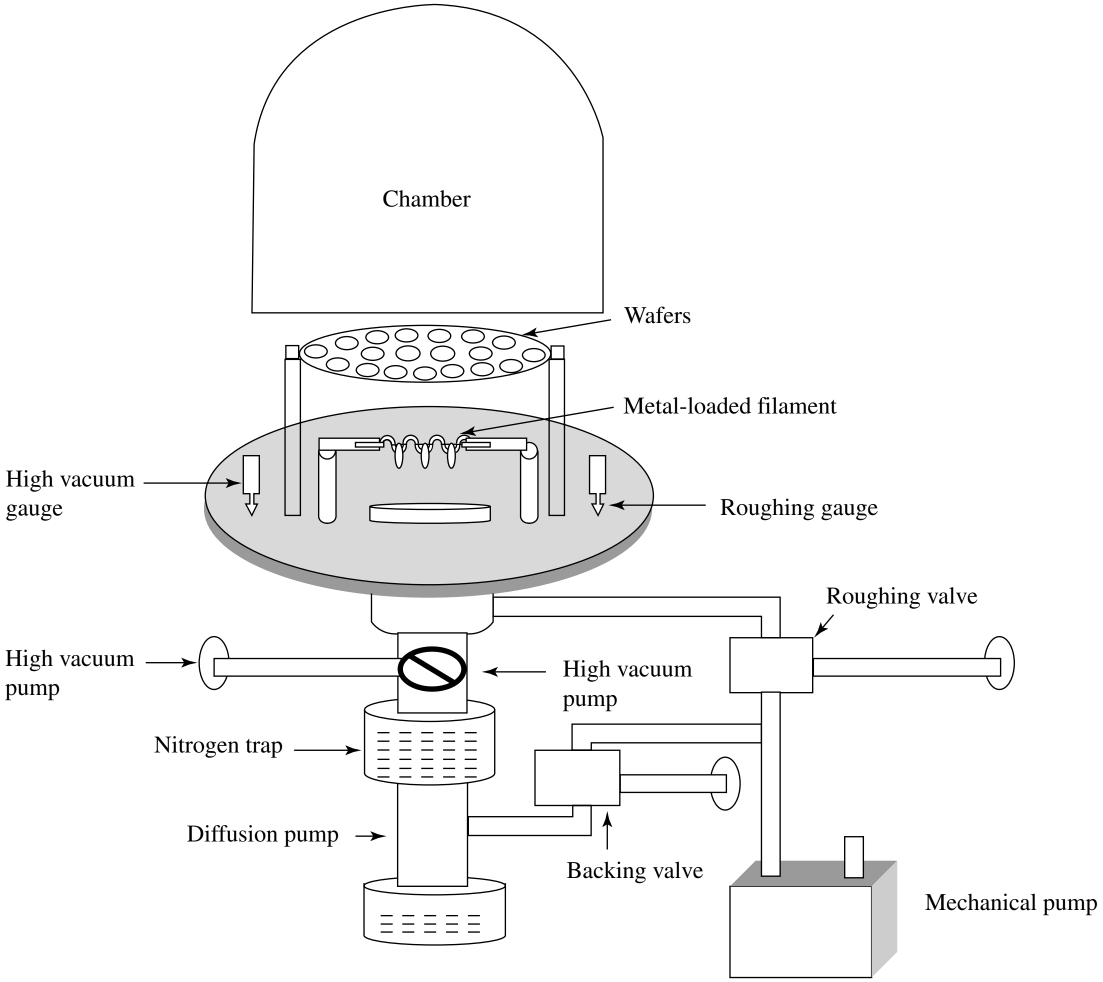


Fig. 10.2 Thermal evaporation unit with lifted chamber

source. A typical schematic of the thermal evaporation equipment is shown in Fig. 10.2. In the metallisation process, initially, the vacuum chamber is evacuated to a pressure in the range of 1 × 10^{-2} Torr and 1 × 10^{-3} Torr using a mechanical vacuum pump. This mechanical pump is also known as the **fore pump**, the **roughing pump**, the **rotary pump** or the **mechanical pump**. Thereafter, the vacuum chamber is further evacuated to a pressure of around 1 × 10^{-6} Torr using a high vacuum pump backed by the mechanical pump. The high vacuum pump may be a **diffusion pump**, a **cryo-pump,** or a **turbomolecular-pump**. The diffusion pump contains silicon oil and when it is heated from the bottom, the silicon oil vaporises and moves upwards. Above the diffusion heater, the diffusion pump is cooled by water from the outside which condenses the silicon oil and converts it into droplets. These droplets fall to the bottom of the diffusion pump and again evaporate. When the silicon droplets come down, they push the air molecules at the bottom of the diffusion pump and then, these air molecules are sucked by the mechanical pump. In the cryo-pump, a refrigeration unit is attached in place of the diffusion pump. The air is condensed and it becomes heavier; then, it comes down and is thrown out by the mechanical pump which is fitted at the bottom of the cryo-pump. In the turbomolecular pump, a large number of blades are attached to a shaft at the centre of the turbomolecular pump. These blades move at a great speed and thrust down the air molecules which are then immediately removed by the mechanical pump in the equipment. The highest vacuum (maximum vacuum) attained by the high vacuum system is called the **base vacuum**. In spite of the high vacuum, the chamber contains traces of water, oxygen, nitrogen and other gas molecules. When the metal is evaporated, the metal atoms collide with the residual air molecules, lose their kinetic energy and change their directions. The average distance travelled by an atom between the collisions is called the "mean free path" and it is expressed as

$$\lambda = \frac{kT}{p\pi\sigma^2\sqrt{2}} \tag{1}$$

where λ is the mean free path, σ is the diameter of the gas molecules, p is the chamber pressure, T is the temperature of the chamber in degrees Kelvin and K is the Boltzmann constant.

It is important to note that the mean free path is inversely proportional to the chamber pressure. If the distance between the wafer and the evaporation source is less than the mean free path then the probability of collision between the metal atoms and the air molecules is zero. If the mean free path is less than the wafer to evaporation source distance, then a fraction of the total metal atoms will collide with the air molecules. Let us say, a fraction of air molecules n_0 does not collide with the metal atoms while travelling a distance d. The ratio of the number of air molecules present in the vacuum chamber to the number of air molecules that do not collide with the metal atoms can be expressed as

$$n/n_0 = exp\,(-d/\lambda) \tag{2}$$

where n is the total number of gas molecules in the chamber.

If $d = \lambda$, only 37% of the n_0 molecules do not undergo collision. This is an important parameter for the sputtering deposition technique.

From the above calculation, it is found that the mean free path at 1 Pa vacuum is 0.7 cm, and for vacuum at 10^{-2} Pa, the mean free path comes out to be 1 m.

Thermal Evaporation Techniques

Generally, there are two types of evaporation sources that are used for metal evaporation. These techniques are: **resistive heating** and **e-beam heating**.

(a) Resistive Heating Evaporation In the resistive heating technique, the bulk of the material is heated using a heating filament in a vacuum chamber. Depending on the material, the heating element and its shape are designed. These heating elements are made of highly pure metals such as tungsten and molybdenum, etc. Different types of heating filaments are shown in Fig. 10.3. In the filament heating technique, the bulk metal pieces are placed on the heating element and then, a high current is passed through the filament. The filament gets heated and that heats the bulk metal. Once the metal is melted, the vapour of the metal is produced and deposited in all directions. Prior to the metal film deposition, the wafers are placed in the vacuum chamber at a convenient place.

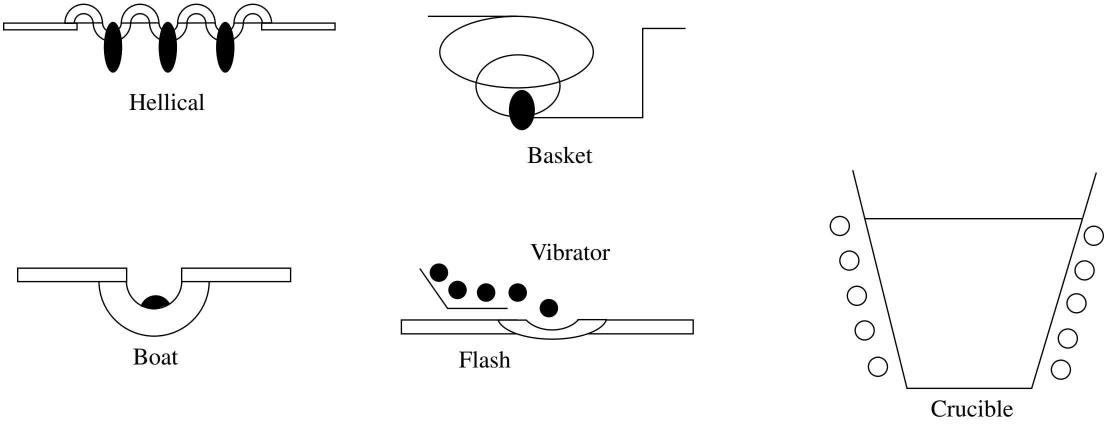


Fig. 10.3 Heating filament techniques

The film deposited by the filament heating evaporation technique suffers from many disadvantages, such as, the limited life of the heating elements, impurities released from the heating element, and limitation of the bulk metal loading onto the heating element that limits the thickness of the deposited film. In addition, the film is deposited at a high vacuum ($>1 \times 10^{-6}$ Torr) and this leads to poor step coverage. This is because the metal atoms move almost straight from the filament to the wafer as shown in Fig. 10.1. To get better film conformity, the wafer is mounted in a domelike jig that rotates in a planetary motion during the film deposition to make sure that the wafer receives metal ions from

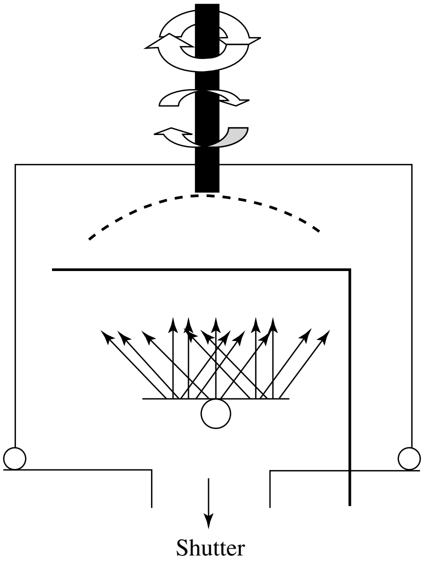


Fig. 10.4 Planetary motion of wafer for uniform coating

all possible angles as shown in Fig. 10.4. The alloy of the film can be deposited by simultaneous heating of different evaporation sources containing different bulk metals; but, maintaining the alloy composition is extremely difficult.

(b) Electron-Beam Evaporation Source In this technique, the metal is heated by an electron-beam (e-beam). The beam of electrons is produced by passing a current through a filament that ejects electrons. These electrons are electrostatically accelerated and focused onto the bulk metal. The e-beam heats the bulk metal and that in turn, produces a cloud of vapour of the metal atoms in the vacuum chamber. The system that produces the e-beam is called an **e-gun** and the container that contains the bulk metal is called the **hearth**. To minimize the radiation exposure of the wafer, the e-beam is diverted by 270º from its original path using the magnetic field as shown in Fig. 10.5.

The electron-beam evaporation technique has many advantages over the resistive heating technique. A large amount of metal can be loaded in the hearth; thus, a thicker film can be deposited. Apart from the thick film, a uniform film can be deposited by increasing the distance between the wafer and the evaporation source. To deposit a metal alloy film, different metals are co-evaporated by different e-beam sources. The main disadvantage of the e-beam evaporation is the X-ray radiation exposure of the wafer, but the radiation damage can be recovered when the wafer is annealed.

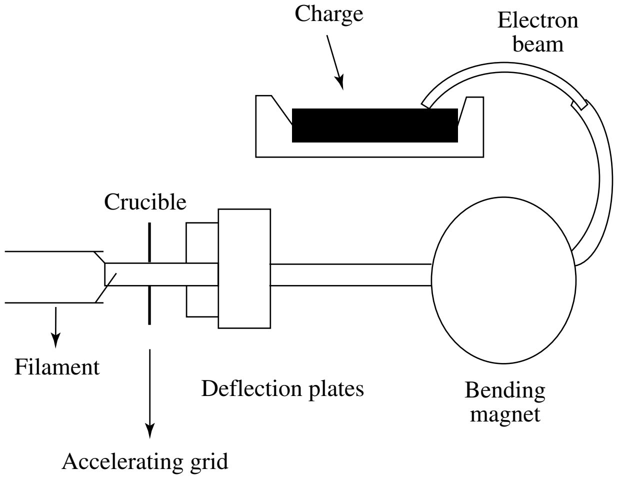


Fig. 10.5 The e-beam heating source

10.2.4 Sputter Deposition Technique

The configuration of the normal sputtering equipment is same as the plasma etching equipment described in Chapter 7; the plasma etching system is illustrated again in Fig. 10.6 for convenience. Two metal electrode plates are kept apart from each other in the vacuum chamber. The material which is to be sputtered is in the form of a plate called the **target** and it is attached to the cathode electrode, whereas, the wafer is attached to the anode electrode. First, the sputtering vacuum chamber is evacuated below 10^{-6} Torr, which is called the **base vacuum**; thereafter, the pressure of the chamber is increased to a value in the range of 10^{-2} Torr and 10^{-3} Torr by introducing argon gas into the chamber. The sputter deposition is carried out at a particular chamber pressure which is called the **sputter pressure**. Once the stable sputtering pressure is obtained, dc or RF power in the range of kilovolts is applied to the electrodes. The power at which the sputter deposition is done is called the **sputtering power**. In the high electric field between the electrodes, the argon gas breaks down electrically, a large number of ions and electrons are generated, and the plasma condition is reached. These ions and electrons are accelerated to a great speed towards the cathode and the anode electrodes respectively due to the high electric field. When the argon ions strike the target with a great velocity, the target metal atoms are knocked out (sputtered) and a vapour of the metal atoms is formed. To enhance the etch rate, the target size is kept significantly smaller with respect to that of the anode (target). This configuration of the electrodes also increases the plasma density and leads

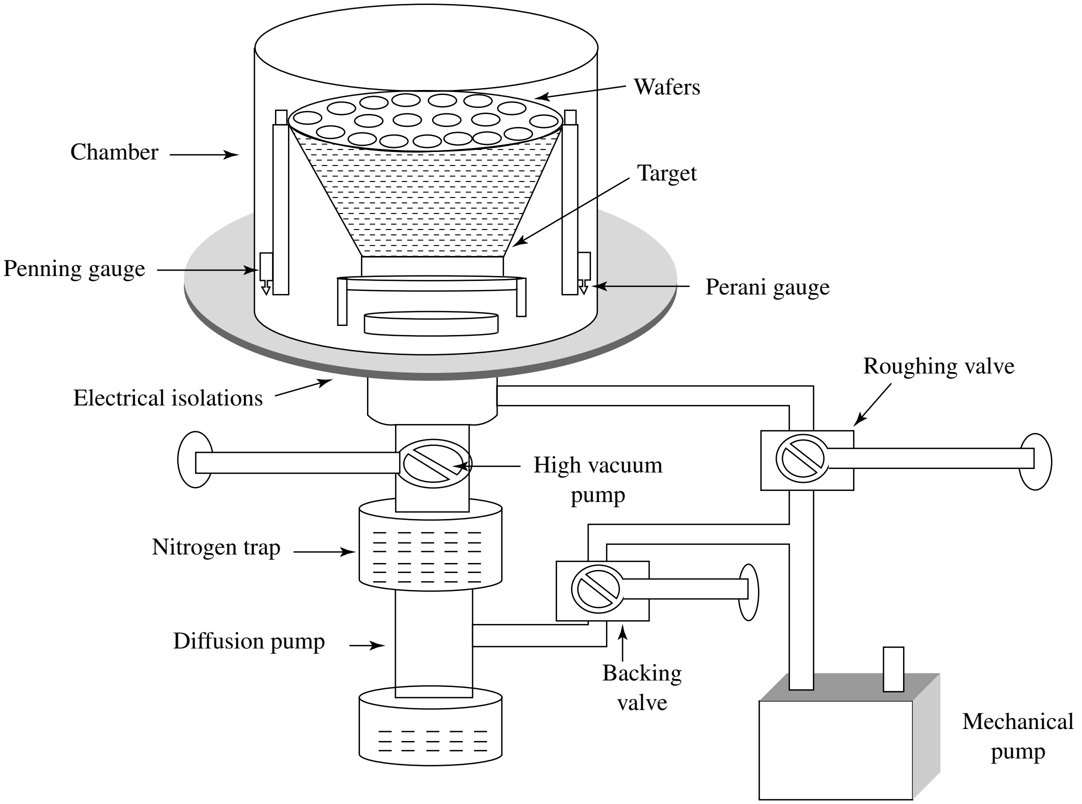


Fig. 10.6 Sputtering unit

to an increase in the ion concentration and that in turn, increases the deposition rate as explained in Chapter 7. On the other hand, the velocity of the electrons, i.e. the kinetic energy is significantly less than that of the argon ions; hence, the wafer etch rate (sputtering) is significantly lower than the rate of film deposition.

The sputter-deposition technique is well suited for the deposition of both metal and alloy metal films. The alloy films can be deposited either by sputtering the alloyed target or by sputtering the individual metal targets simultaneously. The last among the sputter deposition techniques is the **co-sputtering technique**. Furthermore, the dielectric film can be deposited using the RF power.

In sputter deposition, the target is sputtered in the pressure range of 10^{-2} Torr and 10^{-3} Torr in the presence of argon gas. As a large number of argon atoms are in this range of pressure, the mean free path decreases significantly and a large number of argon atoms are ionised. Furthermore, the sputtered metal atoms which are released from the target collide many times with the argon atoms, leading to the scattering of the metal ions in all directions before they reach the wafer and this improves the step coverage. Unfortunately, the quality of the film is inferior as compared to the thermal evaporated film due to the incorporation of gas molecules during film deposition which results in the creation of voids. Furthermore, impurities from the target are also incorporated in the film. In addition, sputter deposition has a disadvantage of radiation effects on the wafer. There are four types of sputtering systems available: **Normal sputtering**, **Magnetron sputtering**, **bias sputtering**, and **collimated sputtering**.

Normal Sputtering

Normal sputtering refers to the simplest form of a sputter deposition system as described above. In the normal sputtering technique, the film deposition rate is very poor and wafer throughput is low. Therefore, this technique is hardly used for film deposition. In addition, the number of gas molecules trapped in the film is high.

Magnetron Sputtering

To increase the film deposition (sputtering) rate, a set of magnets are placed under the target in such a way that the magnetic field is parallel to the target surface as illustrated in Fig. 10.7. In the presence of the magnetic field, the electrons travel in a helical path in the plasma region before they finally strike the anode. The increase in the travelling path of the electrons in the plasma region produces a large number of electrons and ions. The film deposition rate of the magnetron sputter is 10 to 100 times higher than that of the normal sputtering technique. The magnetron sputtering technique is widely used for film deposition in VLSI and ULSI device metallization.

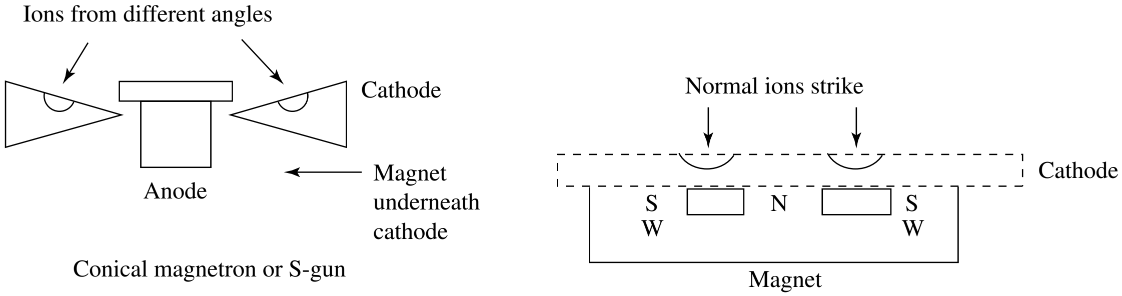

Ions from different angles
Cathode
Anode
Magnet underneath cathode
Conical magnetron or S-gun
Normal ions strike
Cathode
S
W
N
S
W
Magnet

Fig. 10.7 Magnetron sputtering system

Bias Sputtering System (Etch-Deposition Process)

In the bias sputtering technique, the wafer is negatively biased to a voltage of around –50 to 300 volts with respect to the plasma potential and the wafer electrode is electrically isolated from the rest of the equipment. Some of the positive argon ions in the plasma are accelerated towards the negatively charged wafer electrode and they sputter (etch) the wafer or the film on the wafer. Similarly, some of the argon ions are attracted towards the negatively charged target and they sputter the target. If the wafer bias potential is increased substantially with respect to the anode bias, then the etch rate of the wafer will be higher than the target sputter deposition rate. This mode of operation is called the **etching mode** process. Many a time, this mode of operation is required for cleaning the wafer surface prior to the film deposition. This process of wafer cleaning is called **pre-cleaning** or **ion cleaning**. On the other hand, if the wafer bias potential is significantly less than the target potential then the deposition rate of the sputtered material from the target film will be much higher than the etching rate of the wafer. This mode of sputtering is called the **sputter deposition mode**. The film deposition mode which lies between these two etching modes is called the **etch-deposition mode**. Generally, in IC fabrication, a low etch rate and a high deposition rate are preferred for better step coverage (achieved by etch-deposition mode).

Collimated Sputtering System

To combat the nonconformities and to avoid the formation of voids, the collimated sputtering technique or the collimated bias sputter deposition technique is preferred. In the collimated sputter deposition technique, a thin plate containing small holes is placed between the target and the wafer. The holes can be either circular or hexagonal in shape. The sputtered atoms pass almost perpendicularly through the holes (in a collimated fashion) and deposit uniformly over the wafer as shown in Fig. 10.8. Unfortunately, up to 90% of the material is wasted in this technique.

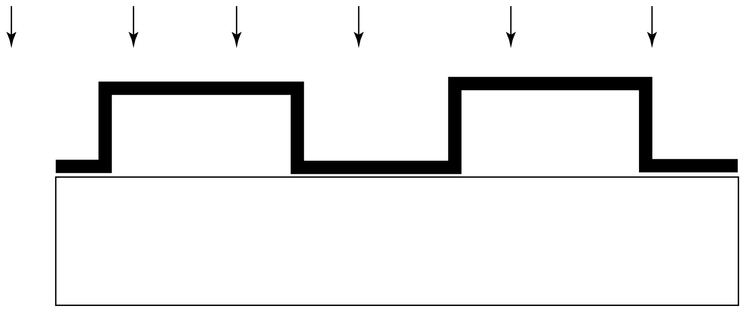


Fig. 10.8 Collimated sputter deposition

Collimated Bias Sputter Deposition Technique In the collimated bias sputter deposition technique, a negatively biased grid is placed between the wafer and the target. The metal ions which sputtered from the target travel in a straight path due to the potential of the negatively biased grid and reach the wafer in a perpendicular fashion; as a result, a uniform film is obtained on the wafer.

