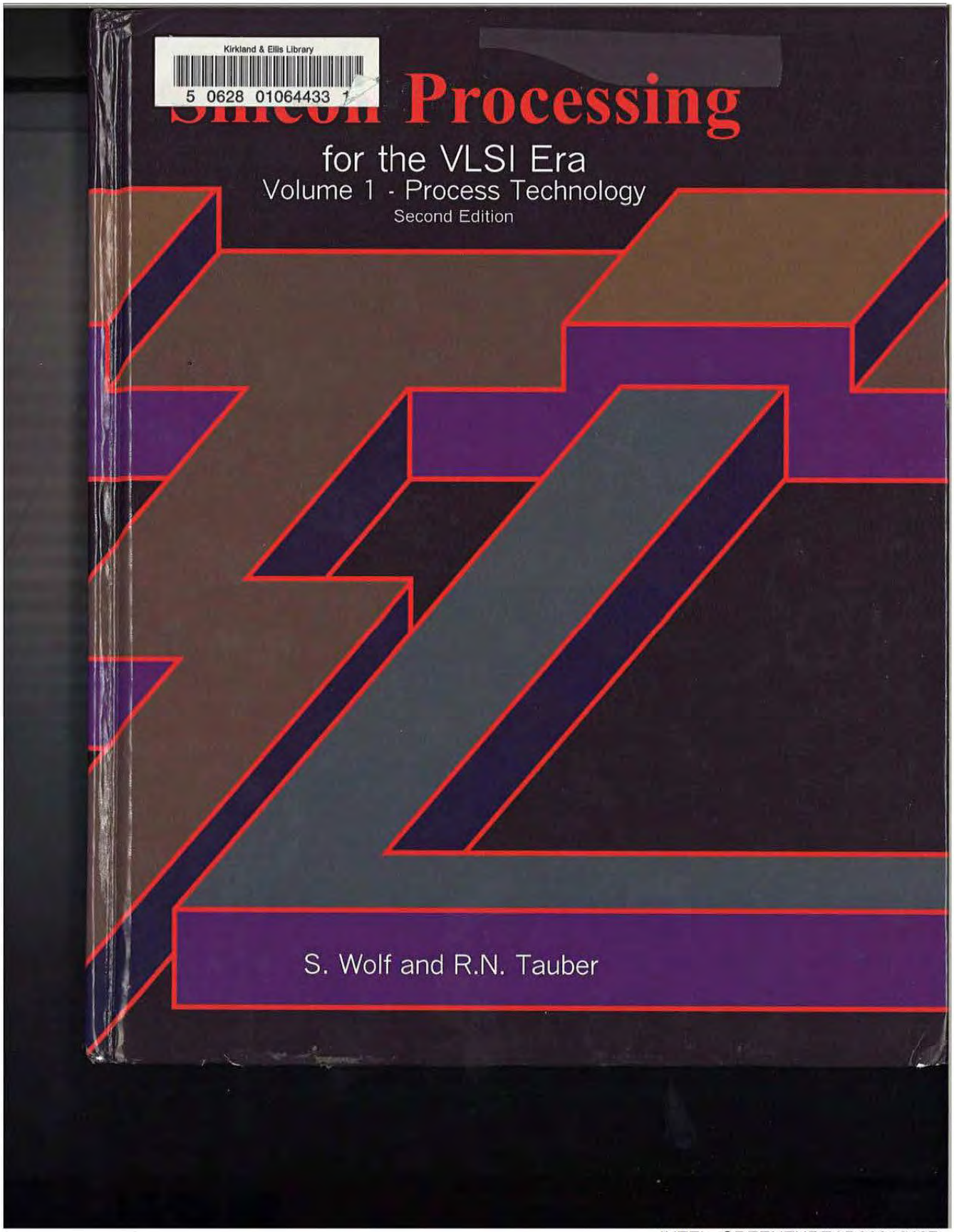


INTEL_GREENTHREAD00016467

SILICON PROCESSING
FOR
THE VLSI ERA

VOLUME 1:
PROCESS TECHNOLOGY
Second Edition

STANLEY WOLF Ph.D.

RICHARD N. TAUBER Ph.D.

LATTICE PRESS

Sunset Beach, California

DISCLAIMER

This publication is based on sources and information believed to be reliable, but the authors and Lattice Press disclaim any warranty or liability based on or relating to the contents of this publication.

Published by:

LATTICE PRESS
Post Office Box 340
Sunset Beach, California 90742, U.S.A.

Cover design by Roy Montibon, New Archetype Publishing, Los Angeles, CA.

Library of Congress Cataloging in Publication Data
Wolf, Stanley and Tauber, Richard N.