10.2.5 Chemical Vapour Deposition (CVD) Technique

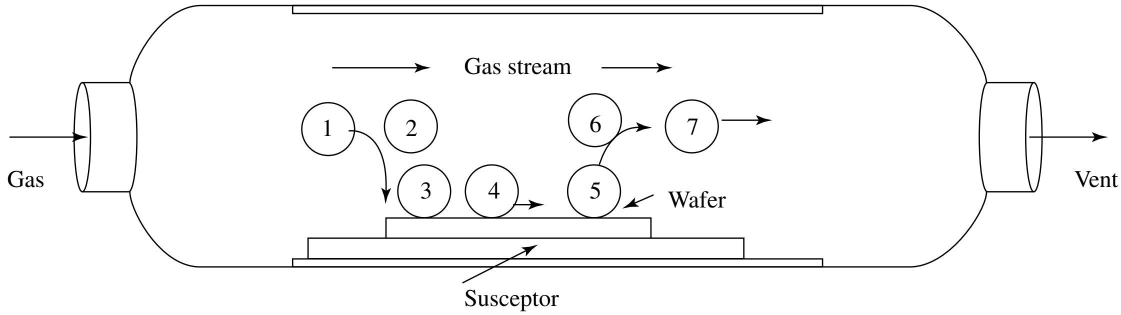
In the Chemical Vapour Deposition (CVD) process, the film is deposited by heating a gas or a mixture of gases called the **precursor(s)** in a furnace. At high temperature, the precursor gas(es) create(s) a vapour of atoms either by chemical reactions or by dissociation processes and then, the vapour of atoms deposits on the wafer. For this reason, this technique of film deposition is called **chemical vapour deposition (CVD)**. As the chemical reaction(s) take(s) place in the furnace, the CVD furnace is called the **CVD reactor** or simply the **reactor**. Several types of metals, metal alloys, nitrides and dielectrics can be deposited by the CVD technique. The CVD technique has an edge over other deposition methods in terms of conformity, high wafer output, and the capability for the deposition of several films. In addition, a pure or doped film can also be deposited on the wafer.

There are two mechanisms by which a film can be deposited. In the first mechanism, the precursors react themselves, produce the vapour atoms and deposit onto the wafer surface. In the second mechanism, the precursor diffuses the silicon wafer and then, reacts with the silicon atoms resulting in the formation of the film. Usually, the deposited atoms are loosely attached to (absorbed by) the wafer and then, they migrate on the wafer surface at high temperature as shown in Fig. 10.9. The movement (migration) of atoms on the wafer surface is called **surface migration**. The surface migration of atoms plays a vital role in making a conformal film deposition.

There are three CVD techniques used for film deposition in IC fabrication, namely, Atmospheric Pressure Chemical Vapour Deposition (APCVD), Low Pressure Chemical Vapour Deposition (LPCVD) and Plasma Enhanced Chemical Vapour Deposition (PECVD).

Atmospheric Pressure Chemical Vapour Deposition (APCVD)

In the APCVD technique, the film is deposited at atmospheric pressure as shown in Fig. 10.9. APCVD technique is the oldest in the CVD family; thus, APCVD is also called the **conventional CVD**. The high deposition rate, low equipment requirements, and simple deposition process are the most attractive features of the APCVD technique.



1. Gas stream 2. Diffusion of precursor through boundary layer towards wafer 3. Adsorption of precursor on the wafer 4. Chemical decomposition or reaction of precursor and attached of specie to proper sites 5. Deadsorption of by-product 6. Diffusion of byproduct into gas stream 7. Vent of by-product out of rector

Fig. 10.9 CVD system and growth mechanism

In the APCVD process, the wafers are placed on a heated graphite plate called the **susceptor**, which is pushed inside the heated horizontal APCVD **reactor** with the help of a conveyer belt. The precursor gases are introduced from one end of the reactor and the unused gas and the by-products are vented out from the other end of the reactor. In the APCVD process, only the susceptor is heated, not the whole reactor; therefore, APCVD reactor is also called the **cold wall reactor**. The film can be doped with gaseous dopants such as Arsenic (AsH_3), Phosphorous (PH_3) and Diborane (B_2H_6) during the film deposition.

(a) Deposition Kinetics When the gas (precursor) is passed through the reactor, it experiences friction at the edges of the wafer and the reactor wall; this retards the gas flow near the reactor wall. The retardation of the gas layer is called the **boundary layer**. This boundary layer increases with the susceptor length as shown in Fig. 10.10. In order to minimise the boundary layer, the conveyer belt is tilted as illustrated in the same figure.

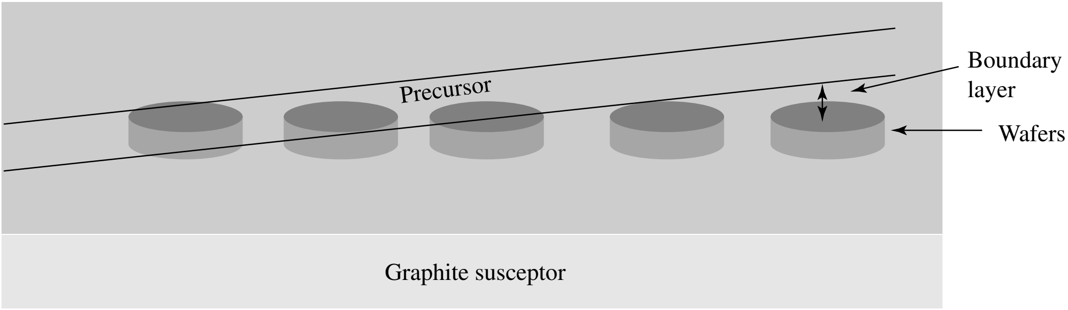


Fig. 10.10 Boundary layer in APCVD reactor

(b) Kinetics of APCVD Let us say that the flux F_1 (molecules cm^{-3} s^{-1}) gas diffuses from the centre of the main gas stream to the reactor wall (across the boundary layer); thus, the flux F_1 can be written as

$$F_1 = h_g (C_g - C_s) \tag{3}$$

where h_g is the mass transport coefficient, C_g is the gas concentration away from the wafer and C_s is the gas concentration at the wafer surface.

Let us consider that flux F_2 is consumed after the reaction with the wafer then the flux F_2 can be written as

$$F_2 = K_s C_s \tag{4}$$

where K_s is the chemical surface reaction rate (cm s^{-1}).

At equilibrium state,

$$F_1 = F_2$$

From the equations (3) and (4), the surface gas concentration can be expressed as

$$C_s = C_g (1 + K_s / h_g)^{-1} \tag{5}$$

If v is the rate of deposition (in cm^{-1}) and N is the number of atoms incorporated per unit volume in the film then, the deposition rate can be written as

$$\begin{aligned} v &= F/N \\ &= (K_s\, h_g / K_s + h_g) \cdot (C_g / N) \end{aligned} \tag{6}$$

To calculate the unknown factor C_g, the mole fraction concept is used. The mole fraction of a particular gas is defined as

$$\text{Mole fraction, } Y = C_g / C_t = P_g / P_\mathrm{t} \tag{7}$$

where P_g is the partial pressure of a particular gas and P_t is the total (sum) of the partial pressure of individual gases i.e. ($P_{g1} + P_{g2} + P_{g3}$ ...).

$$Y = P_g/(P_{g1} + P_{g2} + P_{g3} \ldots)$$

Hence, we get:

$$v = K_s\, h_g / (K_s + h_g) \cdot (C_t / N) \cdot Y \tag{8}$$

From the above equation, one can infer that the deposition rate depends on the reaction rate constant (K_s) and the gas mass transport (h_g). Hence, the film can be deposited in two different cases based on these parameters.

Case 1: Gas Phase Diffusion Controlled Deposition Let us consider that the reaction rate (K_s) is higher than the mass transport coefficient (h_g), i.e. ($K_s >> h_g$)

Then, the deposition rate can be expressed as

$$v = (C_t/N)\, h_g\, Y \tag{9}$$

The above equation is valid only when the gas reacts with the silicon as soon as it reaches the wafer surface. Hence, the deposition rate is limited by the gas arrival (*hg*) at the wafer surface. This case of deposition is called the **gas phase diffusion controlled case** and the film deposition is linear with the h_g parameter.

Case 2: Surface Reaction Rate Controlled Deposition Let us consider that the mass transport coefficient (h_g) is higher than the reaction rate, i.e. ($h_g >> K_s$)

Then, the deposition rate can be expressed as

$$v = (C_t/N)\, K_s\, Y$$

In this case of the film deposition process, the gas is available in abundance at the wafer surface and the deposition rate depends only on the reaction rate constant (K_s). Therefore, this mode of the film deposition is called the **surface reaction controlled case**. The reaction rate (K_s) is a function of temperature and it decreases exponentially with temperature as

$$K_s = K_0 \exp(-E_a/KT) \tag{10}$$

The surface reaction controlled and gas phase diffusion controlled cases and their resultant film deposition rates are shown in Fig. 10.11.

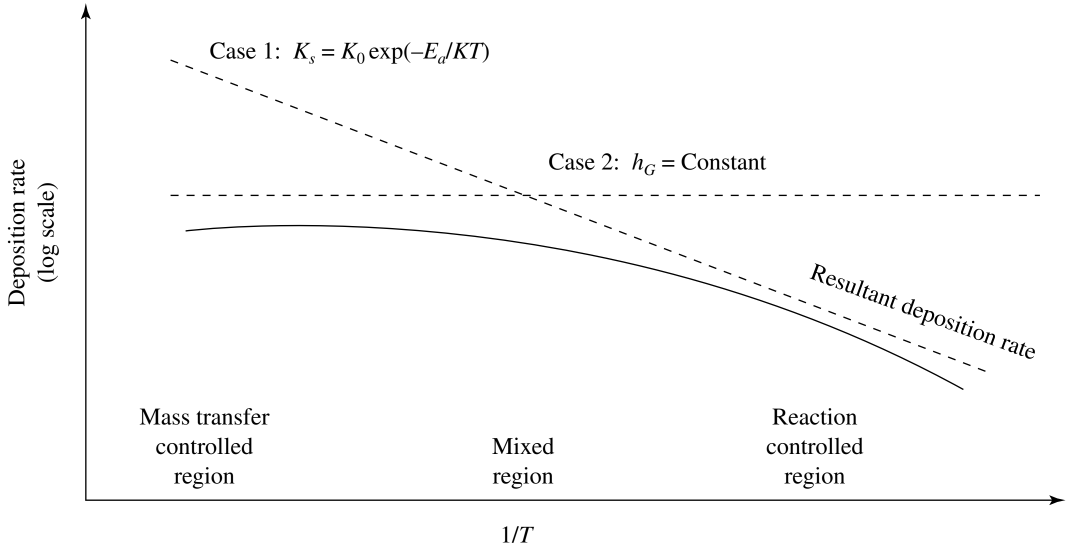


Fig. 10.11 Rate of deposition in two cases and their resultant

The APCVD technique is not commonly used in IC fabrication due to the limitations of film conformity, particulate formation and gas consumption in a large volume. Apart from these disadvantages, the APCVD technique has two more disadvantages; one is called out-diffusion and the other is called auto-diffusion. In out-diffusion, the high concentration dopant diffuses towards the low concentration during the deposition process and in auto-diffusion, the dopant comes out from the susceptor and other parts of the reactor; then, these dopant atoms diffuse into the wafer. Out-diffusion and auto-diffusion are unwanted because they alter the dopant profile in the wafer significantly.

Example 10.1

Silicon dioxide is deposited on a silicon wafer by APCVD technique (at 760 Torr). Find out the deposition rate when the deposition parameters are fixed as: the mass transport coefficient is 1.5 cm per second, reaction rate constant is 15 cm per second, partial pressure of incorporating species is 1 Torr, total gas phase concentration is 1×10^{18} cm^{-3} and density of the depositing film is 1×10^{22} cm^{-3}.

Answer

h_g = 1.5 cm per second,
K_s = 10 cm per second,
$C_r = 1 \times 10^{18}$ cm^{-3}, and
$N = 1 \times 10^{22}$ cm^{-3}.

$$Y = \frac{C_g}{C_t} = \frac{1}{760}$$

Using the formula for film deposition rate,

$$v = \frac{k_s h_g}{k_s + h_g}\frac{C_t}{N}Y$$

$$v = \frac{5.0 \times 1.5}{10.0 + 1.5} \times \frac{1 \times 1 \times 10^{18}}{1 \times 1 \times 10^{22}} \times \frac{1}{760}$$

$$v = \frac{6.5}{11.5} \times \frac{1}{760} \times 10^{-4}$$

$$v = 0.565 \times \frac{1}{760} \times 10^{-4}$$

$$v = 0.74\ \mu/\text{s}$$

Low-Pressure Chemical Vapour Deposition (LPCVD)

In the LPCVD technique, the reactor is evacuated from one end of the reactor using the rotary vacuum pump. This end of the reactor is called the **rear end** and the wafers are loaded from the other end of the reactor, which is called the **front end** as shown in Fig. 10.12. The precursor is introduced from the rear end of the reactor. The deposition starts from the rear end; as a consequence, the concentration of the precursor decreases and the deposition rate decreases as it moves towards the front end of the reactor. In order to keep the film deposition rate constant along the reactor length, the temperature of the reactor is increased by 25ºC–30ºC at the front end of the reactor. The increase in temperature ensures a constant deposition rate along the reactor. In the LPCVD process, the film deposition is carried out in the range of 300ºC to 900ºC at a pressure of 5 to 100 Torr. As the reactor is heated, the LPCVD rector is also called the **hot wall reactor**.

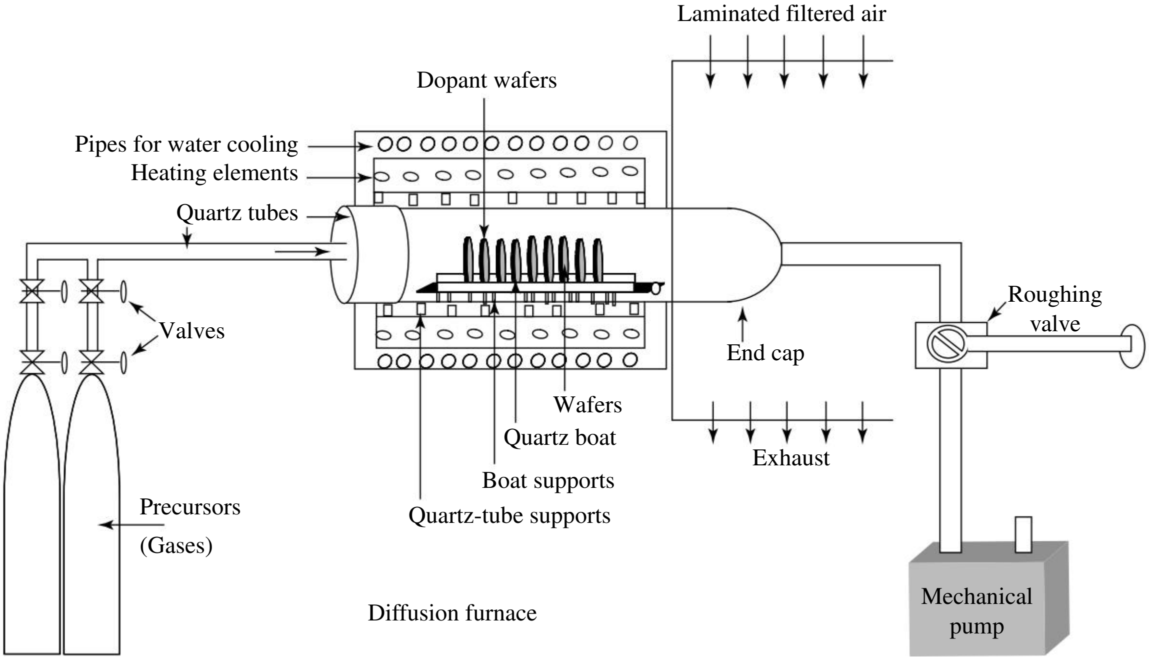


Fig. 10.12 Typical LPCVD system

In the LPCVD deposition process, the auto-doping is insignificant. This is because the out-diffusion dopant as well as the reaction by-products are quickly exhausted out of the reactor by the vacuum system. In addition, a good conformal film is obtained due to the fast movement of the precursor in vacuum. Unfortunately, in the LPCVD technique, the surface migration is much less in comparison to the APCVD technique. This leads to the possibility of creation of voids in the film.

Plasma Enhanced Chemical Vapour Deposition (PECVD)

In the Plasma Enhanced Chemical Vapour Deposition (PECVD) technique, the film is deposited in plasma at a low temperature. This film deposition technique has the flexibility to deposit dielectric films over a metal film; especially, an aluminium film. By this technique, an excellent conformal film is obtained without dopant redistribution. Generally, the film is deposited in the temperature range of 200ºC to 350ºC, at a pressure that ranges between 50 mTorr and 5 Torr using an RF (13.57 MHz) source. The PECVD deposition system is shown in Fig. 10.13. In the PECVD technique, the atoms (or specie) of the film are produced in the plasma and they deposit on the wafer. For example, the precursor tetraethyl ortho silane (TEOS) dissociates in the plasma and SiO_2 specie are produced and they deposit on the wafer. Unfortunately, in the PECVD technique, the physical damage to the wafer is significant as compared to the other deposition techniques, especially in high RF power deposition. In addition, the gas molecules and the by-products are trapped inside the film, leading to the creation of voids in the film. Furthermore, a slight variation in the deposition parameters changes the quality of the film significantly.

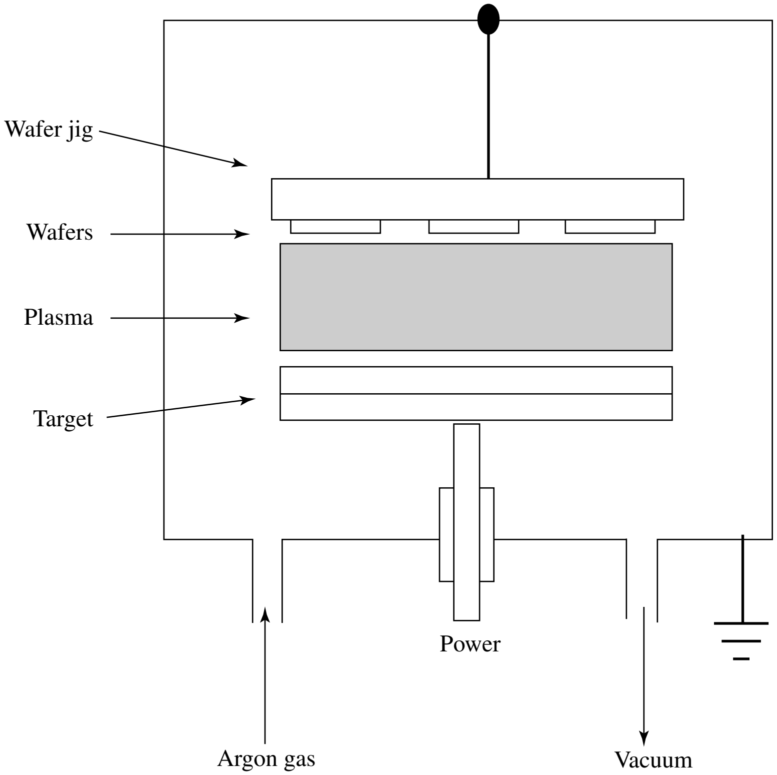


Fig. 10.13 Typical plasma-deposition system

10.3 METAL WIRINGS AND CONTACTS

Films of aluminium (Al), copper (Cu), titanium (Ti), tungsten (W) and metal alloys are commonly used for the electrical wirings and the electrical contacts in IC fabrication. As there are millions of MOS transistors that are electrically wired to realise the IC, the resistances of the contacts as well as the metal wire should be as low as possible; otherwise, it will lower the IC signal processing speed.

10.3.1 Contact Resistance and Wiring Resistance

Metal as well as metal alloy films are used for electrical wiring and electrical contacts. These electrical contacts and interconnecting wires must satisfy the basic electrical requirements, namely, those of an **Ohmic contact** and **interconnecting wires resistivity** or simply **wiring resistivity**.

10.3.2 Ohmic Contacts

There are many types of contacts used in an IC, such as metal to metal, metal to silicon, metal to metal alloy, and metal to silicide, etc. In all these cases, the contacts should be stable with least possible resistance. The contact that obeys Ohm's law regardless of its voltage polarity is called as an **Ohmic contact**, and the resistance that is present due to the contact is called the **contact resistance** or the **Ohmic resistance** as shown in Fig. 10.14. In order to obtain metal-silicon Ohmic contacts, the silicon should be doped more than

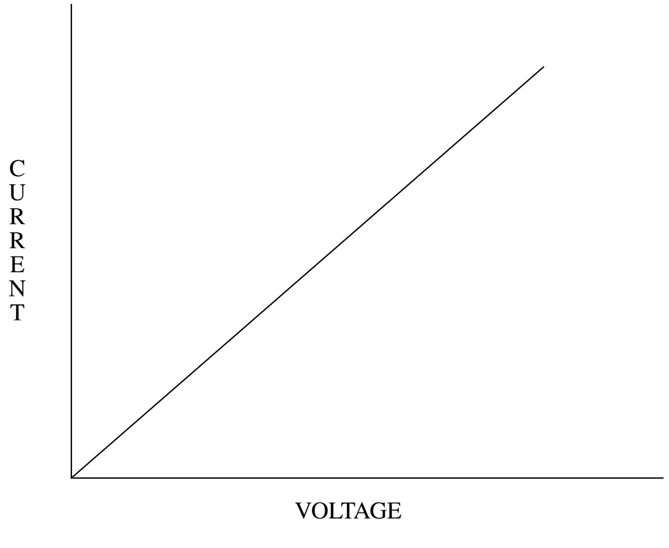


Fig. 10.14 Ideal Ohmic contact

5 × 10^{16} cm^{-3}, so that the depletion width between the metal and the silicon becomes extremely narrow. If the doping concentration is less than 5 × 10^{16} cm^{-3}, then the contact may act as a rectifier (Schottky barrier diode). The contact is characterized by "specific contact resistance" as given below:

$$R_c = \left[\frac{dV}{dJ}\right]_{v=0} \Omega - \text{cm}^2 \tag{11}$$

where R_c is the specific contact resistance, J is the current density and V is the voltage across the metal-silicon contact.

10.3.3 Interconnecting Wiring Resistance

It is essential that the metal wiring should have negligible resistance. The resistance of the film is different from the bulk metal resistance, and is defined by "sheet resistivity". If the film width (*W*), length (*a*), and thickness (*d*) are large compared to the probing distance, then the resistivity ρ of the film can be written as

$$\rho = \frac{V}{I} W\, CF \quad \Omega\text{-cm} \tag{12}$$

where I is the current passing through the film from the outer probes of the four probe measurement equipment, V is the voltage measured by the inner probes of the four probe, and *CF* is the correction factor. If the width (*W*) of the film is very large compared to the probing distance, then *CF* becomes $\pi/In2$ = 4.54. The resistivity measurement technique and the schematic of the four probes are given in Chapter 8 (Diffusion).

10.4 METAL FILM DEPOSITION TECHNIQUES

10.4.1 Aluminium Deposition

The resistivity of an aluminium film is low as compared to that of other materials, except a copper film. The resistivity of an aluminium film is around 2.7 $\mu\Omega$-cm and the aluminium-silicon alloy (contact) resistance is around 3.5 $\mu\Omega$-cm. Apart from the low resistivity, the aluminium film has many advantages, such as it adheres exceptionally well with many IC materials, nor does it react with them, and it does not oxidise very fast. Furthermore, the processing of an aluminium film which includes deposition and etching is relatively easier. Generally, the aluminium film is deposited by the physical vapour deposition (PVD) technique, where both the thermal and the sputter deposition techniques are commonly used. The aluminium-silicon contact exhibits a eutectic nature and it melts around 577ºC, as against their individual melting points of 660ºC (aluminium) and 1420ºC (silicon). Thus, the properties of the aluminium material made it the first choice for IC contacts and interconnections (wiring). Now, copper is replacing aluminium in VLSI and ULSI applications. Unfortunately, spiking and electromigration are the two major disadvantages of the aluminium film.

Spiking

When aluminium is deposited on the contact windows, it migrates into the source and the drain, reaches their junctions and shorts them. This becomes serious in the submicron geometry of VLSI and ULSI, where the junction depth is very shallow. It is also seen that the silicon atoms move into the aluminium film and the aluminium atoms occupy the vacancies created by the silicon atoms. This process continues till it reaches the solid solubility state. But prior to reaching the solid solubility state, the aluminium atoms cross the junction depth in many places and short the junction as shown in Fig. 10.15. This phenomenon of aluminium diffusion is called **spiking**. The spiking phenomenon is significant where the silicon areas are defective.

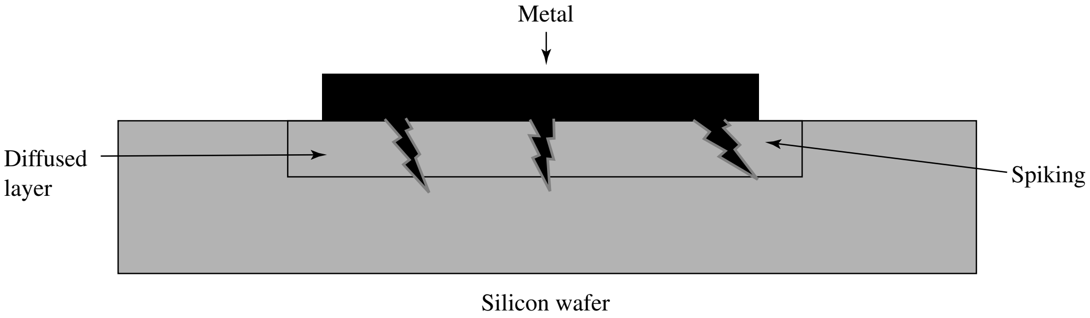


Fig. 10.15 Typical plasma-deposition system

Spiking can be checked by two schemes. In the first scheme, the solid solubility state of aluminium can be achieved by the addition of 1–3% of silicon to the aluminium metal during film deposition or by depositing a thin polysilicon film prior to the aluminium deposition. The polysilicon film provides enough silicon atoms to the aluminium film to reach the solid solubility state; hence, it acts as a barrier layer for spiking. In the second scheme, a tin nitride (TiN) film is also used as a barrier layer that checks well the migration of the aluminium atoms inside the silicon wafer. This type of diffusion barrier is called a **diffusion barrier layer** or in short, a **barrier layer**. Out of these schemes, the TiN barrier layer scheme is most reliable, effective, and it reduces the contact resistance.

Electromigration

During IC fabrication, the surface of the silicon gets modulated in many places as shown in Fig. 10.16. To deposit a uniform film over the modulated surface is difficult, especially at the vertical step, where the thickness of the film is always thin compared to the flat surface of the wafer as shown in the Fig. 10.1. The variation in the thickness of the metal line may also occur due to defective mask, lithography and etching processing errors. When the current is passed through the aluminium wire, the aluminium film gets heated at thin film locations. This leads to localised heating of the film and makes the aluminium viscous and ionised. The generated ions move under the influence of the electric field and accumulate at some other cold places. This phenomenon of the migration of the ions under the influence of the electric field is called **electromigration** and is illustrated in Fig. 10.16. The process of electromigration continues till the aluminium wire gets disconnected at the heated places. The accumulation of the aluminium atoms is called **piling** and the place where the aluminium film gets disconnected is called a **void**.

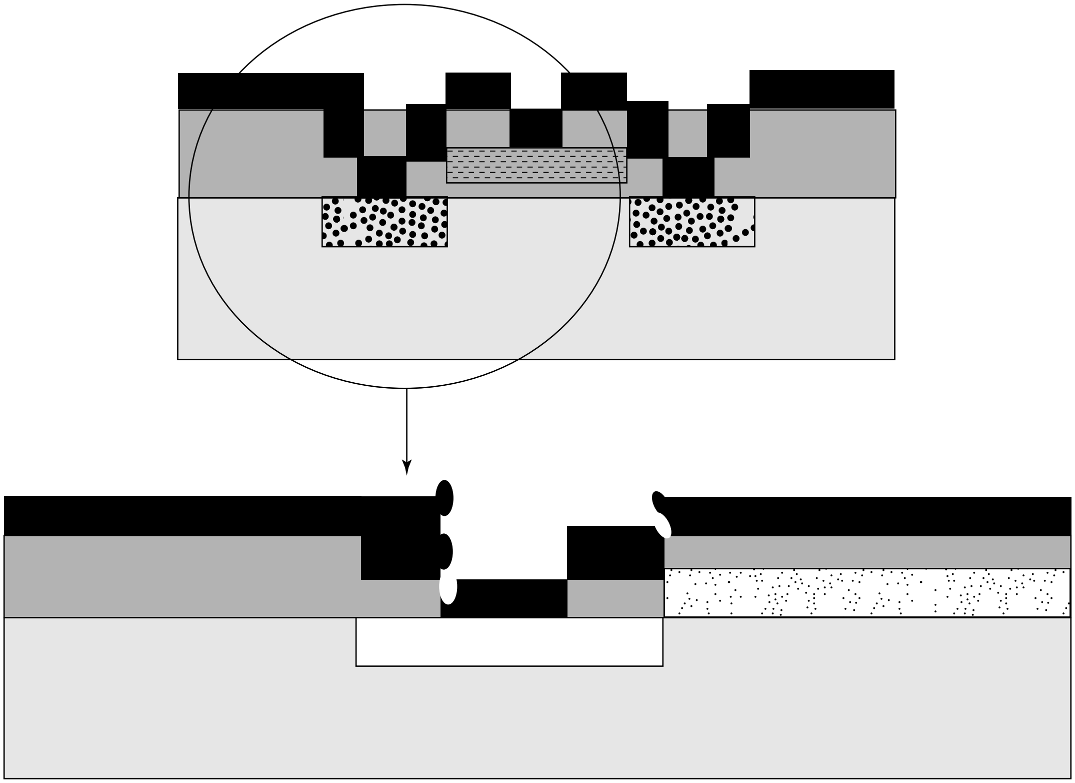


Fig. 10.16 Electromigration of metal

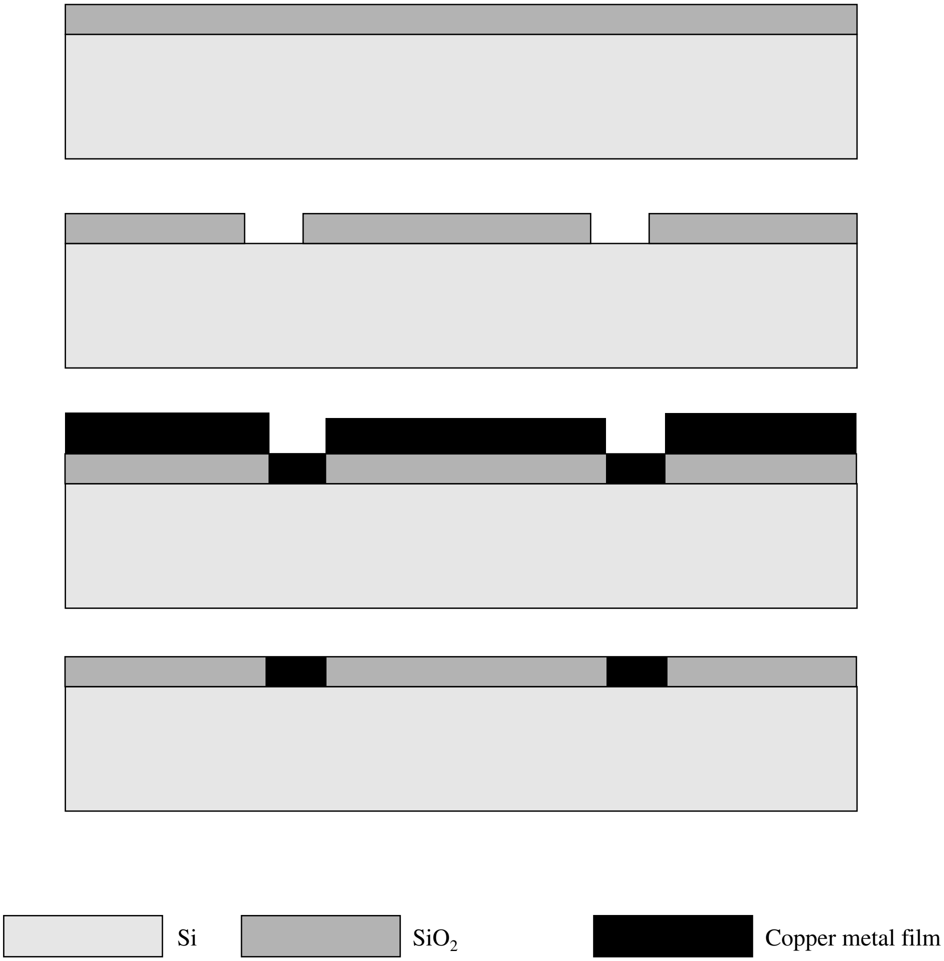


Fig. 10.17 Damascene process for copper deposition

The aluminium electromigration can be checked by incorporating 0.5 to 4 percentage weight of copper during the aluminium evaporation. The electromigration of aluminium can also be checked by sandwiching the aluminium wire between two dielectric layers.

10.4.2 Copper Deposition

The copper film has the advantage of lower resistivity, but it suffers from fast oxidation, adhesion and etching problems. Presently, copper wiring is done by the **damascene process**. The damascene process is a combination of the lift-off lithography and the chemical-mechanical-polish processes. In this technique, the first a layer of dielectric film is deposited and then, PR is coated. Thereafter, PR patterns are made and the unprotected dielectric is removed. Once the dielectric patterns are made, the entire wafer is coated with a copper film. The copper film deposits not only on the bare silicon but also over the patterned dielectric film. Thereafter, the second layer of dielectric is deposited on top of the dielectric and the copper film. Then, the second layer of the dielectric film is completely removed till the dielectric patterns first made are reached by the chemical mechanical polishing as shown in Fig. 10.17.

Copper Deposition Techniques

The copper film is deposited by the CVD process and the copper plating technique. These two techniques are described below.

(a) CVD Technique A copper film with good conformity and excellent via filling can be achieved by the CVD technique. The process of film deposition by the CVD technique is same as discussed previously. For the copper film deposition, Bi-hexafluoroacetylacetonate-Cu[11], metal-organic compounds and β-diketonates are used as precursors.

(b) Electrochemical Deposition Technique The low temperature and low resistivity copper film can be obtained by the electrolyte plating technique and the electroless technique. By the electrochemical deposition technique, which has a better filling, good conformity and a thicker film can be obtained. Unfortunately, the copper films deposited by these techniques are not free from contamination.

(c) Electrolyte Plating Copper Deposition In the electrolyte plating process, two electrodes are dipped in the electrolyte solution containing copper and dc voltage is applied to these two electrodes. In the presence of the electric field, an electrochemical reaction takes place and copper ions are formed. These copper ions move towards the negatively charged electrode to which the wafer is attached. Generally, a thin conducting metal film is deposited over the wafer prior to the copper film deposition for better electrical conduction. The electrochemical reaction can be written as

$$\mathrm{Cu}^{2+} + 2\mathrm{e}^{-} \rightarrow \mathrm{Cu}$$

(d) Electroless Copper Deposition The copper film is deposited by the electroless technique using the formaldehyde (HCOH) and CuEDTA solutions. The overall chemical reaction can be written as

$$(\mathrm{Cu\ EDTA})^{2-} + 2\mathrm{HCOH} + 4\mathrm{OH}^{-} \rightarrow \mathrm{Cu} + 4\mathrm{HCOO}^{-} + 2\mathrm{H_2O} + \mathrm{H_2} + \mathrm{EDTA}^{4-}$$

10.4.3 Tungsten Deposition

A tungsten film can be deposited by the sputtering technique, but tungsten is commonly deposited by the LPCVD technique. The overall chemical reaction is written as follows:

$$WF_6 + 3H_2 \rightarrow W + 6HF, \text{ or}$$
$$2WF_6 + 3SiH_4 \rightarrow W + 3SiF_4 + 6H_2$$

Tungsten (W) is not used for IC wiring (interconnections) due to its high resistivity, but it is used for via (holes in the dielectric) filling in the form of **tungsten plugs**. Apart from via plugging, tungsten is also used as tungsten silicide (WSi_2) for a diffusion barrier. Tungsten silicide is made by depositing a tungsten film on the silicon wafer which reacts with the silicon atoms and forms tungsten silicide.

10.4.4 Titanium Deposition

Good step coverage and better conformity of the film using titanium is obtained by the sputtering technique, where pure Ti is used as a sputtering target. Ti film can also be deposited by the LPCVD technique using titanium pentachloride mixed with hydrogen at 600ºC. The reaction can be expressed as

$$2TiCl_5 + 5\,H_2 \rightarrow 2Ti + 10\,HCl$$

The titanium (Ti) film is extensively used as a diffusion barrier for other materials. The deposited Ti film reacts with the silicon and forms titanium silicide.

10.5 METAL ALLOY DEPOSITION

Metal alloy films such as Titanium–Palladium–Gold (Ti-Pd-Au), Titanium–Platinum–Gold (Ti-Pt-Au), Titanium–Platinum (Ti-Pt), and Titanium–Tungsten (Ti–W) are required in IC fabrication, especially as diffusion barriers. Out of these alloys films, the Ti-W film is proven to be more useful.

10.5.1 Titanium–Tungsten (Ti-W) Alloy Deposition

The titanium–tungsten alloy is used as a metal barrier and an antireflection coating to avoid the reflection of light from the wafer during the lithography process. The Ti-W film is deposited by the sputter technique using a Ti-W alloy target. To improve the barrier quality, nitrogen gas is added sometimes.

10.6 NITRIDE DEPOSITION

The silicon nitride (Si_3N_4) and the titanium nitride (TiN) films play an important role in IC fabrication. For instance, the silicon nitride film is used for the LOCOS process, encapsulation, passivation and as a dielectric; whereas, the titanium nitride layer is used as a metal diffusion barrier and also for antireflection in the lithography process.

10.6.1 Silicon Nitride (SiN) Deposition

The silicon nitride film is deposited by the LPCVD technique and plasma assisted deposition technique. Unfortunately, when the silicon nitride film is deposited directly on the silicon wafer, it produces a large stress due to the mismatch between their lattices. In order to reduce the stress, a silicon dioxide film is grown prior to the silicon nitride film deposition.

It is possible to achieve good uniformity, step coverage, conformity, high throughput and process convenience for the silicon nitride film using the LPCVD technique. The silicon nitride film is deposited using a mixture of ammonia (NH_3) and dichlorosilane (SiH_2Cl_2) or silane (SiH_4) gases at a reactor pressure of around 5 mTorr, in the temperature range of 650ºC–800ºC. The chemical reaction can be written as:

$$3SiH_2Cl_2 + 4NH_3 \rightarrow Si_3N_4 + 6HCl + 6H_2$$
$$3SiH_4 + 4NH_3 \rightarrow SiN_3 + 12\ H_2$$

It is found that nitrogen gas can be replaced by NH_3 gas for better quality of silicon nitride, good silicon to nitrogen bonding, good refractive index, and better film density.

The silicon nitride film is also deposited by the Plasma Enhanced CVD, i.e. the **PECVD** technique at a low temperature (200ºC–400ºC). Generally, SiH_4 and NH_3 gases are used for silicon nitride film deposition. Unfortunately, the film composition and the structure are highly process dependent.

10.6.2 Titanium Nitride (TiN) Deposition

The titanium nitride film is widely used as a diffusion barrier layer. Titanium nitride is better than TiW with respect to film quality, diffusion barrier capability, and adhesion properties.

TiN is deposited by the reactive sputtering technique, where Ti is deposited in the presence of nitrogen gas in the plasma condition; Ti reacts with the nitrogen, TiN is formed, and it deposits on the wafer. This process of sputtering is called **reactive ion sputtering**. The reactive ion sputtering system is same as the sputter deposition system, as shown in Fig. 10.13. The properties of the TiN film are highly dependent on the sputtering parameters.

TiN film is also deposited by the hot wall CVD (LPCVD) technique in the temperature range of 400ºC – 700ºC. The chemical reaction can be written as

$$6TiCl + 8NH_3 \rightarrow 6TiN + 24\ HCl + N_2$$

The TiN film deposited by the LPCVD technique has the advantages of good step coverage and filling properties. When the deposition takes place at higher temperature, the TiN film has lower resistivity, higher density, and lower Cl_2 content. The major disadvantage of this technique is the chlorine content in the film.

10.7 SILICIDE DEPOSITION

The resistivity of a silicide film is around 10 times lower than that of a heavily doped polysilicon film. In addition, silicides are highly stable at high temperature. Therefore, silicide is replacing polysilicon interconnection wirings in ICs. $TiSi_2$, WSi_2, TaSi, MoSi, NiSi, PtSi and CoSi films are the common silicides used for IC fabrication. Among these silicides, the $TiSi_2$ and WSi_2 films are the most popular. The silicide films are mostly deposited by the CVD technique, but in some cases, the sputtering technique is preferred. The chemical reactions that take place when tungsten silicide is deposited by the CVD technique are mentioned below:

$$WF_6 + 2SiH_4 \rightarrow WSi_2 + 6HF + H_2\ (300ºC\ to\ 400ºC)$$

$$WF_6 + 3.5SiH_2Cl_2 \rightarrow WSi_2 + 1.5SiF_4 + 7HCl\ (500ºC\ to\ 600ºC)$$

$$WF_6 + 3.5SiH_2Cl_2 \rightarrow WSi_2 +1.5SiCl_4 + HCl + 6HF\ (500ºC\ to\ 600ºC)$$

10.8 DIELECTRIC DEPOSITION

Generally, the dielectric films are used for electrical isolation, planarisation, capacitor fabrication and dopant diffusion masks. The silicon dioxide and silicon nitride films are the two dielectrics commonly used in IC fabrication.

10.8.1 Silicon Dioxide Deposition

The deposition of the silicon dioxide film is preferably done by the CVD process. To deposit the silicon dioxide film, silane and oxygen gases are used and the deposition is carried out at a low temperature. The chemical reaction can be written as

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

The oxygen can be substituted by nitrous oxide (N_2O), NO or CO_2 gas. At low temperature, the silicon dioxide film deposited by the APCVD technique suffers from inferior step coverage, porosity, low density and low refractive index (~1.44). For better oxide quality, the film is deposited at low temperature by the LPCVD or the PECVD technique. The silicon dioxide film deposition by the LPCVD technique is achieved by the decomposition of tetraethylorthosilicate (TEOS) precursor in a temperature range of 650ºC – 800ºC. The decomposition of TEOS can be written as

$$Si(OC_2H_5)_4 \rightarrow SiO_2 + \text{gaseous form of by-products}$$

The silicon dioxide film deposited by the LPCVD process has a good step coverage, higher density, and higher refractive index. Unfortunately, this deposition process cannot be used after aluminium film deposition, due to the higher temperature of the deposition process.

Better step coverage and a conformal silicon dioxide film can be obtained by the reaction between Tetraethylorthosilicate (TEOS) and ozone (O_3) in the PECVD reactor. The film can be deposited at low temperature (around 500ºC), but the silicon dioxide film suffers from a high carbon level and a high density of porosities. The excellent filling and planarisation of the silicon dioxide film is obtained by the High-Density-CVD technique at a very low temperature.

10.9 SILICON DEPOSITION

In IC fabrication, there are two types of silicon films that are used, namely, polysilicon and epitaxial silicon.

10.9.1 Polysilicon Deposition

Polysilicon is mainly used for MOS gate electrodes, short interconnections (wiring) and resistor fabrication. The silicon film is made of small grains with different orientations;

therefore, it is called **polysilicon** or in short, **poly**. The grain size of the poly ranges from 0.03 μm to 0.3 μm.

The polysilicon film can be deposited by the sputtering or the LPCVD technique, but to get a good conformal film and in-situ poly doping, the LPCVD technique is preferred. The polysilicon film deposition is done in the LPCVD reactor at around 5 mTorr pressure, and the deposition temperature ranges from 600ºC to 650ºC. Generally, the SiH_4 precursor is used for the polysilicon film deposition. The SiH_4 can be diluted by 20% to 30% by adding nitrogen gas. The pyrolytic decomposition of silane can be written as

$$SiH_4 \rightarrow Si + 2H_2$$

10.9.2 Epitaxial Silicon Deposition

The word **epitaxy** is a combination of two Greek words, **epi** (meaning **on**) and **taxis** (meaning **arrangement**). In the context of IC fabrication, the epitaxial silicon stands for a crystalline silicon film deposited over the silicon wafer or on a dielectric film. If the epitaxial silicon film has the same orientation as the silicon wafer, then the epitaxial film is called a **homoepitaxial film**. It is important to mention here that the epitaxial deposition is carried out in the surface reaction (K_s) mode in a high temperature range between 1000ºC and 1200ºC. The thermal energy provides enough kinetic energy to the silicon atoms to move (surface migration) and align with the silicon wafer orientation. This process of crystallisation is called **solid phase epitaxy**. The silicon wafer orientation provides seed to the epitaxial silicon film and makes it aligned with the wafer orientation. If the epitaxial film is to be deposited on the dielectric film, then the crystallisation seed is provided by etching a portion (window) of the dielectric film so that the deposited epitaxial film gets the seed of crystal orientation from the wafer. The epitaxial film can be doped after or during film deposition. Usually, AsH_3, PH_3, and B_2H_6 gases are used for doping the epitaxial film.

The epitaxial silicon is deposited by the pyrolytic decomposition of silane (SiH_4) at a high temperature by the APCVD (conventional) technique. The pyrolytic decomposition of silane gas can be written as

$$SiH_4 \rightarrow Si + 2H_2$$

Silicon tetrachloride ($SiCl_4$), dichlorosilane (SiH_2Cl_2), and trichlorosilane ($SiHCl_3$) mixed with hydrogen gas are also used for epitaxial film deposition. The overall reaction between the tetrachloride ($SiCl_4$) precursor and hydrogen gas can be written as

$$SiCl_4 + 2H_2 \rightarrow Si + 4HCl$$

It is worth mentioning that the HCl which is produced during the chemical reaction cleans the reactor as well as the wafers, and this leads to a good epitaxial film.

10.10 FILM THICKNESS MEASUREMENTS

The deposited film must be evaluated either at the time of deposition or after the film deposition. The thickness and the film properties must be optimised prior to IC fabrication. In this section, the evaluation of the film thickness is briefly described.

The thickness of the film can be measured (*in situ*) during the physical vapour deposition. The common technique used for thickness measurement is the quartz crystal resonance technique. A quartz crystal is made to oscillate at its resonance frequency. During the film deposition, the mass of the crystal increases, and that in turn shifts the resonance frequency of the quartz crystal. The thickness of the film is separately calibrated against the resonance frequency using other techniques; then, by determining the resonance frequency shift, one can estimate the thickness of the film. The shift of the resonance frequency gives the direct information about the film thickness.

There are a number of other techniques that are used for film characterisation. These film characterisations are described in Chapter 4 (Oxidation).

Summary

Films used for shallow trench, LOCOS process, gate electrodes, and metal and silicon silicide for electrical contacts and wiring, etc., are required in IC fabrication. The films should have the qualities of good adhesion, inertness to other IC materials, ease of deposition and patterning, and good composition; besides, they should be free from contamination, there should be no pinholes or voids (empty spaces), and the stress should be minimum. In addition, metal, alloy and silicide should have low resistance, whereas, the dielectric films should have high density, high breakdown, conformity, and good step coverage.

The film can be deposited by Spin–On–Glass (SOG) Deposition, Electrochemical Deposition, Physical Vapour Deposition (PVD) and Chemical Vapour Deposition (CVD) techniques. Generally, SOG deposition and electrochemical deposition are not used for ICs because of certain limitations. Usually, the sputtering and CVD techniques are extensively used, especially the LPCVD technique.

A large number of metal films such as aluminium (Al), copper (Cu), titanium (Ti), tungsten (W), and metal alloys are commonly used in IC fabrication, for electrical connections. The film and contact resistances should be minimum to avoid the lowering of the signal processing speed, and they must satisfy the requirements of an Ohmic contact.