Silicon Processing for the VLSI Era
Volume 1: Process Technology

Includes Index
1. Integrated circuits-Very large scale integration. 2. Silicon. I. Title

ISBN 0-9616721-6-1

9 8 7 6 5 4 3 2

PRINTED IN THE UNITED STATES OF AMERICA

DETAILED TABLE OF CONTENTS

S 149

[redacted]

PREFACE
SECOND EDITION

This second edition of *Silicon Processing for the VLSI Era: Volume 1 - Process Technology* follows the 1986 publication of the first edition. More than 20,000 volumes of that version were sold, attesting to its widespread acceptance throughout the microelectronics community. Professionals and students of this vibrant technology should find the new edition equally useful.

Many new processes and materials have been incorporated into IC fabrication in the fifteen years since the first edition was written. The new materials and processes covered in the second edition include: 300-mm wafers, DUV lithography, chemically-amplified resists, high-energy ion implantation, high-density plasma sources for CVD and etching, step-and-scan aligners, chemical-mechanical polishing, dual-damascene interconnects, copper metallization, and low-*k* dielectrics. Altogether the material of our book comprises more than 900 pages, 600 illustrations, and 1500 references (with over half of the citations from 1996 to 1999). It also includes those topics from the first edition that are still in use.

The effort of writing and producing this new edition was predominantly carried out by Stanley Wolf. He revised eight chapters (Chaps. 1, 2, 3, 4, 6, 11, 12, and 14), and entirely wrote the five new ones (Chaps. 5, 13, 15, 16, and 17). Three chapters were revised by Richard N. Tauber (Chaps. 7, 8, and 9), and Michael Current revised Chap. 10.

Valuable input to Stanley Wolf was also provided by many persons. Foremost of these were R. N. Tauber and C. A. Wolf. They each carefully read the entire manuscript and offered extensive editorial assistance (Tauber: primarily technical, and Wolf: grammatical). Others who provided important technical assistance and critical reviews were: Jerry Healey, Richard Cohen, Chris Mack, Bruce Smith, Robert Simonton, Dennis Hess, Moshe Preil, Wilbur Krussell, Bob Climo, Kathleen Perry, Peter Loewenhardt, and Donald Smith. Roy Montebin, of Vox Mundi, designed the book's cover.

Stanley Wolf, Ph.D.
Professor Emeritus, Department of Electrical Engineering, CSULB

I would also like to acknowledge the support of my former colleagues at Applied Materials for their helpful discussions and for kindly providing some of the figures used herein: Chris Fulmer, Gary Miner, Gary Xing, Kelly Truman, Majeed Foad, Norma Riley, Parvin Katebi, and Satheesh Kuppurao.

In addition, I would like to thank the following persons for providing additional background material and figures used in this textbook: Al Geise - SVG/Thermco; Clyve Hayzelden - KLA/Tencor; Gerhardt Kneissal - ADE; Jim Cable - Peregrine Semiconductor; Pat Shanks - Semitool; Sang-Shin Kim - GaSonics; and Terry Ma - Avant!/TMA.

Richard N. Tauber, Ph.D.
Consultant to the Microelectronics Industry
Clovis, California

PREFACE TO THE FIRST EDITION

SILICON PROCESSING FOR THE VLSI ERA is a text designed to provide a comprehensive and up-to-date treatment of this important and rapidly changing field. The text consists of three volumes of which this book is the first, subtitled, *Process Technology*. Volume 2, subtitled, *Process Integration* was published in 1990. Volume 3, subtitled *The Submicron MOSFET* was published in 1995. In this first volume, the individual processes employed in the fabrication of silicon VLSI circuits are covered in depth (e.g., epitaxial growth, chemical vapor and physical vapor deposition of amorphous and polycrystalline films, thermal oxidation of silicon, diffusion, ion implantation, microlithography, and etching processes). In addition, chapters are also provided on technical subjects that are common to many of the individual processes, such as vacuum technology, the properties of thin films, and CMOS process integration. The topics covered in the book are listed in more detail in the Table of Contents.

The purpose of writing this text was to provide professionals involved in the microelectronics industry with a single source that offers a complete overview of the technology associated with the manufacture of silicon integrated circuits. Most other texts on the subject are available only in the form of specialized books (i.e., that treat just a small subset of all of the processes), or in the form of edited volumes (i.e., books in which a group of authors each contribute a small portion of the contents). Such edited volumes typically suffer from a lack of unity in the presented material from chapter-to-chapter, as well as an uneven writing style and level of presentation. In addition, in multi-disciplinary fields (such as microelectronic fabrication), it is difficult for most readers to follow technical arguments in such books, especially if the information is presented without defining each technical "buzzword" as it is first introduced. In our book, we hope to overcome such drawbacks by treating the subject of VLSI fabrication from a unified and more pedagogical viewpoint, and by being careful to define technical terms when they are first used. The result is intended to be a user friendly book for workers who have come to the semiconductor industry after having been trained in but one of the many traditional technical disciplines.