Low resistance aluminium films can be deposited by the sputtering and LPCVD processes. Spiking and electromigration are the two major disadvantages of the aluminium film, whereas, copper suffers from fast oxidation, adhesion and etching problems. Presently, copper wiring is done by the damascene process. Tungsten (W) films are deposited by the sputtering and the LPCVD techniques. Tungsten is preferred for via plugging. Apart from via plugging, tungsten is also used as tungsten silicide (WSi_2) for diffusion barriers and local connections. Good step coverage and better conformity of the titanium film is obtained by the sputtering and LPCVD techniques. Many metal alloy films such as Titanium–Palladium–Gold (Ti-Pd-Au), Titanium–Platinum–Gold (Ti-Pt-Au), Titanium–Platinum (Ti-Pt), and Titanium–Tungsten (Ti–W) are used for diffusion barriers. Out of these alloy films, the Ti-W film has been proven to be more useful.

Titanium nitride layer is used as a metal diffusion barrier and also for antireflection in the lithography process. Other nitride and silicide films are $TiSi_2$, WSi_2, TaSi, MoSi, NiSi, PtSi, and CoSi, but among these silicides, the $TiSi_2$ and WSi_2 silicide films are the most popular. The silicide films are mostly deposited by the CVD technique, but in some cases, the sputtering technique is preferred.

Silicon nitride and silicon dioxide are used for the LOCOS process, encapsulation, passivation, and as dielectric films. The deposition of the silicon dioxide film is preferably done by the CVD process, but PECVD and sputtering are also powerful techniques. Polysilicon is mainly used for MOS gate electrodes, short interconnections (wiring) and resistor fabrication; and, for a good conformal film and *in-situ* poly doping, the LPCVD technique is preferred. The deposition of an epitaxial silicon film is carried out at a high temperature range using the CVD technique by the pyrolytic decomposition of silane (SiH_4).

References

- J D Plummer, M Deal and P B Griffin; *Silicon Fundamental Technology: Fundamentals, Practice and Modeling*, Prentice Hall, 2000
- S Wolf and R N Tauber; *Silicon Processing for VLSI Era*, Vol 1: Process Technology, Lattice Press, Second Edition, 2000
- S M Sze; *VLSI Technology*, Second Edition, McGraw-Hill, 1988
- S K Gandhi; *VLSI Fabrication Principles*, Second Edition, Wiley, 1994
- D Nagchoudhuri; *Principles of Microelectronic Technology*, Wheeler, 1998
- S A Campbell; *The Science and Engineering of Microelectronic Fabrication*, Oxford University Press, 1996

Multiple-Choice Questions

10.1 What is the minimum order of vacuum that is recommended for evaporation deposition?

(a) More than 10^{-6} Torr
(b) Less than 10^{-6} Torr
(c) 10^{-6} Torr

10.2 The sputtering target size with respect to the anode is

(a) smaller (b) same (c) bigger

10.3 Polysilicon is deposited at a temperature of about

(a) 620°C (b) 920°C (c) 1020°C

10.4 Polysilicon grains are oriented

(a) perpendicular to the silicon wafer
(b) parallel to the silicon wafer
(c) randomly

10.5 Oxide film is deposited on the wafer prior to the nitride film

(a) to reduce the stress at the silicon interface
(b) to improve the adhesion with the silicon
(c) to promote photoresist adhesion.

10.6 Generally, what is the thickness of aluminium that is deposited for electrical connections of device components?

(a) 5 microns
(b) 1 micron
(c) 3 microns

10.7 Relatively uniform film can be obtained by

(a) evaporation technique
(b) sputtering technique
(c) LPCVD technique

10.8 Relatively uniform film can be obtained by

(a) evaporation technique
(b) sputtering technique
(c) e-beam technique

10.9 Why is annealing done after aluminium metallisation?

(a) For better adhesion
(b) To reduce stress
(c) To get a uniform aluminium film

10.10 Generally, epitaxial silicon is deposited by
   (a) APCVD technique
   (b) LPCVD technique
   (c) physical deposition technique

10.11 To check aluminium electromigration, which metal is usually deposited prior to the aluminium deposition?
   (a) Chrome (b) Copper (c) Gold

Descriptive Problems

10.1 Calculate the mean free path at 1 Pa, 10^{-2} Pa and 10^{-3} Pa.

10.2 What is the basic difference between the etching and the film deposition processes?

10.3 Why is metal deposited in high vacuum ($>1 \times 10^{-6}$ Torr), and sputtering is done at higher pressure than metal deposition?

CHAPTER

11 ULSI (nano) Fabrication

11.1 INTRODUCTION

The market demand for higher speed, lower power and maximum number of functional ICs on a chip, especially in the areas of DRAM memory, logic and analogue ICs, microprocessors and microcontrollers compel the IC industry to scale down the MOS transistor dimensions significantly. This scaling down of the MOS transistor dimensions has led to the generation of Small-Scale Integration (SSI), Medium-Scale Integration (MSI), Large-Scale Integration (LSI), Very Large-Scale Integration (VLSI), and Ultra Large-Scale Integration (ULSI) circuit eras. Generally, these generations of ICs are based on the gate length, and also on the number of MOS transistor in a chip.

The ULSI generation started with a gate length of less than 1 micrometre and it will continue for a few years. The ULSI generation can be divided into three sub-divisions, namely, the submicron technology, the deep submicron technology, and the ultra-deep submicron technology. These subdivisions are not based on a specific gate length, but on a range of the gate length. For instance, the deep submicron gate length ranges from 90 nm to 65 nm, and the ultra-deep submicron gate length ranges from 45 nm to 32 nm.

In the previous chapters, MOS fabrication techniques and processing equipments have been described in general. These techniques and basic equipments were good enough for the fabrication prior to the ULSI generation. When the transistor gate length gradually enters the ULSI generation, the device fabrication becomes complex and complicated manifold and so do the fabrication techniques and equipment.

ULSI device fabrication is complex because of the features incorporated to control drain junction breakdown, punch through, short-channel effects and hot electron generation, current leakage, MOS transistor V_T fluctuation, contact resistance, parasitic resistance and capacitance, and the mismatch of MOS transistor threshold voltage and other device parameters to a certain limit.

To achieve the above parameters, one of the essential fabrication process constraints is the processing temperature. As the device is shrinking, in the ULSI generation, the low thermal budget (temperature multiplied by time) process is extremely essential. The low thermal budget process may be either for high temperature and less time or for low temperature and longer time. The ideal condition of low thermal budget is low temperature and short time. Presently, the ULSI device fabrication process cannot be carried out in the ideal thermal budget due to the non-availability of process equipments, processes technique and material requirement. At present, many fabrication processes are carried out at low temperature and longer time. Many researchers are contemplating to replace high thermal budget to low thermal budget processing, especially in silicon oxidation, metal deposition and dielectric deposition. Some of the existing equipments and processes have been modified to suit the ULSI device fabrication. In addition, new precursors (recipes), and new MOS structures are coming up to meet the challenges of ULSI device fabrication. Fortunately, dopant diffusion is done by ion-implantation, which is a process carried out at room temperature. In addition, ion-implantation can be done in an oblique angle that is required for the fabrication of many VLSI and ULSI devices.

To replace the high-thermal-budget processes to low-thermal-budget processes, the Rapid Thermal Processing (RTP) equipment has been introduced. The RTP equipment has been conceptualised from the Rapid Thermal Annealing (RTA) system. The RTA was used for wafer annealing to reduce the wafer defects, especially after ion-implantation (see Chapter 9). The name RTA has been coined due to its fast processing time; where the wafer is heated by intense light for a few seconds. Even though the RTP (and RTA) equipment has the limitation of single wafer processing that results in a low wafer throughput, it is the trend of today’s technology. Derivatives of the RTP system are used for different process applications and are named after their specific fabrication process steps; for example, if the RTP equipment is used for oxidation, then it is called **Rapid Thermal Oxidation (RTO)**, and when it is used for CVD then the system is called **Rapid Thermal Chemical Vapour Deposition (RTCVD)**. In the RTO chamber, the annealing process is performed immediately after oxidation, hence, this process is abbreviated as RTO-RTA. Similarly, RTCVD followed by RTA annealing is called RTCVD-RTA. The challenging problems in the RTP and RTA processes are uniform wafer heating and temperature measurement in case of wafers of large diameter (800 mm and above). Apart from thermal budget, other process-related issues, such as lithography resolution, exposure systems, etching, film deposition in holes, multi-level wiring, contact, etc., in ULSI device fabrication are still to be overcome.

In this chapter, silicon oxidation, film deposition of different dielectrics, epitaxial film deposition, silicon nitride film deposition, film deposition of different silicides, aluminium and tungsten film deposition, lithography, and etching, in view of the ULSI (nano) fabrication technology is discussed.

11.1.1 Silicon Dioxide

The ULSI transistors require very thin silicon dioxide gate film; for example, around 70 Å of gate thickness is needed for 0.25 μm of the MOS gate. This thickness comprises only a few atomic layers and reaches the fundamental limits of silicon dioxide as a gate dielectric. Decrease in the oxide thickness leads to an exponential increase in the gate leakage current. For this reason, one of the high gate dielectric materials such as silicon nitride or silicon oxynitride (SiOxNy), zirconium oxide (ZO_2), hafnium oxide (HfO_2), aluminium oxide (Al_2O_3), or lanthanum oxide (La_2O_3) dielectric films may replace the silicon dioxide in the ULSI device.

The cleaned silicon reacts with environmental oxygen molecules almost instantaneously at room temperature (Chapter 4), and forms a silicon dioxide film of 20 Å thickness. This oxide film is sometimes called **native oxide**. This native oxide is not device worthy. To avoid growing the native oxide (to some extent), the wafer is usually cleaned in the HF. The hydrogen atom from the HF binds with the silicon atom and that checks the growth of the native oxide. At the oxidation temperature, the silicon atom unbinds from the hydrogen atom and reacts with the oxygen atom to form silicon dioxide. Generally, dry oxidisation is carried out in the oxygen gas environment from 700 to 1000°C temperature range for a short period of time. Extreme care is taken during the dry oxidation process, so that water molecules more than 125 parts per million (ppm) are not present in the oxygen gas, otherwise a good-quality oxide film cannot be obtained (see Chapter 4). To fabricate the ULSI device, it requires a good quality and uniform thin silicon oxide across the wafer; therefore, a stringent oxidation control process is required. For instance, a 0.25 μm MOS transistor needs around 70 Å gate thicknesses with not more than 7 Å thickness variations across the wafer, otherwise, the transistor threshold voltage will vary from transistor to transistor across the wafer. Furthermore, the **defect density** in the oxide should be less than $0.5/cm^2$ to keep the leakage current under limit.

11.1.2 Silicon Oxidation by Rapid Thermal Oxidation (RTO/RTA)

In the rapid thermal oxidation technique, the wafer is exposed to a very intense light for a very short time in the presence of oxygen gas, as shown in Fig. 11.1; and the typical heating sequence is shown in Fig. 11.2. This technique has gained more popularity than the furnace oxidation technique, mainly because of the following reasons: (a) to get a thin oxide film, (b) there is no increase in junction depth, (c) there is no horizontal dopant diffusion, and (4) there is no change in the diffusion profile. Thus, the RTO technique for growing the thin oxide film is better than the furnace oxidation technique for ULSI device fabrication.

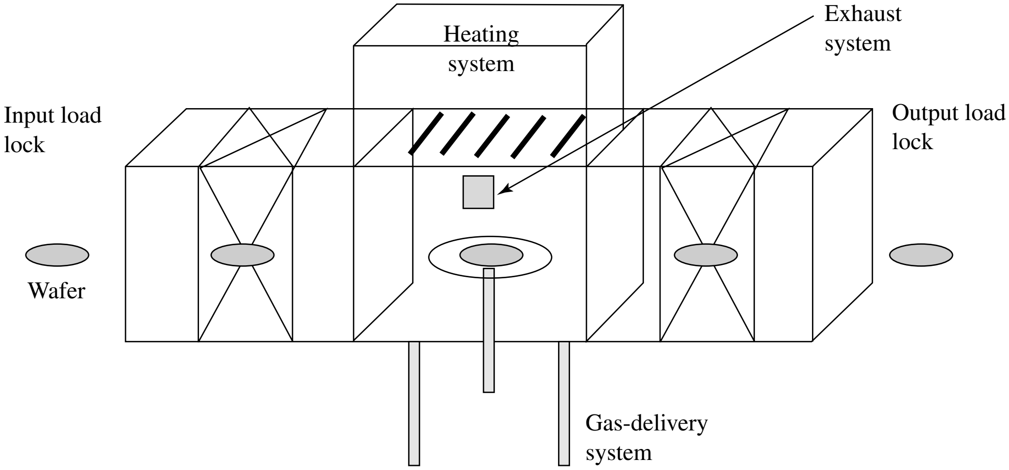


Fig. 11.1 Schematic of RTO system

The RTO/RTA dry oxidation is carried out at around 1000°C in the pure and particulate free oxygen gas. The wafer is heated gradually by increasing the light intensity, which is called the **ramp up**, for a few seconds in the presence of nitrogen gas, to avoid thermal shock to the wafer. Thereafter, the light intensity is kept constant for a few minutes in the oxygen gas for oxidation. The oxidation time is chosen according to the desired oxide thickness. Thereafter, light intensity is gradually decreased for a few seconds in the nitrogen gas. The last process is called the **ramp down**. The ramp down process is also referred to as the **Annealing** process. During ramp down, the wafer is annealed after the oxidation process, to reduce the oxide charges (see Chapter 4). The entire sequence of steps from the ramp up to the ramp down is called the **heat cycle** and shown in Fig. 12.

The RTO oxidation mechanism is not well understood till today, but it is believed that in the presence of intense light, the oxygen molecules dissociate into the oxygen radicals

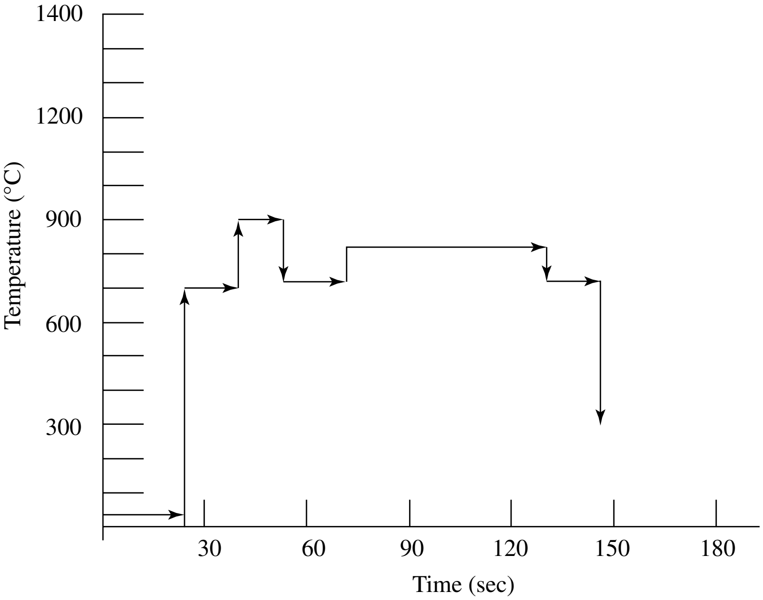


Fig. 11.2 Typical RTO heating sequence

O and then these oxygen radicals react with the silicon atoms and form silicon dioxide. The generation of oxygen radicals can be described as

$$O_2 \rightarrow O + O$$
$$O + O_2 + O_2 \rightarrow O_3 + O_2$$
$$O + O_3 \rightarrow 2O_2$$
$$O_3 \rightarrow O + O_2$$

It has been found that the oxide growth rate by the RTO process is initially linear at high temperature till 20 seconds and then it becomes nonlinear. It has also been found that the oxidation rate increases with an increase in the oxidation temperature. In the linear region, the activation energy of RTO oxidation is around 1.44 eV which is close to the activation energy of furnace oxidation which is 1.76 eV. The RTO processed oxidation has less oxide charges and high charge-to-breakdown of around 80 C/cm^2 compared to all 20 C/cm^2 in the case of furnace oxidation. It is believed that RTO needs relatively higher oxidation temperature than the furnace oxidation of thin silicon dioxide.

11.2 DIELECTRIC FILM DEPOSITION

The basics of the dielectric film deposition technique have been mentioned in Chapter 10. Generally, dielectric films are deposited by the Chemical Vapour Deposition (CVD) process and are mainly used for electrical isolation and planarisation of surface morphology, gate wall passivation and diffusion mask. In addition, a dielectric is deposited in between the metal levels. This dielectric is called the **Interlevel dielectric (ILD)**. To satisfy these requirements the dielectric film has to be of uniform thickness and good step coverage, no void, having high dielectric constant, and free of contaminations.

Silicon dioxide and silicon nitride dielectrics are extensively used in the ULSI devices. Generally, the deposition is preferred to be carried out at lower temperature, so that these dielectrics can be deposited over the metals having low melting point like aluminium. The dielectric film deposition process at low-temperature suffers from one or the other problem of less step coverage, non-uniform deposition, selective deposition, chemical residual, void and particulate and less planarisation properties. In this section, the silicon dioxide and silicon nitride dielectric deposition techniques, their advantages and disadvantages are covered in brief.

11.2.1 Silicon Dioxide Deposition by the RTO Process

Silicon dioxide can be deposited by the pyrolysis of Tetraethoxysilane (TEOS) precursor. The chemical reaction is shown below.

$$Si(OC_2H_5)_4 \rightarrow SiO_2 + 2H_2O + 4C_2H_4$$

It has been found that below 800°C, the oxide deposition rate is slow, and above this temperature, the oxide deposition rate significantly increases to 1000 Å/min. This silicon dioxide is good as a dielectric film for the purpose of multilevel IDL where aluminium metal is not present.

Silicon dioxide deposition can be done with silane gas as a precursor added with O_2 or N_2O_2 gas. Chemical reactions of these gases are mentioned below.

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

$$SiH_4 + 2N_2O_2 \rightarrow SiO_2 + 2N_2 + 2H_2O$$

Silane and oxygen reaction takes place at the temperature of 300°C to 450°C, but at this temperature, the particulates are generated; thus, this technique is generally avoided.

At higher temperature around 800°C, a good-quality stoichiometric oxide film is formed, if silane is mixed with nitrous oxide. This process is almost 100 times faster than the furnace thermal oxidation, but cannot be used in IDL where aluminium is present, because of high deposition temperature.

Silicon dioxide can also be deposited by the APCVD and LPCVD techniques, but cannot be used for ILD due to high deposition temperature. Furthermore, this CVD deposited silicon dioxide is not used because it is a porous oxide and has water absorbing property.

Silicon dioxide can be deposited in the high aspect ratio structures with good step coverage by the APCVD technique and sub-atmospheric CVD (SACVD) technique using TEOS as the precursor in the temperature range of 250°C to 400°C. Till today, the CVD deposition chemistry is not well known, but it is believed that in the CVD reactor, TEOS decomposes and releases ozone that in turn reacts with TEOS and forms silicon dioxide. The deposition rate of silicon dioxide depends on the ozone concentration and temperature. It has been found that the silicon dioxide deposition rate increases till 1% of the ozone concentration, and thereafter, it declines with an increase in the ozone concentration. The oxide deposition rate also increases in the range of 100°C to 250°C. It has been found that at higher temperature and with higher ozone concentration, a good quality oxide film is obtained. In addition, the oxide deposition rate is 30% higher on bare silicon than on silicon dioxide. Unfortunately, the TEOS deposited oxide is porous and absorbs water molecules. If silicon dioxide contains water then the contact metal corrodes with time. This phenomenon is known as **contact poisoning**. To prevent contact poisoning, a combination of PECVD and CVD deposited oxides are generally used.

A good-quality silicon dioxide can be obtained using TEOS by the plasma assisted (PECVD) technique. In the presence of plasma, TEOS reacts with oxygen below 400°C and silicon dioxide is formed, as

$$Si(OC_2H_5)_4 + O_2 \rightarrow SiO_2 + \text{by-product}$$

The deposition rate of silicon dioxide is shown in Fig. 11.3 and by this technique, a good step coverage is obtained. It has also been found that oxide deposition takes place more at the edge of the metal line than at other places. This leads to void formation, as shown in Fig. 11.4. To avoid extra deposition at the corners of the metal lines, the **deposition-etch-deposition (dep-etch-dep)** process is used.

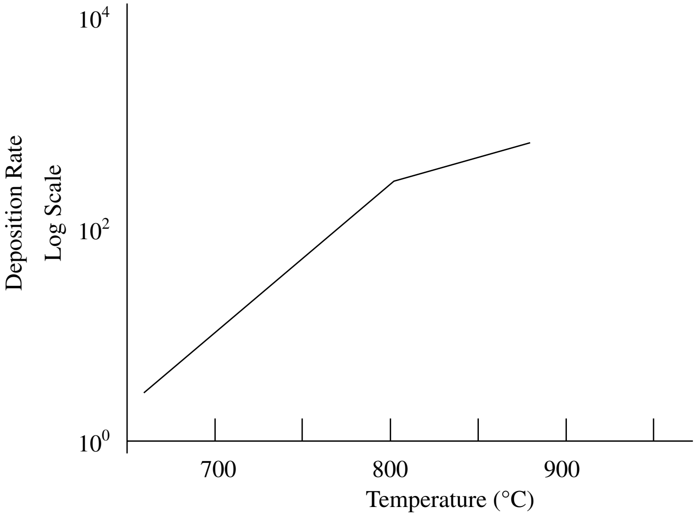


Fig. 11.3 Silicon dioxide deposition (rate Å/min) using TEOS in RTO process

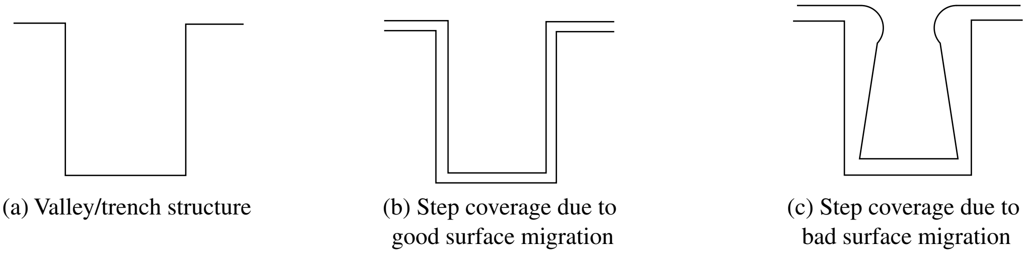

(a) Valley/trench structure
(b) Step coverage due to good surface migration
(c) Step coverage due to bad surface migration

Fig. 11.4 Step coverage and surface migration

Silicon dioxide can be deposited by the N_2O reduction of silane in the presence of plasma, but it suffers from poor step coverage although it can be improved by high-density plasma process.

11.2.2 Quartz Sputtered Deposition

High-quality oxide can be deposited in the narrow gaps and undercuts using quartz target, for the ILD application, by the High-Density Plasma (HDP) sputtering technique at low temperature, if the wafer is electrically biased. In this HDP technique, qualities like high deposition rate, less plasma radiation exposure, less compressive stress, and no particulates of silicon dioxide film, is obtained. It seems that for deep submicron device fabrication, the substrate biased HDP technique may be useful in the future.

11.2.3 Spin-on-Glass

Silicon dioxide can be deposited by the liquid spin-on-glass (SOG) chemical. The coating of liquid SOG is similar to the coating of the photoresist. Once SOG is coated on the wafer, it is solidified to silicon dioxide by heating. This process is called the **curing** of SOG. Silicon dioxide deposition by the SOG technique is simple, has good step coverage and is low cost, but it is generally not used in the ULSI device fabrication. There are two types of SOGs used for low-cost IC fabrication. One type of SOG is inorganic based and another type of SOG is organic based. Both of these SOGs can be modified by adding boron or phosphorous. These types of SOGs are called **modified SOGs**. SOGs are available in a large range of molecular weight and viscosity. This gives flexibility of use in the device fabrication process. Unfortunately, SOG oxide absorbs water and poisons the metal. In addition, the SOG film oxide has a low dielectric constant ranging from 3.0 to 3.6, which is much less than the dielectric constant of 4.5 obtained by furnace oxidation.

The silicon dioxide film obtained by inorganic (silicate) SOG is thermally stable and does not absorb water significantly, but it suffers from volume shrinkage during curing that leads to cracks due to high stress (around 400 MPa). This problem can be solved by repeatedly applying a very thin coating of SOG and curing. This process is continued till the desired thickness is achieved. The inorganic SOG technique has the ability to fill the narrow gaps between the patterns. Silicate SOG can be modified by introducing phosphorus in quantities less than 4%. The silicon dioxide obtained by this process is softer, has less stress (around 200 MPa), thus less cracks. Unfortunately, the phosphorus absorbs water that results in metal poisoning and poor mechanical and electrical properties.

The organic SOG (siloxane) deposited oxide has less stress (around 150 MPa), hence, a thick oxide can be deposited. On the other hand, siloxane oxide absorbs water molecules, contains carbon, has poor mechanical strength, has low melting point (400°C), and cannot stand plasma exposure.

11.2.4 Nitride Deposition

The silicon nitride (Si_3N_4) film is mainly used for the LOCOS process, for the oxidation mask, for the passivation of the IC, as diffusion barrier, and for the gate of the Metal Nitride Oxide Semiconductor (MNOS). The LOCOS process is not used in the ULSI device fabrication, as it consumes the silicon area due to bird beak formation (see Chapter 3). Generally, in a submicron MOS device, a silicon nitride film is deposited by RTCVD, but by this process, a thick silicon nitride deposition is obtained. However, a good-quality thin film can be obtained using a mixture of silane and ammonia gases in the ratio of 1:120, at a temperature of around 800°C with high deposition rate (around 100 Å/min). The chemical reaction takes place as shown:

$$3SiH_4 + 4NH_3 \rightarrow Si_3N_4 + 12H_2$$

Silicon nitride deposition can be done using trichlorosilane (SiH_2Cl_2) and ammonia gases by the LPCVD technique at a temperature of 700°C to 800°C. The chemical reaction can be expressed as

$$3SiH_2Cl_2 + 10NH_3 \rightarrow Si_3N_4 + 6NH_4Cl + 6H_2$$

These gases are usually used in the APCVD and LPCVD deposition techniques, but in the RTCVD technique, the particles of NH_4Cl deposit on the cooled places of the chamber.

Nitride deposition by APCVD at 700°C to 800°C temperature (similar to RTCVD) using silane precursor mixed with ammonia gas can also be used. The chemical reaction of ammonia reduction is mentioned below.

$$3SiH_4 + 4NH_3 \rightarrow Si_3N_4 + 12H_2$$

11.2.5 Silicon Deposition

Polysilicon Deposition

The polysilicon is mainly used for the MOS gate, short electrical connections, the resistor, barriers of metal diffusion and metal silicide formation. Generally, doped polysilicon is used for IC fabrication. The doping can be done during the polysilicon deposition (in situ) or after the polysilicon deposition. The details of polysilicon deposition have been mentioned in Chapter 10.

Polysilicon can be deposited by simple pyrolysis of silane gas as

$$SiH_4 \rightarrow Si + 2H_2$$

Generally, saline gas is diluted either with nitrogen or with argon gas. The surface quality of the deposited poly film is dependent on the ratio of the silane and the nitrogen or argon gases.

For a reasonable deposition rate, pyrolysis of TEOS is done in the RTP process; where the temperature is kept above 700°C for a very short time. The polysilicon surface comes out to be smoother when polysilicon deposition is done by the RTCVD process at high temperature.

Epitaxial Silicon Deposition (Growth)

The word *epitaxial* comprises two Greek words: *epi* (upon) and *taxy* (order). In the context of the MOS transistor, epitaxy means a crystalline (order) film deposited over (upon) the wafer. The process of making a crystalline film on the wafer is called the **epitaxial film deposition or growth**. The doping profile and doping concentration in the epitaxial (epi) film can be tailored according to requirements; hence, the designer has more freedom to design the desired MOS transistor characteristics. Therefore, the epitaxy process is much better than the diffusion and the ion-implantation processes. Furthermore, epitaxial film deposition rate is much higher on bare silicon than on silicon oxide or silicon nitride, and that helps in planarisation of the wafer surface. For instance, the epitaxial growth is faster in the silicon trench than at other place of the wafer. Trench isolation has many advantages over the LOCOS process such as: no bird's beak formation, better latch-up prevention, is suitable for submicron electrical isolation, has good planarity, and good buried layer.

The epitaxial film material can be deposited either on the substrate material of the same type or it can be deposited on the substrate material of different type. The first type of epitaxy is called **homoepitaxy** and the latter is called **heteroepitaxy**. A good-quality homoepitaxial layer can be grown that has insignificant stress, is thermodynamically stable, and has less defects.

The homoepitaxial layer is deposited by the CVD technique at high temperature (see Chapter 9), but it is not suitable for submicron device fabrication because it is a high-temperature (above 1000°C) deposition process. In addition, the high-temperature process introduces unwanted auto-doping, washout and pattern shift distortion. The washout and pattern shift phenomena become more prominent when the epitaxial film is deposited over the buried layer, especially for bipolar devices.

Epitaxial Deposition Techniques

There are two main criteria needed for the epitaxial film deposition: (a) the silicon atoms should loosely adhere to the substrate, and (b) the deposited silicon atoms should migrate to the substrate at an appropriate place to form the crystalline film. The phenomenon of epitaxial atom migration on the substrate surface is called **surface migration**, and it needs sufficient thermal energy.

In general, for submicron ULSI device fabrication, a low temperature around 500°C is used for processing in the RTCVD. This low temperature epitaxial deposition does not cause a significant alteration of junction depth, auto-doping and the doping profile.

There are two techniques used for the silicon epitaxial film deposition. In the first technique, the epitaxial layer deposition is done in an Ultrahigh Vacuum Chemical Vapour Deposition (UHVCVD) system, where the deposition is carried out below 10^{-9} Torr. For submicron ULSI device fabrication, it is essential that the silicon substrate has been ultra cleaned prior to the epitaxial film deposition. The gases used for the epitaxial film deposition are silane (SiH_4) or disilane (Si_2H_6). Out of these gases, silane is preferred because of the low deposition temperature. A typical sketch of UHVCVD is shown in Fig. 11.5. The UHVCVD system contains a reactor (furnace), load lock systems, high vacuum systems, exhaust system, a mass spectrometer, gas-delivery lines, a gas detector, and other service supports. To drive away moisture and other gases, the inlet load lock system is heated at around 100°C. Then, the silicon wafer is loaded into the inlet load lock system and thereafter, a high vacuum of the order of 10^{-6} Torr is created, and finally the wafer is transferred onto the reactor plate. Prior to wafer transferring, the reactor is heated at around 550°C (or the epitaxial deposition temperature). Once the wafer is transferred onto the reactor plate, an ultra high vacuum is created below 10^{-9} Torr, and thereafter, the reactor vacuum is reduced to around 10^{-3} Torr by introducing well-regulated silane gas through the gas flow controller. If the epitaxial layer is to be doped then diborane (B_2H_6) or phosphine (PH_3) gases are introduced in the system during the epitaxial film deposition (*in situ* doping). Once the epitaxial film is deposited, the wafer is transported to the outlet load lock system and finally, the wafer is taken out in reverse order of how it was loaded onto the reactor plate. Prior to epitaxial growth, the silicon wafer is thoroughly cleaned using the RCA cleaning procedure.

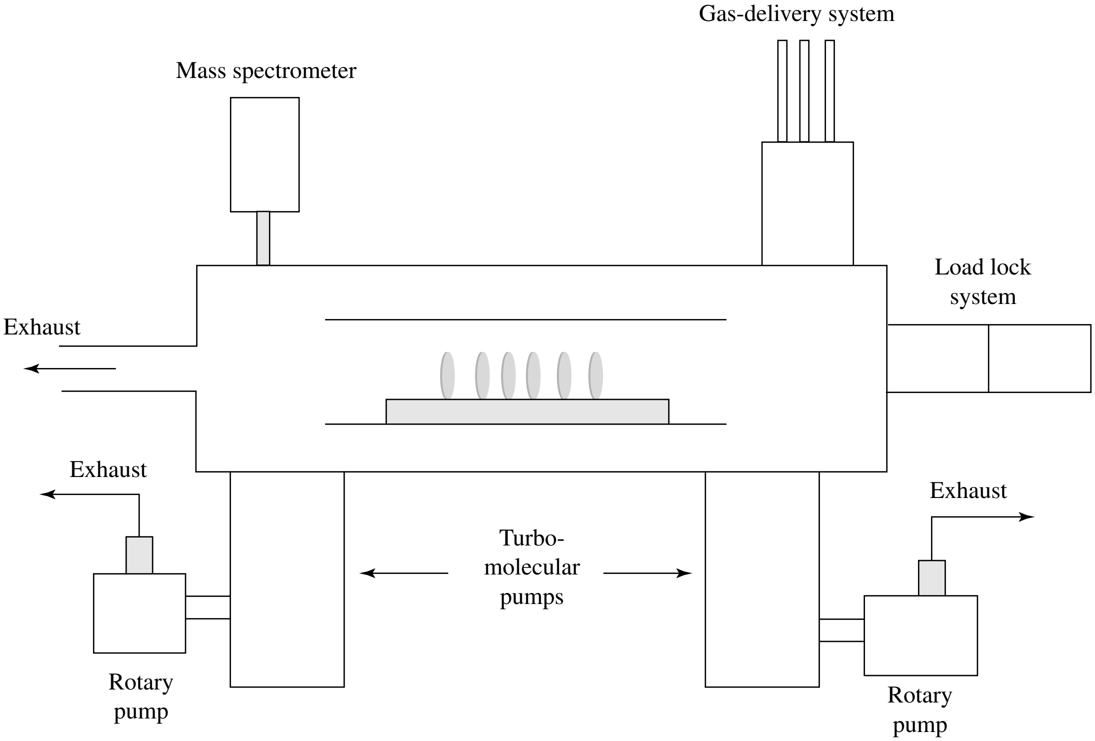


Fig. 11.5 Block diagram of UHV-CVD

The other method of epitaxial deposition is the Molecular Beam Epitaxy (MBE) technique. The Molecular Beam Epitaxy machine is more complicated than the UHCVD reactor, but it has the advantages of having a precise control over the epitaxial thickness and doping. The principle of MBE deposition is the same as the e-beam evaporation process (see Chapter 10). The desired epitaxial film doping can be done by evaporating the silicon and the dopant simultaneously (*in situ*). Film deposition in the MBE system is done better at ultra-high vacuum than at a pressure of 10^{-11} Torr, in the temperature range of 500°C to 900°C. Solid silicon is heated by an e-beam that generates vapours of silicon atoms. These silicon atoms deposit in the form of a crystalline film on the wafer. The silicon epitaxy deposited by the evaporation of silicon by the e-beam suffers from cluster formation and spitting problems. To overcome these defects, a gaseous source (GS) is used as shown in Fig. 11.6. This type of MBE is similar to UHVCVD and is called **GSMBE**, except that GSMBE is a cold wall system, and UHVCVD is a hot wall system. It is expected that epitaxial film deposition by GSMBE will be the technique of the future for deep submicron device.

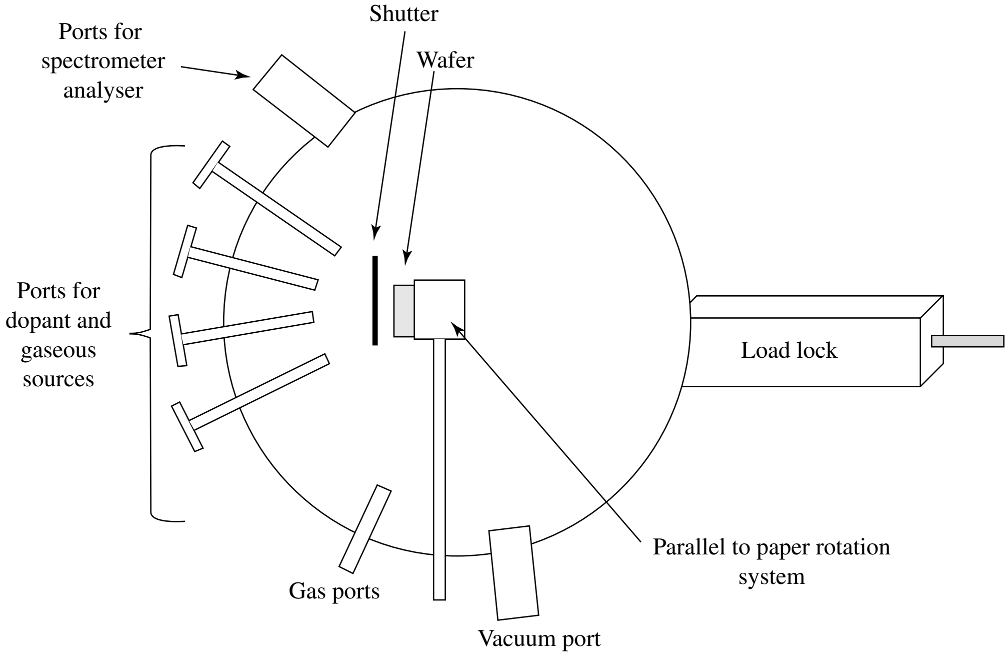

Shutter
Ports for spectrometer analyser
Wafer
Ports for dopant and gaseous sources
Load lock
Parallel to paper rotation system
Gas ports
Vacuum port

Fig. 11.6 Symbolic diagram of GSMBE

11.2.6 Metal Deposition Techniques

Metal deposition (metallisation) is done for device contact and electrical wiring. The basics of metallisation have been described in Chapter 10. Ideally, the metal contact with gate, source, and drain should have minimum possible resistance and must follow Ohm's law; whereas, the metal wiring should have minimum possible resistance and capacitance. Presently, all device contacts and metal (circuit) wirings are not possible on a single layer (surface) because of insufficiency of space. Therefore, contacts and wirings are done in between the dielectric layers. This scheme of electrical wiring is called **multilevel wiring**. In present days, two- to three-level wiring is done.

The Physical Vapour Deposition (PVD) and the Chemical Vapour Deposition (CVD) metallisation techniques have been discussed in Chapter 10. These two metallisation techniques are used for the submicron device applications. It seems that these techniques can be extended for deep submicron device fabrication also, with some modifications or by inventing a new technique. In this section, the deposition techniques of aluminium, tungsten, titanium, copper, and cobalt in the context of ULSI device fabrication are discussed.

Aluminium Deposition

Aluminium can be deposited by PVD, especially the sputtering technique. This deposition technique has the advantages of uniform deposition, controlled thickness deposition, and pre-deposition wafer cleaning.

The reflow of aluminium for plugging has been reported to have little success. Aluminium is first deposited on the wafer and then heated by the RTP, excimer laser or UV laser to plug the contact holes. At high temperature, aluminium melts and fills the holes by the reflow process due to surface migration. If aluminium is deposited on the barrier layer and thereafter heated by the RTA, the molten aluminium reacts with the barrier layer and damages its quality. In the excimer laser heating technique, the molten aluminium remains there only for microseconds and does not get sufficient time to react with the barrier layer; hence, no barrier film damage takes place. The UV laser based aluminium reflow needs a very intense beam as aluminium has high reflectivity in the UV range; whereas, with the high laser beam heating technique, the ablation of aluminium takes place and it also reacts with Ti that results in a high resistivity. High temperature *in situ* reflow of aluminium is also reported. In this process, aluminium is first deposited and then it is subjected to a high temperature. Generally, the temperature is kept more than 500°C for a few minutes under vacuum at around 1×10^{-8} Torr. As the process is carried out at a high temperature for a long time, a robust TiN barrier layer is needed. To increase the wetting between Al and TiN, a layer of Ti, $TiSi_2$ or Si is sandwiched. Aluminium plugging can be done by the low and high deposition technique. In this process, a thin aluminium film is sputtered at a low temperature and the next layer of aluminium is deposited at a high temperature of around 450°C. The low-temperature aluminium film provides the seed for surface diffusion for the high-temperature aluminium deposition. In this process, the barrier film requirement is less stringent, but a high vacuum is essential. A low temperature and high pressure is also suggested for aluminium plugging. A thick aluminium layer is deposited at low temperature and then the wafer is subjected to a high pressure of around 600 MPa at a temperature of 350°C to 400°C. The temperature softens the aluminium and the applied pressure pushes the aluminium into the hole. It has been found that aluminium plugging depends on the shape and the aspect ratio of the holes.

Aluminium can also be deposited by the CVD technique using several precursors, one of them being tri-isobutyl-Al (TIBA). The TIBA decomposes in the presence of hydrogen at a temperature of 40°C to 50°C and produces DIBAH. Thereafter, DIBAH further decomposes at 150°C to 300°C to form AlH_3, and finally AlH_3 decomposes into pure aluminium and hydrogen. The chemical reaction is as follows:

$$\text{TIBA} + H_2 \rightarrow \text{DIBAH} + 2C_2H_8$$

$$\text{DIBAH} + H_2 \rightarrow AlH_3 + 2C_4H_8$$
$$AlH_3 \rightarrow Al + 3/2\ H_3$$

With the TIBA precursor, a low aluminium resistivity (~2.8 μΩ-cm) that is close to bulk aluminium, is obtained when the TiN barrier layer is used in between the aluminium metal and the silicon. On the other hand, the TIBA precursor has some disadvantages like (a) low vapour pressure, (b) less utilisation of TIBA, (c) explosive, (4) prone to fire, (5) reacts violently with water, and (6) copper metal cannot be deposited simultaneously to check the aluminium electromigration. The problem of electromigration can be solved if half of the aluminium thickness is deposited using the TIBA precursor and the rest half is deposited by sputtering the copper-aluminium target. An aluminium film having 2.5% copper along with Dimethyl Aluminium Hydride (DMAH) and Cyclopentadienyl Copper Triethylphosphine (CpCuTEP) precursors can also be used. A low deposition rate is not preferred for aluminium deposition for the ULSI applications.

Tungsten Deposition

Tungsten is widely used for plugging vias (holes) and short electrical connections. Usually, the tungsten metal is deposited by the CVD technique. The wafer is heated between 400°C and 500°C in the presence of WF_4 precursors, as mentioned below.

$$WF_6 + 3H_2 \rightarrow W + 6HF \quad \text{(hydrogen reduction)}$$
$$2WF_6 + 3Si \rightarrow 2W + 3SiF_4 \quad \text{(silicon reduction)}$$
$$WF_6 + SiH_4 \rightarrow W + SiF_4 + 2HF + H_2 \quad \text{(silane reduction)}$$
$$2WF_6 + 3SiH_4 \rightarrow 2W + 3SiF_4 + 6H_2 \quad \text{(silane reduction)}$$
$$WF_6 + 2Al \rightarrow W + 2AlF_3 \quad \text{(aluminium reduction)}$$
$$2WF_6 + 3\ Ti \rightarrow 2W + 3TiF_4 \quad \text{(titanium reduction)}$$

The advantages and disadvantages of these techniques are described as follows. The Si, Al, and Ti reduction is not favourable. Tungsten consumes around 20 nm silicon and that may lead to junction leakage, especially when W is directly deposited on the source and the drain contacts. To avoid junction leakage, the first step is to deposit W by the SiH_4 precursor followed by the WF_6 precursor. To check the junction leakage, TiN or TiW barrier is used. Out of these barrier layers, TiN is found to be most suitable. These layers also provide good adhesion of W with SiO_2. Resistivity of the W film is around four times higher than the Al film; for this reason, W is only used for short connections.

The W plugging by hydrogen reduction suffers from selective nucleation (growth). It has been found that W deposits in the contact holes as well as on few places on the silicon dioxide surface. To increase selectivity, the TiN or TiW layer is deposited on the contact holes prior to W deposition. This scheme has advantages of uniform plugging and better adhesion irrespective of the height of the holes.

W plugging can be achieved by global W deposition using silane reduction. Prior to W deposition, TiN is deposited. Thereafter, both W and TiN are removed by reactive ion etching (RIE) except from the contact areas, and then, hydrogen reduction of WF_6 follows. The W deposition and etching and deposition (dep-etch-dep) technique (see Chapter 10) is also used for the W plug. The TiW barrier layer for W deposition is generally not used because of end point detection problem, as TiW also contains W.

Titanium Deposition

The Ti metal can be deposited at a temperature of 400°C by the ion sputtering technique or the Electron Cyclotron Resonance (ECR) technique. The ECR-enhanced CVD process shows good results in case of the submicron device, but Ti deposited by the CVD process suffers from electronegative and thermodynamical due to unstable precursor.

Copper Deposition

In early days, copper was not used because of its problems of etching and having fast reaction with oxygen. Copper has the advantages of less resistivity than aluminium, and no electromigration. The oxidation of copper is checked by the damascene process (see Chapter 10), as depicted in Fig. 10.7. Better quality, low resistivity of 2 μΩ-cm (bulk resistivity 1.9 μΩ-cm), and reasonable deposition rate of copper can be obtained by the hydrogen reduction of Cu(hfac) in the CVD in the temperature range of 350°C to 450°C, as

$$Cu(hfac)_2 + H_2 \rightarrow Cu + 2H(hfac)$$

Copper deposition for electrical wiring by electroless plating is under study. To deposit copper, $CuSO_4$ and formaldehyde chemicals are used and their chemical reaction is written below.

$$Cu^{2+} + 2HCHO + 4OH^- \rightarrow Cu^0 + 2HCOO^- + 2H_2O + H_2$$

Prior to copper plating, a Ti film is usually deposited to make a conducting wafer surface.

Nickel Deposition

It has been found that nickel deposits in the submicron contact and plugs very well. The electroless deposition of nickel using nickel sulphate by hypophosphite reduction is mentioned below:

$$Ni^{2+} + (H_2PO_2)^- + H_2O \rightarrow Ni + 2H^2 + H(HPO_3)^-$$

Till now, metal plating (or electroless plating) has not been applied in IC processing, but it may prove to be a good option in the near future because it is an easy deposition and low-temperature process. Furthermore, thicker metals can be deposited with less stress. Presently, it has the disadvantage of contamination, but future researches may overcome this disadvantage.

Silicide/Polysilicide and Nitride Film Deposition

Silicide and nitride deposition is required for (a) promoting good electrical connection of devices, (b) hole plugging for electrical connection, and (c) barrier layer for electromigration. To obtain a silicide/polysilicide film, generally, the Ti, Co, Ta, and W metals are used. The resistivity of $TiSi_2$ and CoSi can be obtained as low as 10 to 15 μΩ-cm. If the thickness of silicide is reduced, one can get an even lower resistive film. These silicide films are especially used for the gate polysilicon, the source and the drain contacts. Prior to Ti (or Co) metal deposition, the dielectric layer(s) from the gate, source, and drain is removed. Apart from the low contact resistance, silicide is used for the self-align process, as shown in Fig. 11.7. Presently, these materials are used for the submicron device application, but they have reached their limits of thickness variation and thermal stability. It is very unlikely that these silicides can be used for deep submicron device. Therefore, in place of Ti and Co, some other suitable material has to be explored.

The deposition of these materials is done either by the CVD or by the sputtering technique. In this section, silicide and nitride deposition is discussed.

TiN Deposition

TiN is widely used as a barrier for aluminium electromigration, to check the reaction of W with silicon and for better electrical contact with silicon. TiN can be deposited by the reactive sputtering technique, and the CVD technique.

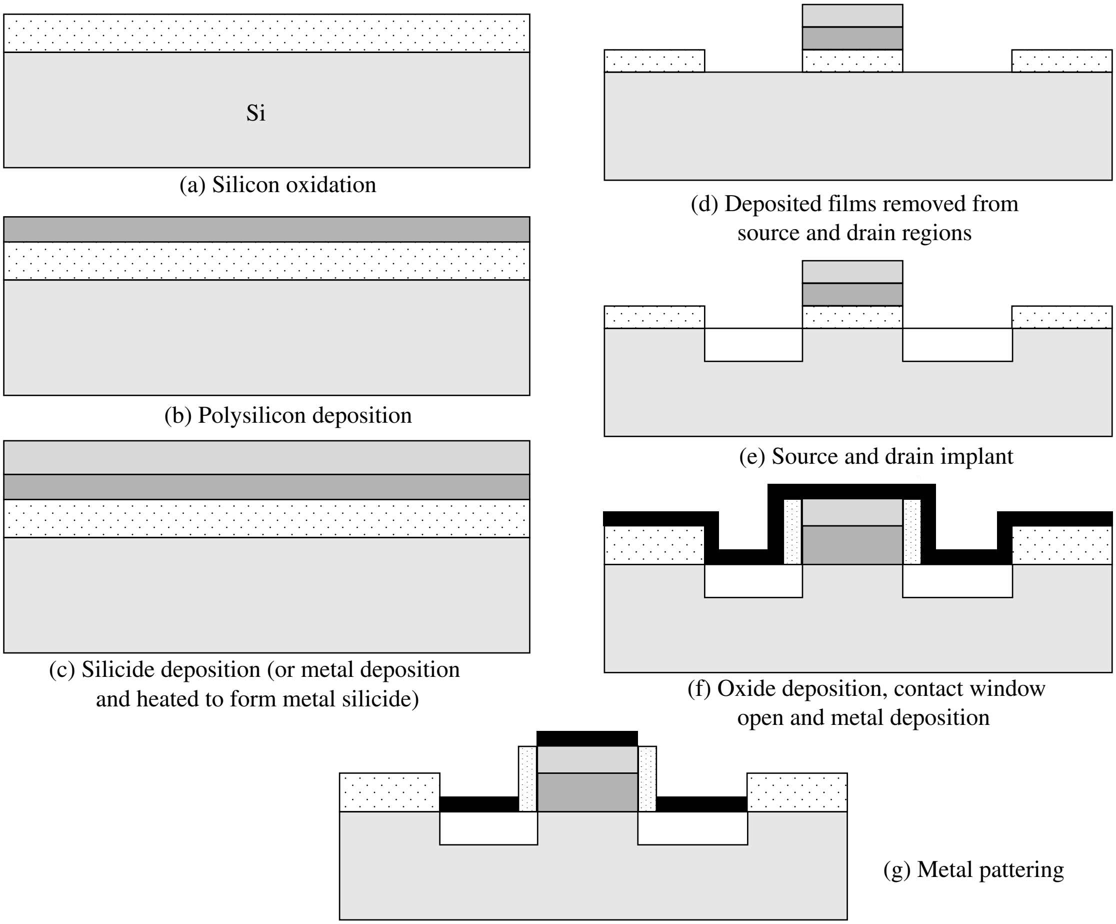
In the RF reactive sputtering, Ti target is sputtered in the presence of nitrogen gas at the chamber pressure of 10^{-2} to 10^{-3} Torr. The sputtered Ti atoms react with nitrogen and deposit on the wafer in the form of TiN. Sputtering of the TiN target is another option for TiN deposition.

Generally, TiN is deposited by the molecular-organic chemical vapour deposition (MOCVD) technique. It has better step coverage than the collimated sputtering technique. The following precursors and gases are used for TiN deposition.

$$6TiCl_4 + 8NH_3 \rightarrow 6TiN + 24\ HCl + N_2 \text{(ammonia reduction)}$$

$$2TiCl_4 + 2NH_3 + H_2 \rightarrow 2TiN + 8HCl \quad \text{(ammonia reduction)}$$

$$2TiCl_4 + N_2 + 4H_2 \rightarrow 2TiN + 8HCl \quad (H_2 \text{ and } N_2 \text{ reduction})$$



(a) Silicon oxidation

(b) Polysilicon deposition

(c) Silicide deposition (or metal deposition and heated to form metal silicide)

(d) Deposited films removed from source and drain regions

(e) Source and drain implant

(f) Oxide deposition, contact window open and metal deposition

(g) Metal pattering

Fig. 11.7 Self-align silicide process

TiN deposition using ammonia reduction is done in the temperature range of 400°C to 700°C; whereas H_2 and N_2 reduction is done above 700°C temperature. For a higher TiN deposition rate, around 0.5% of chlorine is incorporated in the TiN film; whereas, for the ammonia gas reduction, around 5% of chlorine is incorporated in the TiN film. The incorporated Cl reacts with H_2O and forms HCl that in turn reacts with the metal and corrodes it.

TiN can also be deposited at low temperature without chlorinating, but by using metal-organic precursors such as tetrakis (dimethylamido)-Ti (TDMAT) or tetrakis(diethylamido)-Ti (TDEAT) mixed with ammonia gas (NH_3), as

$$6Ti[N(CH_3)_2]_4 + 8NH_3 \rightarrow 6TiN + 24HN(CH_3)_2 + N_2 \text{ (below 450°C)}$$

The TiN deposited with metal-organic precursor has low density, is unstable, has high resistivity due to the incorporation of carbon and oxygen, and has less step coverage in high aspect ratio. A better step coverage is reported by the TDEAT precursor with low resistivity (180 μΩ-cm) against the resistivity of 500 μΩ-cm obtained by the TDMAT precursor.

Plasma-assisted TiN deposition using NH_3 reduction at a temperature of around 550°C has been explored. By this deposition technique, TiN contains 1% of Cl. A good-quality TiN film with 40 μΩ-cm resistivity can be deposited by the Electron Cyclotron Resonance (ECR) technique using the $TiCl_4$ precursor. TiN can be deposited by the ECR technique with the TDMAT precursor at lower temperature (<400°C) with resistivity ranging from 200 to 300 μΩ-cm. TiN does not make good contact with silicon that is heavily doped with boron (n^+) or phosphorous (p^+) and that results in a high resistance. To overcome this problem, around 40 nm Ti is deposited prior to the deposition of TiN.

$TiSi_2$ Deposition

Generally, Ti is sputtered from the Ti target and then heated in RTP in the temperature range of 620°C to 680°C and, thereafter, RTP temperature is set to around 750°C in the nitrogen environment. The chemical transformations are mentioned below.

$$Ti + 2Si \rightarrow TiSi_2 \text{ (C-49)}$$

$$TiSi_2 \text{ (C-49)} \rightarrow TiSi_2 \text{ (C-54)}$$

Similarly, Co can be used in place of Ti. The Co deposition process is almost the same as the Ti deposition process.

$TiSi_2$ deposited by CVD is not widely accepted, though it has 3 to 4 times lower resistivity than WSi_2 and $TaSi_2$. This is because $TiSi_2$ is not a good material for the self-align process. $TiSi_2$ can be deposited using $TiCl_4$ in the presence of hydrogen and silane reduction at 650°C to 700°C in the LPCVD reactor.

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

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

$TaSi_2$ Deposition

Generally, $TaSi_2$ is deposited by the standard sputtering process. $TaSi_2$ can also be deposited by the CVD techniques.

In the LPCVD technique, the Ta_5Si_3 precursor reacts with bare silicon (or polysilicon) and forms $TaSi_2$, as mentioned below:

$$Ta_5Si_3 + 7Si \rightarrow 5TaSi_2$$

$TaSi_2$ can also be deposited by the $TaCl_5$ precursor by the same process. Deposition of $TaSi_2$ by the LPCVD process is not uniform and develops notches that in turn damage the gate oxide. For this reason, $TaSi_2$ is only deposited by the sputtering technique, especially in the gate area. Reaction of $TaCl_5$ with silicon is given below:

$$4TaCl_5 + 13Si \rightarrow 4TaSi_2 + 5SiCl_4$$

Another precursor $TaCl_5$ mixed with dichlorosilane can be used for the deposition of $TaSi_2$ at 650°C in LPCVD, but damages the gate oxide in a similar way as described previously. Reaction of the gases is given below.

$$TaCl_5 + 2SiH_2Cl_2 + 2.5H_2 \rightarrow TaSi_2 + 9HCl$$

The other method by which $TaSi_2$ deposition can be done, is by the silane reduction of $TaCl_5$ at a temperature of 600°C using the CVD technique as given below:

$$5TaCl_5 + 3SiH_4 + 6.5H_2 \rightarrow Ta_5Si_3 + 25HCl$$

This process of $TaSi_2$ deposition is popular.

It is essential to mention that if a proper recipe is not chosen then the quality of the $TaSi_2$ film will be impacted and its properties will be different.

$TaSi_2$ and $TiSi_2$ Silicides and Displacement Reaction

One of the purposes of silicide deposition is to bring low contact between the metal and especially, source and drain contacts. It has been found that silicide precursors react faster with silicon and transform into silicide, as

$$4TaCl_5 + 13Si \rightarrow 4TaSi_2 + 5SiCl_4$$

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

Once a thin silicide film is formed, the precursor diffuses through it and reacts with silicon at the silicide and silicon interface. This process is called **displacement reaction**. As the silicide film thickness increases, the diffusion of the precursor decreases, and at some thickness, the silicide formation stops. In the silicide displacement reaction, silicon atoms also migrate inside the silicide film and create defects below it. In recent days, the junction depth of the source and the drain is about 0.15 µm. Out of this junction depth, $TaCl_5$ and $TiCl_4$ consume 150 to 250 nm and 300 to 600 nm of silicon respectively. The amount of silicon consumption depends on the CVD process parameters. The consumption of silicon leads to serious problems of junction damage, leakage current, and dopant profile. In addition, during silicide formation, defects are created in the silicon that enhance the junction depth of the source and the drain. One option to control the dopant profile is the implantation of the source and the drain after silicide formation. Till today, silicide is used for IC fabrication, but in the near future, a new scheme like raised source and drain MOS structure may be developed.

WSi_2

Both the CVD and sputtering techniques are extensively useful for the deposition of WSi_2. Low-resistance WSi_2 polycide is made on top of the polysilicon gate to reduce the gate resistance, and is also found useful in the submicron device fabrication. In the CVD process, silane reduction of WF_6 in the temperature range of 300°C to 400°C is carried out. Generally, the WF_6/SiH_4 ratio is kept more than 10 to get a good WSi_2 film. The chemical reaction that takes place is given below:

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

The other method of WSi_2 deposition is by dichlorosilane (SiH_2Cl_2) reduction, at a temperature range of 500°C to 600°C, which has 4 to 5 times faster deposition rate and better step coverage. The reaction of WSi_2 with SiH_2Cl_2 is given below.

$$WF_6 + 3.5SiH_2Cl_2 \rightarrow WSi_2 + 1.5SiF_4 + 7HCl$$

11.3 LITHOGRAPHY

The basics of lithography have been described in Chapter 6. The main criteria of lithography is the resolution, and that depends mainly on (a) the exposure wavelength, (b) the exposure system, (c) the numerical aperture, (d) the pattern placement, (e) the photoresist (resist) resolution, (f) depth of focus of the exposure system, (g) the modulation of wafer surface, and many other factors.

There are three types of exposure techniques used in submicron device fabrication, namely, the optical exposure technique, the electron-beam exposure technique, and the X-ray exposure technique. Out of these three exposure techniques, the optical exposure technique and the electron-beam exposure technique are in use extensively. The X-ray exposure technique is not popular mainly due to fragile mask and complicated mask fabrication.

11.3.1 Optical Lithography

Optical lithography is the oldest among all the lithography techniques, but still used for the submicron device fabrication extensively. The optical exposure system is simple, low-cost, easy to maintain, has high field of view and high wafer throughput. There are three modes of optical exposure systems available, namely, (a) the contact mode, (b) the proximity mode, and (c) the projection mode. The working principle, merits, and demerits of all these modes of the optical lithography system have been mentioned in Chapter 6. In the contact mode, the mask and the wafer are made in contact with each other by applying pressure ranging from 0.05 atm to 0.3 atm. The mode that applies less contact pressure is called the **soft contact mode lithography** and the mode that applies high contact pressure is called the **hard contact mode lithography**. The contact mode lithography has resolution better than 0.5 μm with the I-line (λ = 365 nm) of mercury source. Resolution can be further enhanced, if the exposure wavelength is decreased. The contact mode lithography introduces defects both in the mask and in the wafer during every cycle of contact. In addition, resolution variation occurs due to wafer surface modulation and wafer bow. Despite all these disadvantages, the contact mode lithography is used for submicron devices till today.

The disadvantages of contact mode lithography can be overcome if the mask and the wafer are kept apart during exposure, but at the cost of resolution. This mode of lithography is called the **proximity mode lithography**. Usually, the separation (gap) between the mask and the wafer is kept a few micrometres. This mode of lithography is hardly used in submicron devices because of poor resolution.

Presently, the optical projection mode lithography is extensively used in submicron devices. In this mode of lithography, the mask is kept at the field plane and the wafer is placed at the image plane, as shown in Fig. 6.13(b). In projection lithography, the numerical aperture (NA) is high and it plays an important role in resolution. The numerical aperture is the solid angle that is formed from the last lens of the projection system to the focal point on the image plane, as shown in Fig. 6.7. If the numerical aperture is small then the depth of focus is large and hence, higher resolution can be obtained. On the other hand, if the resolution is high and the depth of focus is large then the numerical aperture is short (see Chapter 6). Therefore, the numerical aperture and the depth of focus is optimised for better resolution. This is because the depth of focus cannot be made too small due to the distance required between the lens and the wafer. The distance between the lens and the wafer is called the **working distance** of the projection lens.

Three types of optical projection systems are available in the market. These are the reflection optical projection system, the refraction optical projection system, and the catadioptric projection system. The principle of reflection projection system is based on the reflection of light from the mask patterns onto the wafer.

In the refraction projection system, an image of the mask is made on the wafer using the principle of optical refraction. The principle of refraction optical projection system is more or less the same as that of the step and repeat camera. In this projection system, the reticle is generally used for lithography. The size of the reticle may be 1 X, 4 X, or 5 X; where X denotes the magnified size of the actual die. The catadioptric system is based on the combination of the reflection and the refraction projection systems. Using the catadioptric system, one can obtain resolution up to 0.4 μm.

11.3.2 Lithography Resolution

Lithography resolution mainly depends on the resolution of the optical system, the photoresist (or resist) resolution, the mechanical system, and the exposure system. The resolution also depends on other factors, such as the temperature, the composition of the developer, and the wafer stirring during developing.

Optical Resolution

Well-designed projection systems with I-line of high-pressure mercury lamp, has resolution up to 0.4 μm. Resolution can be further increased up to 0.3 nm with 248 nm wavelength of the excimer laser light source (KrF laser). In addition, resolution can be further increased by mask modification, called **mask engineering**. When the transparent and the opaque patterns of the mask are illuminated, light gets diffracted from the edges of the transparent patterns. This diffracted light significantly alters the light distribution on the PR and that reduces the resolution, as shown in Fig. 11.8. To nullify the diffracted light, an alternative transparent pattern of mask is coated with a transparent film. The thickness of the transparent film is chosen such that the light coming out from the transparent film has 180° phase difference with respect to the light coming out from the uncoated adjacent transparent patterns. When these two lights fall on the PR, they overlap each other and cancel out, as shown in Fig. 11.8. The thickness (t) of the transparent film can be determined by the equation

$$t = \frac{\lambda}{2(n-1)}$$

where n is the refractive index of the film and λ is the wavelength of the light.

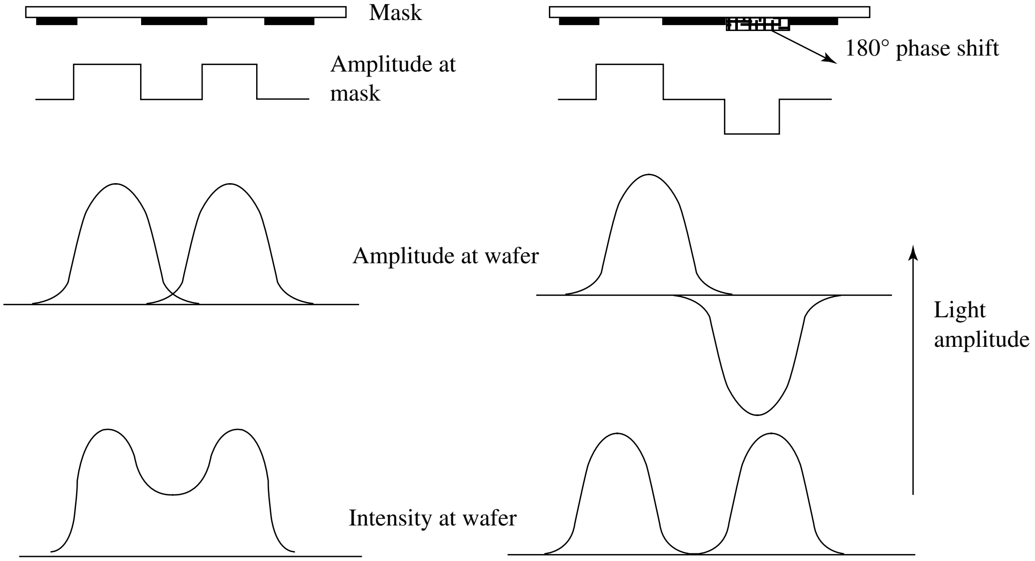


Fig. 11.8 Phase-shift techniques on mask to improve resolution (mask engineering)

Photoresist Resolution

The resolution also depends on the contrast of the photoresist and the Modulation Transfer Function (MTF) of the optical projection system. In the first case, resolution depends on the material property of the PR and in the second case, it depends on the imperfection in the image quality produced by an optical system. PR contrast is related to the wavelength of light, the pre-bake and post-bake temperature, the developer temperature, concentration of the developer, and the underlying material. The PR contrast is experimentally determined by exposing the light energy and simultaneously developing in steps while measuring the reduction of the PR thickness keeping all other lithographic parameters constant. This tells how the PR responds to the intensity of light.

For negative PR, the contrast is defined as

$$\gamma = \frac{1}{\log_{10} \dfrac{Q_f}{Q_0}}$$

and for positive PR,

$$\gamma = \frac{1}{\log_{10} \dfrac{Q_0}{Q_f}}$$

where Q_f is last light dose that is required to expose the PR completely, and Q_0 is the first light energy dose given to the photoresist. Q_f can be obtained by drawing a tangent at E_T (total light dose from Q_0 to Q_f) that meets the point where 100% photoresist is removed as shown in Fig. 11.9. E_T is the dose of light energy that is sufficient to expose a whole layer of (thickness) the PR and it is called the **threshold dose of light energy**. A typical graph of the positive PR, after removal of the PR, with steps of exposure and development is shown in Fig. 11.9. Similarly, a typical graph of the negative PR after removal of the PR and the PR profile with thickness depth. One can see that the higher the contrast, the better is the resolution of the PR.

The Modulation Transfer Function (MTF) or the Critical MTF (CMTF) defines the optical property and the imperfection of the lens projection system. The Modulation Transfer Function (MTF) defines the light contrast in the aerial image that falls on the PR. MTF of an optical system can be expressed as

$$\text{MTF} = (I_{\text{Max}} - I_{\text{Min}})/(I_{\text{Max}} + I_{\text{Min}})$$

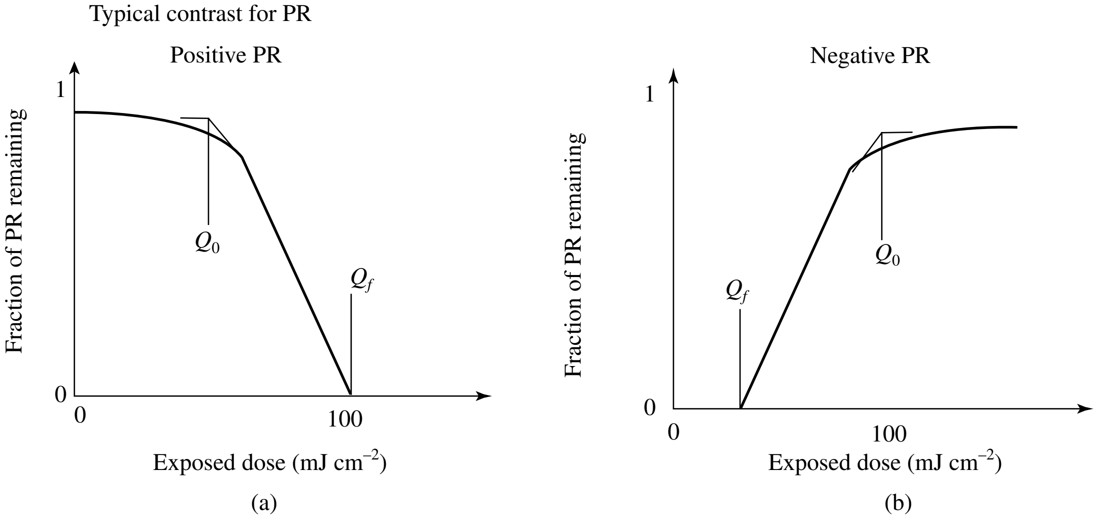


Fig. 11.9 Typical light sensitivity curves

where I_{Max} and I_{Min} are the maximum and the minimum light intensities in the image respectively. The image formed by the projection system does not give the ideal representation of the mask in terms of the light distribution (on the PR), when light passes through the mask. The diffraction of light and the imperfection of the optical system both distort the ideal representation of the mask. To get the maximum possible high resolution, an ideal image of the mask is needed. To get an ideal image representation of the mask, the optical system has to be corrected for the diffraction of light. This type of optical system is called the **diffraction limited** optical system. The imperfection of the optical system cannot be corrected beyond a certain point. The Critical MTF (CMTF) can be mathematically expressed as

$$\text{CMTF} = \frac{Q_f - Q_0}{Q_f + Q_0}$$

$$\text{CMTF} = \frac{10^{1/\gamma} - 1}{10^{1/\gamma} + 1}$$

where Q_0 is the first dose of light exposure and Q_f is the last dose of light exposure. It is found that a typical CMTF value for a g-line (or i-line) is around 0.4.

Generally, the PR is exposed to monochromatic light to get a good resolution, but when the wafer is exposed, a part of light reflects back from the PR/silicon interface and interferes with the incoming light and manifests standing interference, parallel to the wafer surface and perpendicular to the PR thickness. The light intensity of the constructive (maxima) interference is much higher than the destructive (minima) interference, as shown in Fig. 11.10. When the PR is developed, it is removed relatively more from the constructive interference position than the destructive interference position. This creates modulation in the side wall of the PR that decreases the resolution. This phenomenon is more prominent, when the PR is thick. Therefore, a thin PR coating is done for ULSI device fabrication.

In the submicron device technology, the alignment of mask patterns with the wafer is more crucial. The alignment inaccuracy is introduced mainly due to personal errors, exposure system, mask aligner stability, vibration of the alignment machine and floor vibration, and inaccuracy of lithography alignment; all these factors bring down the resolution.

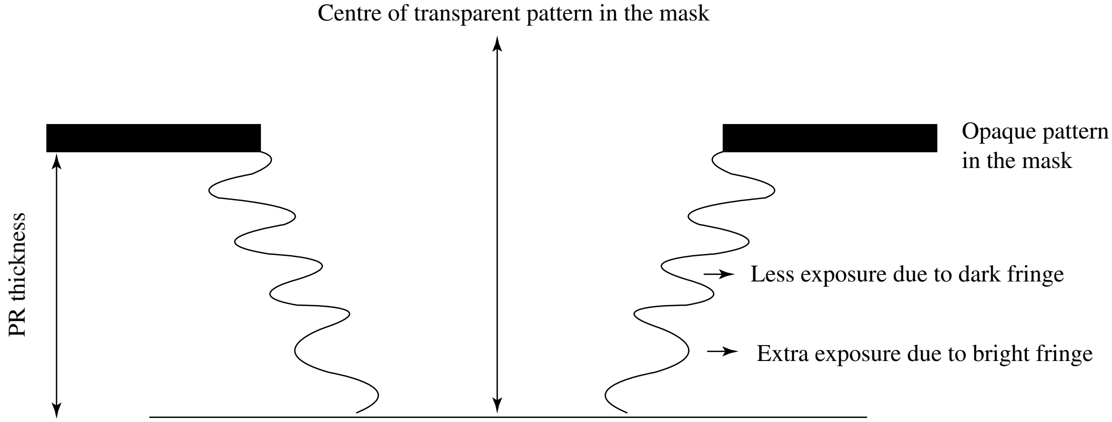

Centre of transparent pattern in the mask
Opaque pattern in the mask
PR thickness
Less exposure due to dark fringe
Extra exposure due to bright fringe

Fig. 11.10 A typical exaggerated light interference effects within the positive PR

11.3.3 Electron-Beam Lithography

The fundamentals of the electron beam equipment and the lithography process are described in Chapter 6. In this section, electron-beam lithography in the context of submicron lithography is covered in brief.

Resolution increases when the exposure wavelength decreases and the numerical aperture increases (Eq. 6.1). One can get the electron wavelength below 1 Angstrom, an electron beam spot size less than 0.1 μm and a numerical aperture around 0.01 with 10–15 keV. As compared to optical lithography, the e-beam wavelength is around 400 times shorter than the I-line and the numerical aperture is 50,000 times smaller than 100 X of the optical numerical aperture. For this reason, the electron beam lithography is extremely useful in ULSI lithography.

Electron-beam lithography is used for many purposes like optical and X-ray mask fabrication and direct patterning on the resist. The latter application is called **Direct Writing on Wafer**. The writing on the wafer is extremely useful for application-specific integrated circuits (ASICs), where a large number of devices are not in demand. This technique of lithography saves both fabrication cost and time significantly.

There are two modes of e-beam exposure, namely, raster scanning mode and vector scanning mode. These modes of exposure are described Chapter 6. In both these modes, the exposure time is significantly longer and this lowers the wafer throughput. In order to increase the wafer throughput and reduce the cost, proximity and projection e-beam lithography systems are coming up. In the proximity e-beam lithography, the mask die is exposed to a beam of 1 mm diameter; then, the wafer is translated to an appropriate position and then, the next die is exposed. This technique is similar to the step and repeat camera used for mask fabrication (see Chapter 5). In the e-beam projection lithography, the mask is made of opaque and transparent patterns made out of a silicon membrane by etching a bulk silicon wafer. The opaque patterns of the silicon membrane are attached to the bulk silicon in a few places for mechanical support (which are unwanted), as shown in Fig. 11.11. The resist that is unexposed due to the mechanical supports is exposed by making another identical mask where the supports are attached at different locations and then the wafer is exposed for the second time. Generally, the distance between the mask and the wafer is kept around 0.5 mm. A typical electron proximity system is sketched in Fig. 11.11. This equipment is not yet fully developed but constant research is going on to improve the system.

High resolution and high throughput can be obtained by a 1:1 electron projection system. This projection system is still in the development stage. In this projection system, a quartz plate is coated with a chrome metal film that provides the opaque patterns and the transparent patterns are coated with CsI, as shown in Fig. 11.12. Once the wafer alignment is completed, the mask is illuminated with UV light from behind the coated mask. Intense UV light interacts with the CsI material and generates electrons. A very high electric field is applied between the mask (cathode) and the wafer. This electric field accelerates the electrons towards the wafer and then, the electrons expose the resist. This technique has the disadvantage of short life span of the CsI film, but this technique works well for better throughput with high resolution.

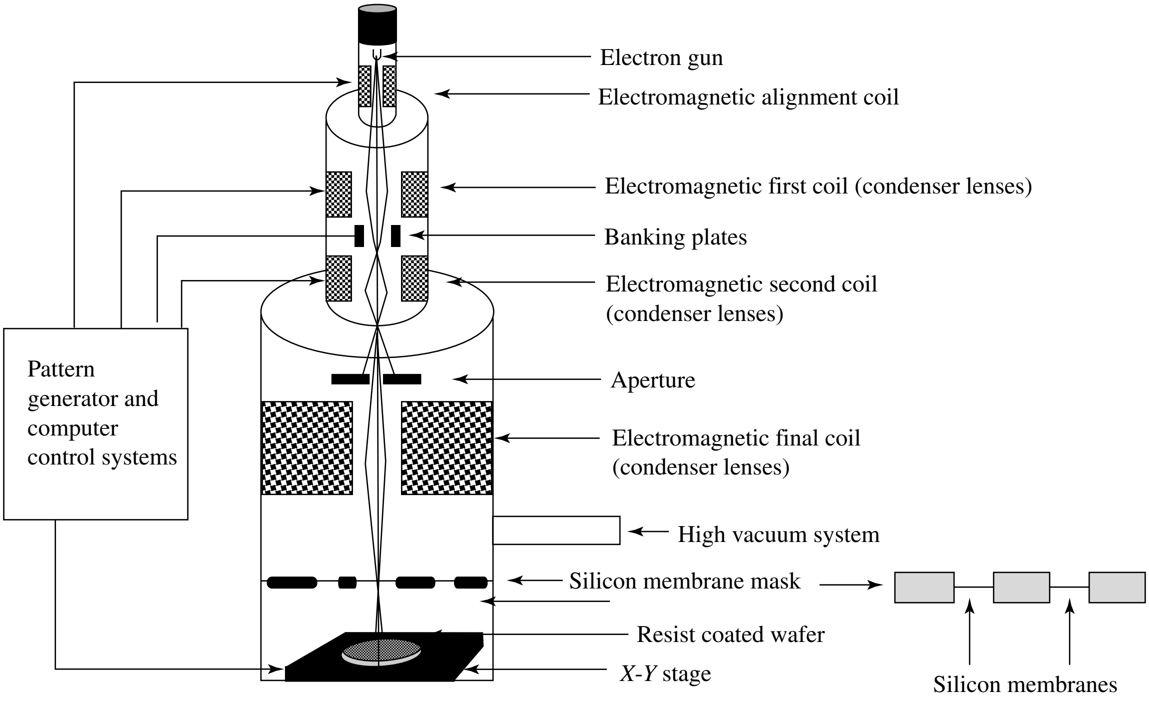
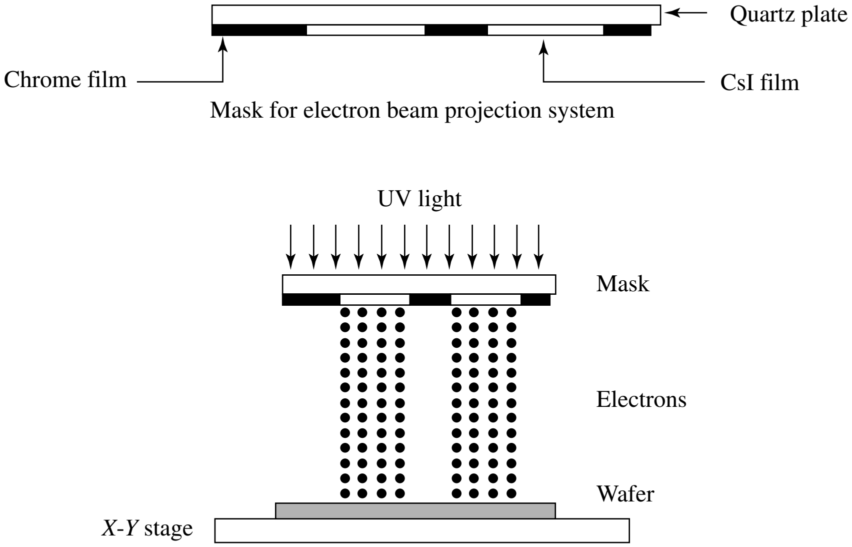

Electron gun
Electromagnetic alignment coil
Electromagnetic first coil (condenser lenses)
Banking plates
Electromagnetic second coil (condenser lenses)
Pattern generator and computer control systems
Aperture
Electromagnetic final coil (condenser lenses)
High vacuum system
Silicon membrane mask
Resist coated wafer
X-Y stage
Silicon membranes

Fig. 11.11 Electron-beam lithography system



Mask for electron beam projection system

Fig. 11.12 1:1 Electron wafer exposure projection system

In another version, the electron beam lithography is done through a stencil mask as shown in Fig. 11.13. The stencil mask contains several transparent apertures of different shapes. The mask is placed between the electron beam and the wafer, while exposure electrons are passed through the stencil mask patterns and the wafer is exposed. The size and positioning of the image are controlled by the electrostatic (electron) lenses. This version of electron beam lithography is in the development stage.

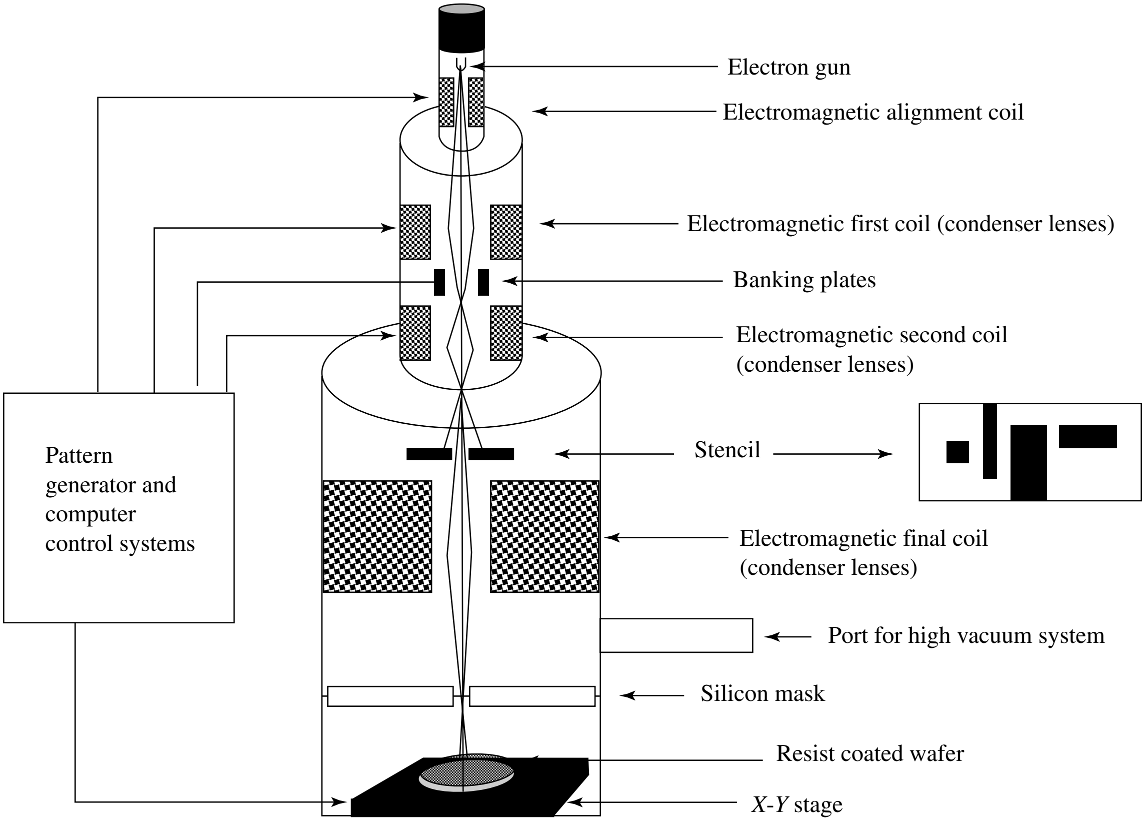

Electron gun
Electromagnetic alignment coil
Electromagnetic first coil (condenser lenses)
Banking plates
Electromagnetic second coil (condenser lenses)
Pattern generator and computer control systems
Stencil
Electromagnetic final coil (condenser lenses)
Port for high vacuum system
Silicon mask
Resist coated wafer
X-Y stage

Fig. 11.13 Electron-beam lithography system

11.3.4 Electron Resist Resolution

The resolution of electron lithography is found to be around 0.1 μm, but the overall resolution is much less than 0.1 μm as it has the total effect of the exposure wavelength, the numerical aperture, the mechanical stability of the machine, the alignment accuracy, the mask alignment stability, the vibration of the alignment machine, the floor vibration, and the lithography process, etc. In addition, the resolution further degrades because of two more reasons. The first reason is the swelling of the resist (negative resist) while developing, similar to the negative photoresist that affects the resolution in two ways: (a) loss of adhesion with the wafer and undercut, and (b) closely spaced patterns swell and come closer to each other and prevent etching between the patterns. The second reason for losing resist resolution is the scattering of the electrons in the resist. When high energy electrons enter the resist, they collide with the resist molecules and deviate in different directions as shown in Fig. 11.14. In addition, it is found that the isolated patterns need more exposure intensity than the denser patterns; for example, an isolated 0.5 µm pattern needs around 25% more exposure than an identical dense pattern. This problem is solved by exposing the isolated patterns for a longer time to the electron beam as compared to the denser patterns. This is not possible in either optical or X-ray lithography. Scattering of the electrons depends mainly on the electron beam energy, the thickness of the resist and the material below the resist. If the thickness of the resist is reduced then scattering is reduced.

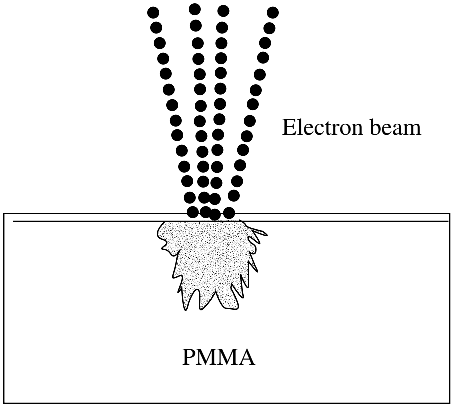


Fig. 11.14 Electron trajectory inside the PMMA

11.3.5 X-ray Lithography

To increase the lithography resolution, a shorter wavelength is imperative. For this reason, optical lithography UV and deep UV wavelengths are used in the optical range. Beyond the deep UV wavelength, all materials including glass become opaque, whereas, many materials are transparent to X-ray wavelength. In addition, X-ray is even transparent to contaminations such as organic material and dust, and that reduces the defects in X-ray lithography. Furthermore, X-ray has the advantages of small wavelength, large depth of focus, high resolution (around 0.2 µm) and placement accuracy (around 0.03 µm). On the other hand, X-ray has the disadvantages of complex mask fabrication, high mask fragility and the need for mask engineering. The X-ray mask is made of Au film patterns on the silicon membrane. The X-ray mask process is shown in Fig. 11.15. An X-ray is opaque to gold film patterns; but, it is transparent to the thin silicon membrane.

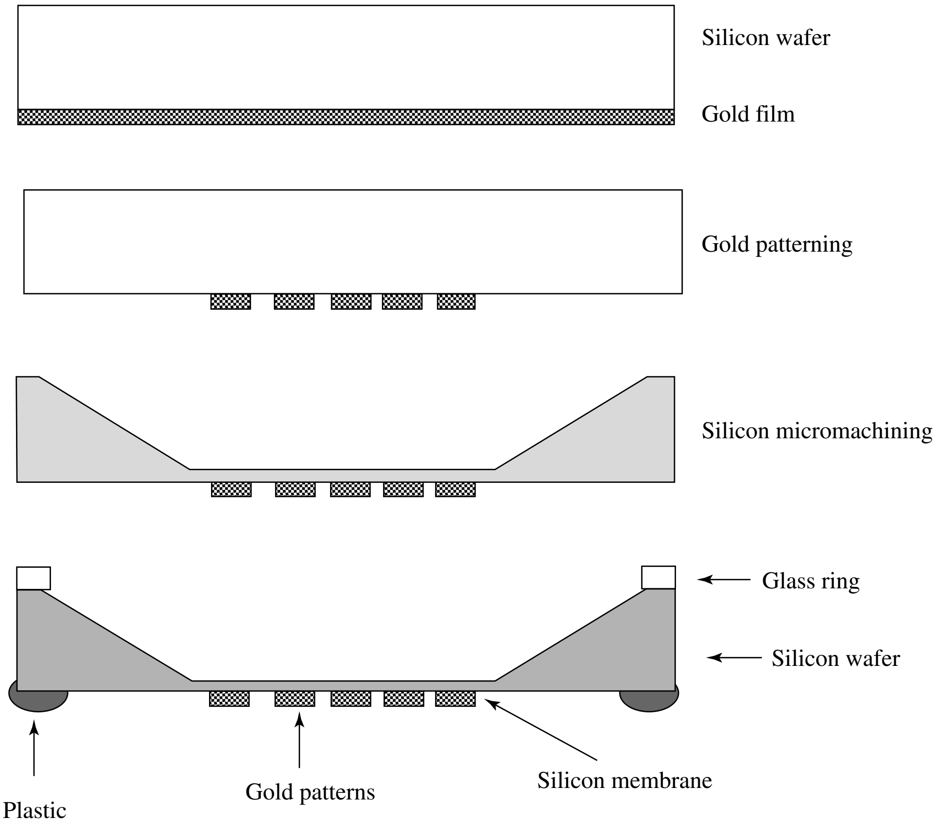


Fig. 11.15 X-ray mask fabrication sequence

An X-ray has a wavelength in the order of a few Angstroms. An X-ray is generated from a point source and this introduces a deviation in R by ∂R, and as R increases, ∂R also increases as shown in Fig. 11.16. In addition, there is an increase of Δ due to the pattern away from the X-ray point source, as depicted in Fig. 11.16. These deviations from the actual pattern size can be expressed as

$$\partial R = \frac{Sg}{D} \text{ and,}$$

$$\Delta = g\frac{R}{D}$$

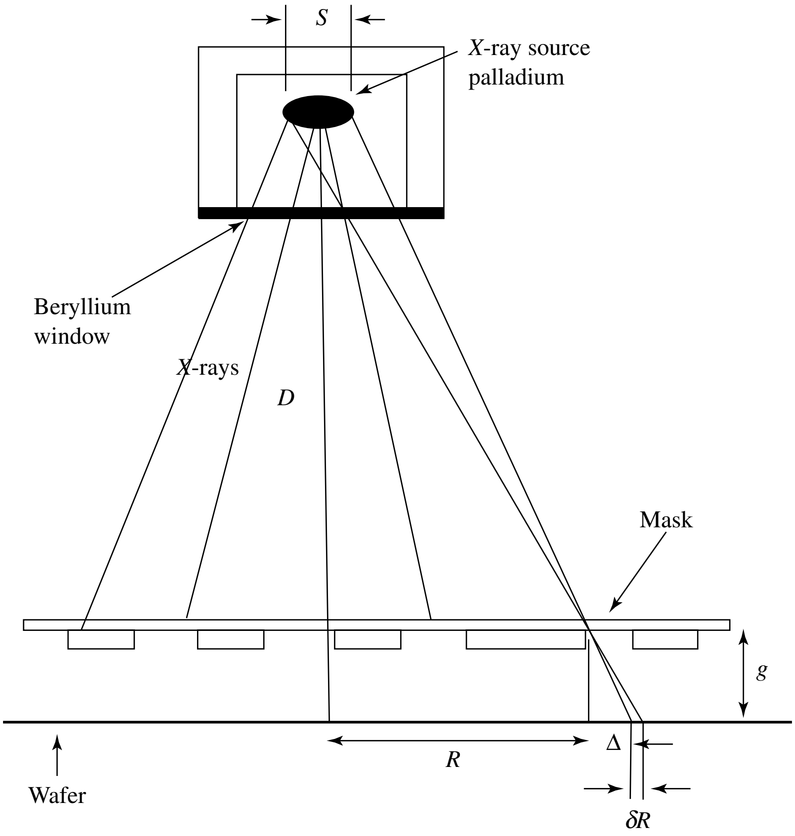


Fig. 11.16 Error in X-ray lithography

where *R* is the distance between the X-ray point source perpendicular to the mask and a particular pattern of the mask, *δR* is the shadow casted on the wafer due to the gap between the mask and the wafer at a particular pattern *R*, *S* is the source size, *g* is the gap between the mask and the wafer, *D* is the distance between the point source and the wafer. This variation is corrected by mask engineering during mask fabrication, but it is a complicated process.

To overcome these disadvantages of point source X-ray, projection mode X-ray lithography has been developed. In this technique, the full mask is illuminated by the extended X-ray source and patterns of the mask are projected on the wafer by a combination of convex and concave mirrors. Generally, laser-induced plasma is used to produce soft X-ray, which has low radiation effect on the wafer. This projection X-ray system is in the development stage.

11.3.6 X-ray Mask

X-ray mask fabrication is a difficult task. The process sequence of X-ray mask fabrication is depicted in Fig. 11.15. One of the sides of the polished silicon wafer is coated with gold and thereafter, gold patterns are made by the lithography process. Once the gold patterns are made, the other side of the silicon wafer is etched till a thin silicon membrane is left. The opaque Au film patterns over the silicon membrane block the X-ray, whereas, the rest of the thin silicon membrane is transparent to the X-ray. Thereafter, the mask is protected from both the sides for handling and fixing into the mask holder.

11.4 ETCHING

Etching becomes very critical in the domain of ULSI device fabrication because of the requirement of a uniform and good etching selective. Uniform etching encounters two major issues: (a) Aspect Ratio-Dependent Etching (ARDE), and (b) pattern density variation. ARDE phenomenon occurs due to isotropic etching and it becomes an even more serious issue for deep etching like trench isolation etching. It is well known that the etch rate for closely placed (high density) patterns is different from that for isolated patterns of identical sizes. This phenomenon is called **micro-loading**.

There are two techniques used for etching, namely, wet etching and dry etching. The fundamental of etching is described in Chapter 7. These two etching techniques have their merits and demerits, especially in the range of submicron patterns.

11.4.1 Wet Etching

Wet etching has the main advantages of high selectivity, uniformity, ease of the process, low cost, and high throughput. But, wet etching cannot be used below 1.5 μm due to the restriction of the etchant, and the movement of the reacted product from the etching surface. As the patterns are becoming denser, the problem with wet etching is aggravating further. In addition, the etch rates for bigger patterns are different from those of the smaller patterns, and the etch rates for denser patterns are different from those for isolated patterns. Furthermore, wet etching suffers from pattern dimensional fidelity due to isotropic etching. These disadvantages restrict the use of wet etching in ULSI device fabrication. At present, ULSI device fabrication is done using dry etching. Dry etching has significant issues related to deep etching and micro-loading effect, especially for denser patterns. In addition, dry etching is approaching its limit below 0.5 μm. Hence, for deep submicron etching, dry etching needs to be improvised or a new etching technique needs to be devised. This limitation of dry etching has instigated researchers to explore the possibility of using wet etching with a new chemistry.

Generally, wet etching of any material is heterogeneous with a complicated chemistry. As explained in Chapter 7, ions, or molecules of an etchant react with the material atoms and form a complex product (or products). This product either dissolves in the etchant or reacts with one of the etchant constituents. The reacted product then dissolves in the etchant solution. There are two techniques used for wet etching: (a) emersion, (dip) technique, and (b) spray technique. Among these two etching techniques, the spray technique is more suitable because the reactant product is forcefully removed away from the etching surface by the pressure of the spray. In addition, the spray technique requires lesser amount of etchant and the etching rate is higher than that of the dip etching technique.

11.4.2 Silicon Etching

Generally, silicon etching is done by HF-HNO_3 solution. Water or acetic acid is used for dilution to get an appropriate etch rate and better etch control. Silicon first converts into silicon dioxide and then, silicon dioxide is dissolved in the solution. The chemical reaction of silicon with HF-HNO_3 can be written as given below:

$$Si + HNO_3 + 6HF \rightarrow H_2SiF_6 + HNO_2 + H_2O + H_2$$

Presently, silicon is etched by the dry etching process. Generally, CF_4 precursor gas is used for silicon etching, but the etch rate of silicon and silicon dioxide is almost the same. In order to increase the etch rate of silicon with respect to SiO_2, around 12% oxygen is added into the CF_4 gas.

11.4.3 Silicon Dioxide Etching

Silicon dioxide etching is done either by using liquid HF or gaseous (vapour) HF. In the liquid wet etching, the wafer is dipped in the HF and etching is done on the basis of time duration, keeping all other etching parameters such as temperature, agitation and concentration constant. To maintain a constant etchant rate, the fluoride ions are kept constant (to maintain the pH value) by adding the NH_4F chemical. Recently, it has been found that silicon dioxide can be etched for submicron patterns using the spray etching technique, if the side etching walls are protected (passivated) by hydrogen atoms. The etch reaction of silicon dioxide and HF vapour is given below.

$$SiO_2 + 6HF \rightarrow H_2SiF_6 + H_2O$$
$$H_2SiF_6 + H_2O \rightarrow SiF_4 + 2HF$$

Presently, the dry etching technique is used for submicron silicon dioxide etching. For deep submicron device fabrication, a faster oxide etching technique has to be developed.

11.4.4 Polysilicon Etching

Polysilicon etching is done by concentrated KOH chemical at around 80°C using the dip etching method. After polysilicon etching, the wafer is treated with HCl to remove the potassium ions from the wafer completely. One of the advantages of KOH etching is the high selectivity between polysilicon and silicon dioxide.

11.4.5 Silicon Nitride Etching

Silicon nitride can be etched in HF at room temperature, but there is hardly any selectivity between silicon nitride and silicon dioxide. In fact, HF is widely used for silicon dioxide etching. Heating in 85% H_3PO_4 at around 180°C is another option for silicon nitride etching, but at 180°C temperature; hence the PR cannot be used as a mask. The advantage of H_3PO_4 etching is the high selectivity between silicon nitride and silicon dioxide. For example, the etch rate of CVD deposited silicon nitride is around 100 Å per minute, whereas, the etch rate of thermally grown silicon dioxide is in the order of a few angstroms. Though the etch rate of silicon nitride is very low, it is the best option for the global removal of silicon nitride from the wafer. At present, nitride is etched by the dry etching technique.

11.4.6 Dry Etching

The physics of low-pressure discharge (plasma), the dry etching equipment and the etching processes are described in Chapter 7. Presently, dry etching is widely used for submicron etching, but it needs improvements for deep submicron etching. Dry etching is a single wafer process; hence, the wafer throughput is less. However, processing a larger diameter single wafer is the trend in IC fabrication.

Dry Etching Chemistry, Selectivity and Etch Profile

The dry etching technique can be divided into three categories: (a) the physical etching technique, (b) the reactive ion etching technique, and (c) the physical-reactive ion etching technique. In the physical etching technique, highly energetic argon ions are generated in the plasma and these energetic ions are made to strike the wafer surface that physically knock out (sputter) the surface material of the wafer (see Chapter 10). In the reactive ion etching technique, highly reactive radicals are generated in the plasma and these radicals react with the material and produce a volatile product, which is pumped out instantaneously from the etching chamber. In the third dry etching technique (also known as chemical-physical etching), both physical and reactive ion etching techniques are used to etch out the material.

To fabricate the ULSI device, high etch rate, high selectivity and perfect anisotropic etching are imperative. Till today, the chemistry of dry etching is not well understood, so the perfect etching model is not available; therefore, etching optimisation is done by experimentation; which is a time-consuming and costly affair.

In the past, stable nontoxic CF_4 precursor gas was used to etch silicon and its compounds. In the plasma, CF_4 produces reactive fluorine atoms and these atoms react with the silicon (or its compound) and produce a volatile product as written below:

$$SiO_2 + 4F \rightarrow SiF_4 + O_2$$

The etch rate of silicon and silicon dioxide increases by 12% and 20% respectively if O_2 is added into the CF_4 gas; but, the etch rate of silicon and silicon dioxide decreases if the oxygen is further increased. This is because, the excess O_2 reacts with CF_4 and produces COF_2, CO, and CO_2, and this decreases the CF_4 concentration.

It is found that the silicon etch rate decreases significantly when H_2 is added to CF_4, but the etch rate of silicon dioxide does not change much. The reasons may be: (a) H_2 reacts with F and forms HF that reduces the concentration of F, (b) CFx (when x is higher than 4) reacts with SiO_2 and forms SiF_4, CO, CO_2 and other products, and (c) CF_4 reacts with H_2 and forms either a carbon or hydrocarbon polymer, which passivates the silicon surface and reduces the F concentration and that in turn reduces the silicon etch rate. The overall reaction of CF_4 with silicon dioxide and silicon is mentioned below:

$$CF_4 + SiO_2 \rightarrow SiF_4 + CO + CO_2 + COF_2$$

$$CF_4 + Si \rightarrow SiF_4 + C + CF_y + CH_xF_y$$

Apart from CF_4, CHF_3 and a combination of (CHF + CF_4) are also used for silicon dioxide etching.

Selectivity is defined as **the etch rate ratio of different materials**. In the ULSI device, the MOS gate is in the order of a few nanometres; hence, the requirement of etching selectivity becomes stringent. For example, the ratio of etching selectivity between polysilicon and gate oxide should be more than 10:1. On the other hand, the electrical connection of the source and the drain needs deep silicon dioxide via (hole) etching; hence, a high selectivity between silicon dioxide and polysilicon is needed.

Etching selectivity depends on two factors: (a) etching should be thermodynamically favourable for the reaction between the etchant and the etch material, and (b) the reactant product should be volatile in nature. For instance, in the first case, the reaction of chlorine gas with polysilicon is more thermodynamically favourable than silicon dioxide. Hence, the etch rate of polysilicon is much higher than that of silicon dioxide. In the second case, the reactant product should be volatile and immediately exhausted from the etching system. Otherwise, the reactant product may be adsorbed (or may stay) on the material surface and block the further etching process. Many a time, this phenomenon is utilised when one wants to protect the sides of the etching walls. For example, the etch side walls gets passivated by the fluoride atoms while silicon dioxide etching is done in CF_4.

The etch profile (the etch slope with respect to the etch depth) is very important in a ULSI device, especially for high aspect ratio (vertical to horizontal etch ratio). In principle, the etching should be anisotropic, but this never happens in reality. In the dry etching process, a majority of the ions strike vertically, but a significant number of ions also strike from oblique angles (off axis), and this results in horizontal etching (undercut) as shown in Fig. 11.17. In order to reduce the horizontal etching, the etching system is designed in such a way that the generation of oblique-angled ions is minimised. There are also techniques which are used to make the ions strike perpendicularly, but they have their own limitations (see Chapter 7). In deep etching, in order to protect from undercut, another mechanism is chosen. In this mechanism, the sputter material or the reactant product deposits on the side wall and protects from etching. For example, when silicon is etched by reactive ion etching using HBr, the side walls are deposited with the reactant product, $SiBr_xH_y$.

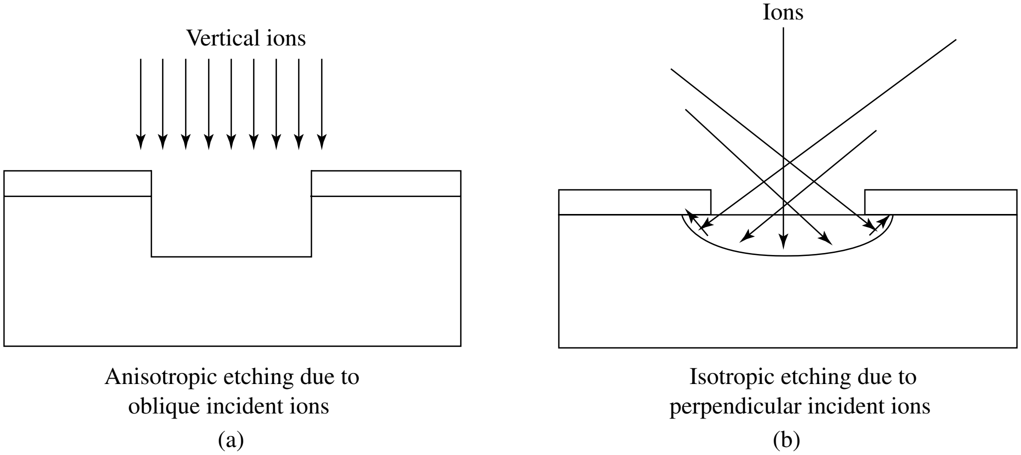

Vertical ions
Ions
Anisotropic etching due to oblique incident ions
(a)
Isotropic etching due to perpendicular incident ions
(b)

Fig. 11.17 (a) anisotropic etching due to oblique incident ions, and (b) Isotropic due to vertical ions strike

Side wall etching is a function of the aspect ratio; at a higher aspect ratio, the side wall (undercut) etching is more, whereas, vertical etching is less. In addition, the etch profile also depends on micro-loading, and it becomes worse when the aspect ratio is high. This aspect-ratio-dependent etching is abbreviated as **ARDE**.

In the parallel plate reactive ion etching system, highly energetic ions strike the wafer and damage it. This wafer damage is a serious issue in a ULSI device. In order to reduce the wafer damage, Electron Cyclotron Resonance Plasma Etcher (ECR) and Inductively Coupled and Helicon Wave RF Plasma Etcher have been developed. In the ECR system, microwave energy is used to create the plasma and to increase the plasma density, a magnetic field is applied. The magnetic force makes the electrons to rotate in a circular path at a particular frequency called the **cyclotron frequency**. When this cyclotron frequency is tuned to the microwave frequency, a resonance occurs, and high plasma density is created. The ECR system is designed in such a way that a large number of electrons move away from the plasma zone towards the low magnetic field. While travelling, the electrons collide with the atoms and produce low energetic ions, which then strike the wafer as shown in Fig. 11.18. The wafer is negatively biased either by high dc voltage or RF voltage in order to attract ions. The ion striking velocity can be further minimised by tuning the wafer bias voltage; hence, the wafer damage can be further reduced. This technique is most suitable for deep submicron devices, but the use of ECR may be restricted because the system is most complex and its cost is exceptionally high.

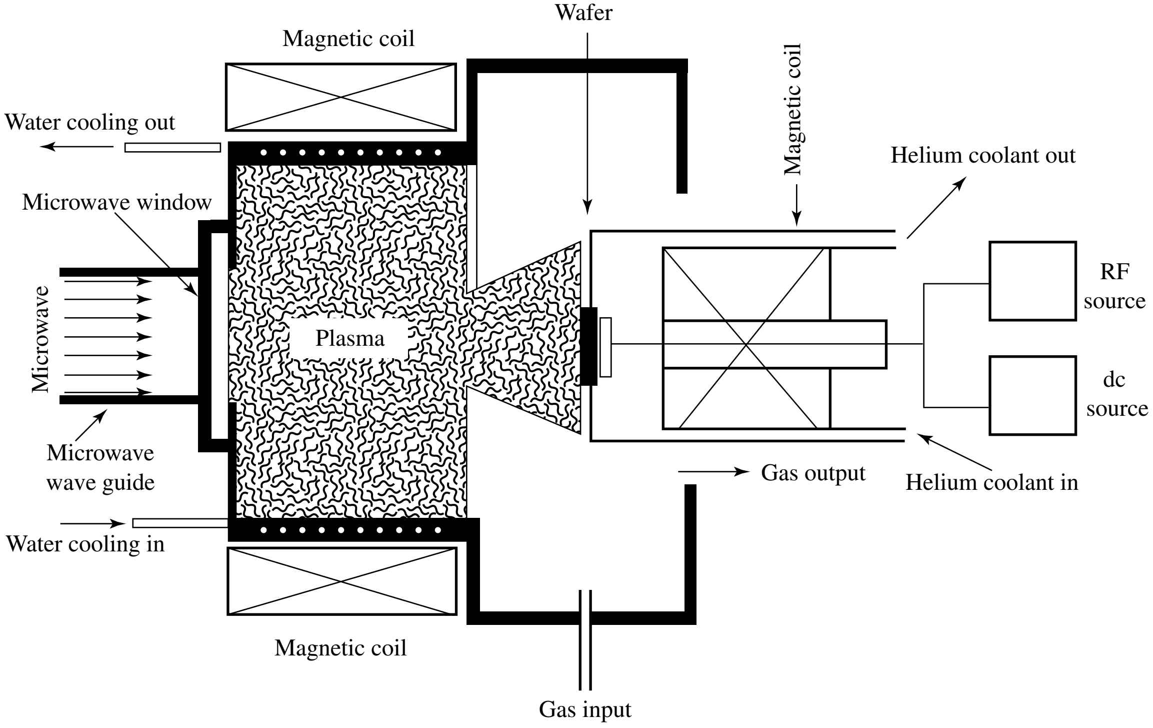


Fig. 11.18 Typical electron cyclotron resonance etcher

The other option to reduce the wafer damage is the **inductive coupled etching or the helicon wave RF plasma etching technique**. In both the techniques, the wafer is decoupled from the plasma source so that only the low energetic ions are allowed to strike the wafer. These techniques can be used even for deep submicron etching. An inductive coupled plasma etcher is produced by passing an RF current through an induction coil located outside the plasma chamber, as shown in Fig. 11.19. In order to increase the electron path, magnets are placed outside the plasma chamber. In order to attract ions, the wafer is ac biased, and is several skin depths away from the coil to avoid the loss of plasma density and the electromagnetic effect.

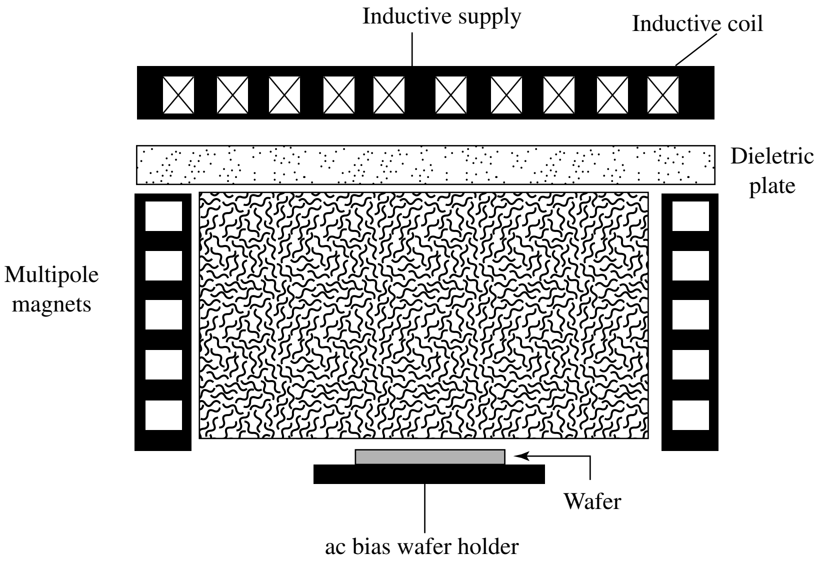


Fig. 11.19 Typical inductive couple plasma reactor

In the helicon wave RF plasma etcher system, dense plasma is created inside the plasma chamber by applying an RF frequency to the loop antennae, which are placed outside the plasma chamber. In order to make the electron travel in the helicon path in the plasma region, a longitudinal magnetic field is applied, as shown in Fig. 11.20. When the helicon wave is the same as the antenna length, a resonance occurs and that in turn creates the high plasma density. The electrically biased wafer is placed a little away so that only the low energetic ions strike the wafer.

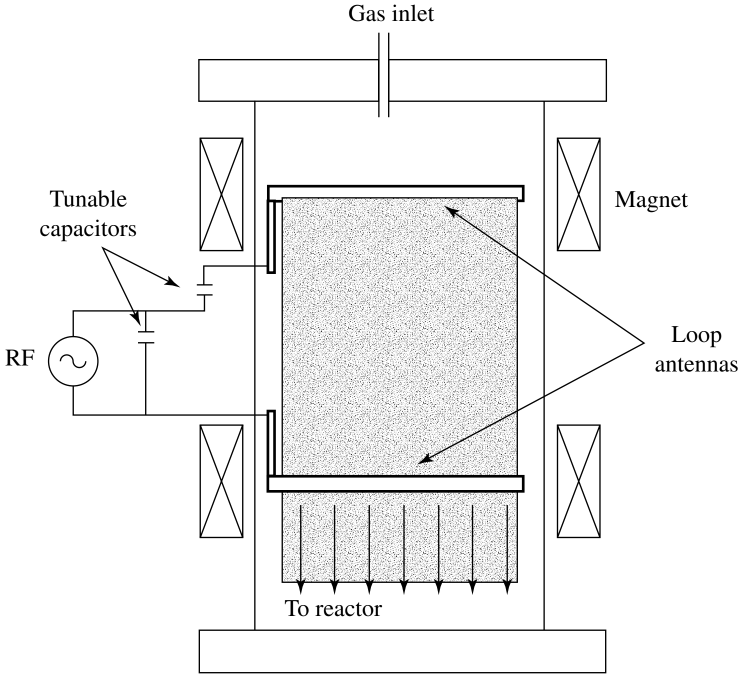


Fig. 11.20 Typical helicon wave plasma reactor

11.4.7 Silicon Etching

To make a trench in the silicon for electrical device isolation and storage capacitor for the memory cell, perfect vertical etching is needed. For submicron electrical device isolation, a shallow trench around 1 μm is sufficient, but for capacitor fabrication, a deep trench (> 5 μm) is usually required. Once the trench is made, it is filled with silicon dioxide insulator.

Generally, for the etching of shallow trench, the fluorocarbon gas is used. This gas has a high silicon etch rate whereas the etch rate of SiO_2 is low, although it etches the silicon horizontally (undercut). However, it works well for the shallow submicron device isolation. For deep trench, bromide-based etching is much more suitable as it has less undercut and higher selectivity between silicon and silicon dioxide as compared to the fluorocarbon and chlorine-based gases. In deep trench etching, the sidewall is protected either by a carbon containing gas or by a carbon free precursor.

11.4.8 Polysilicon Etching

The gate of the MOS transistor is made of polysilicon material; hence, dimensional integrity is very important. Any change in the gate dimensions will reflect on the MOS transistor characteristics. To enhance the electrical conductivity between the metal and the polysilicon, a metal silicide film is deposited over the polysilicon. For this reason, polysilicon etching must have high selectivity with respect to gate oxide and metal silicide. These criteria cannot be obtained by the ion-enhance etching process. Presently, high etching selectivity and multistep etching process is being used. In this process, different etch steps are optimised for the desired requirements.

Generally, a bromide precursor such as HBr is used for polysilicon etching, because bromide etching has a high etch rate as well as high selectivity between the polysilicon and the oxide. Prior to polysilicon etching, the native polysilicon oxide is removed by the plasma in the presence of a fluorine-based precursor.

11.4.9 Silicon Dioxide Etching

Silicon dioxide etching is required for the diffusion window, the metal contact window, interlayer deposition, and may other purposes. It has been found that plasma etching in CHF_3 and CF_4 gases has high selectivity between silicon dioxide and silicon. This is because the carbon polymer deposit passivates the silicon surface and that restricts etching. Once the etching process is complete, the deposited carbon is removed and cleaned thoroughly.

11.4.10 Nitride Etching

Silicon nitride is used for the LOCOS process, isolation between the metal layers, IC passivation, and sidewall protection for the polysilicon gate. Silicon nitride is generally deposited on silicon dioxide. Silicon nitride can be etched either by CF_4 or a mixture of oxygen and CF_4 gases in the plasma; but the etching selectivity between silicon nitride and silicon dioxide is very poor. Fortunately, silicon nitride can be etched anisotropically with high selectivity by O_2 mixed with CH_2F_2 or CHF_3 gases in the RIE mode. It is not understood till today why these two gases result in good selectivity, and not CF_4. One of the reasons seems to be the deficiency of the fluorocarbon polymer.

11.4.11 Metal Etching

Aluminium and tungsten metals are extensively used in IC fabrication. Aluminium is used for the gate electrode and electrical circuit wiring, and tungsten is generally used for hole (via) filling (plugging) for the contact.

Aluminium Etching

Normally, aluminium is etched by a chlorine-based gas in the plasma. The etch rate of the chlorine-based gas is high, but chlorine also etches horizontally to a great extent. The chlorine-based aluminium etching is good for devices of the order of microns, but not suitable for ULSI device fabrication. It is important that after the aluminium etching, chlorine is removed thoroughly with deionised water immediately; otherwise, chlorine will react with water and produce HCl and with time, HCl will corrode the aluminium metal. The fluorine-based etching is not suitable, as its reactant product AlF_3 has low vapour pressure. Therefore, to etch the aluminium, researchers are exploring the use of bromine based etching. It is well known that aluminium surface gets oxidised and forms Al_2O_3; hence, prior to aluminium etching, Al_2O_3 is etched by the ion sputtering process.

Tungsten Etching

A uniform etching of tungsten is obtained by the PCVD process. A chlorine-based precursor reacts with the tungsten film and forms a volatile product. There is good selectivity between tungsten and silicon dioxide. The fluorine-based precursor also makes a volatile product with tungsten, but the selectivity between tungsten and silicon dioxide is much poorer. It has been found that when tungsten is etched with a high etch rate, a recess (dip) is formed in the plug. Generally, tungsten etching is preferred in two steps. In the first step, a uniform tungsten is etched with a high etch rate to get higher throughput, and in the second step, slow etching is carried out to avoid dip formation. The tungsten metal works as a barrier layer for TiN, where the selectivity between tungsten and TiN is taken care of at the time of tungsten etching. If the photoresist mask is used for tungsten etching, then SF_6 or Cl_2 gases are added with N_2 for good selectivity.

Summary

To fabricate a ULSI device, parameters like control of drain junction breakdown, punch through, short-channel effects and hot electron generation, current leakage, MOS transistor V_T fluctuation, contact resistance, parasitic resistance and capacitance, and the mismatch of MOS transistors threshold voltage and other device parameters are important. The requirement of these parameters has led to the evolution of low thermal budget Rapid Thermal Processing (RTP) equipment. Though the RTP (and RTA) equipment has the limitation of single wafer processing that results in low wafer throughput, it is the trend of today's technology. Different derivatives of the RTP system are used for different process applications and are named after their specific fabrication process steps, for example, if the RTP equipment is used for oxidation, it is called Rapid Thermal Oxidation (RTO) and when it is used for CVD, then the system is called Rapid Thermal Chemical Vapour Deposition (RTCVD). In the RTO chamber, the annealing process is done immediately after oxidation, hence, this process is abbreviated as RTO-RTA. Similarly, RTCVD followed by RTA annealing is called RTCVD-RTA. The challenging problems in the RTP and RTA processes are uniform wafer heating and temperature measurement over a large wafer diameter.

Apart from thermal budget, other process-related issues, such as lithography resolution, exposure systems, etching, film deposition in holes, multilevel wiring, contact, etc., in ULSI device fabrication are still to be overcome. The low thermal budget RTP processes used for silicon oxidation, film deposition of different dielectrics, epitaxial film deposition, silicon nitride film deposition, film deposition of different silicides, aluminium and tungsten film deposition, lithography, and etching specially in view of the ULSI (nano) fabrication technology is discussed. Most of the film depositions have been discussed in Chapter 10, but in this chapter, the merit and demerits of the deposition techniques have been discussed and the recent technological developments in the ULSI generation have been covered.

References

- C Y Chang and S M Sze (Ed); *ULSI Technology*, McGraw-Hill Companies Inc, 1996
- J D Plummer, M Deal and P B Griffin; *Silicon Fundamental Technology: Fundamentals, Practice and Modeling,* Prentice Hall, 2000
- S Wolf; *Silicon Processing for the VLSI Era*, Vol 4: Deep-Submicron Process Technology, Lattice Press, First Edition, 2002
- S M Sze; *VLSI Technology*, Second Edition, McGraw-Hill, 1988
- S K Gandhi; VLSI Fabrication Principles, Second Edition, Wiley, 1994
- S A Campbell; The Science and Engineering of Microelectronic Fabrication, Oxford University Press, 1996

Index