An important technical breakthrough has occurred in book publishing that the authors also felt could be exploited in creating a unique book on silicon processing. That is, revolutionary electronic publishing techniques became available in the mid-1980's, and these can cut the time required to produce a published book from a finished manuscript. The task traditionally took 15-18 months, but can be now reduced to less than 4 months. If traditional techniques are used to produce books in such fast-breaking fields as VLSI fabrication, these books automatically possess a built-in obsolescence, even upon being first published. The authors took advantage of these rapid production techniques, and were able to successfully meet the reduced production-time schedule. As a result, it was possible to include information contained in technical journals and conferences within four months of the book's publication date. Most other books on silicon processing undergo such 15-18 month production cycles. This is the first book on the subject where such time-delay effects have been eliminated from the production process!

Written for the professional, the book belongs on the shelf of workers in several microelectronic disciplines. *Microelectronic fabrication engineers* who seek to develop a more complete perspective of the subject, or who are new to the field, will find it invaluable. *integrated circuit design enegineers, test engineers,* and *integrated circuit equipment engineers,* who must understand VLSI processing issues to effectively interface with the fabrication environment, will also find it a uniquely useful reference. The book should also be very suitable as a text for graduate-level courses on silicon processing techniques, offered to students of electrical engineering, applied physics, and materials science. Problems are inclnded at the end of each chapter to assist readers in gauging how well they have assimilated the material in the text.

The book was an outgrowth of an intensive seminar of the same title as the book, conducted by the authors through the Engineering Extension of the University of California, Berkeley. In the fifteen years that this course has been conducted (1984-1999), several thousand engineers and managers from more than 100 companies have enrolled in it. Its fine reputation is attested to by the fact that many firms have sent participants to subsequent courses, presumably based on the recommendations of earlier participants.

In setting out to create a comprehensive text on VLSI fabrication, the authors each contributed a set of unique and complementary skills to the project. Professor Wolf"s proficiency as a teacher and writer were utilized to produce a clearly written and logically organized book. Some of this expertise was gained in authoring an earlier best-selling text *Electronic Measurements and Laboratory Practice,* Prentice-Hall, 1973. Dr. Tauber brought a technical expertise acquired from his long involvement in the semiconductor industry. He used this background to insure that the most important topics of VLSI fabrication were addressed, and that the information was up-to-date and presented in a technically correct fashion. Note that for over twenty-five years, Dr. Tauber has held positions at Bell Telephone Laboratories, Xerox, Hughes Aircraft Company Microelectronics Research Center, the Microelectronics Center of TRW, and Applied Materials. The labor of the writing effort was divided between the anthors in the following manner: Professor Wolf was responsible for writing Chapters 1, 2, 3, 9, 10, 12, 13, 15, 16, 17, and 18, and Dr. Tauber undertook the writing of Chapters 4, 5, 7, 8, 11, and 14. Material for Chap. 6 was jointly contributed by Andrew R. Coulson and Dr. Tauber.

A book of this length and diversity would not have been possible without the indirect and direct assistance of many other workers. To begin with, virtually all of the information presented in this text is based on the research efforts of a countless number of scientists and engineers. Their contributions are recognized to a small degree by citing some of their articles in the references given at the end of each chapter. The direct help came in a variety of forms, and was generously provided by many people. The text is a much better work as a result of this aid, and the authors express heartfelt thanks to those who gave of their time, energy, and intellect.

Each of the chapters was reviewed after the writing was completed. The engineers and scientists who participated in this review were numerous. Special thanks are given to Warren Flack, Stephen Franz, Kenneth Tokunaga, and Simon Prussin for their critical reviews. Superlative computer support and access to computer resources was generously made availahle by Donald E. Carlile, Harry T. Hayes, and Dale Lambertson of the Personal Computer Support Section of the Electronic Systems Group of TRW. Henry Nicholas was a computer expert and friend who lit the fire of inspiration that led to the undertaking of the project. The management

of the Electronics Systems Group and the Microelectronics Center of TRW are warmly thanked for providing a supportive environment, conducive to producing such an intensive technical project. They made available technical literature and other resources to the authors, especially S. Wolf, who was able to avail himself of this generosity while writing during a Sabbatical leave from his teaching duties at California State University, Long Beach. Roy Montibon and Donald Strout of New Archetype Publishing, Inc., Los Angeles, CA designed the cover. Finally, we wish to thank Shirley Rome, Carrol Ann Wolf, and Barbara Tauber for typing the manuscript.

Stanley Wolf and Richard N. Tauber

P.S. Additional copies of the book can be obtained from:

LATTICE PRESS
P.O. Box 340
Sunset Beach, CA, 90742

An order form, for your convenience, is provided on the final leaf of the book.

PROLOGUE

Since the invention of the first integrated circuit in 1960, there has been an ever-increasing density of devices manufacturable on semiconductor substrates. Silicon technology has remained the dominant force in integrated circuit fabrication and is likely to retain this position in the foreseeable future. The number of devices manufactured on a single chip exceeded the generally accepted definition of *very large scale integration,* or *VLSI* (i.e., more than 100,000 devices per chip) in the mid-1970's (Fig. 1a). By 1986 this number had grown to over 1 million devices per chip, and by 2000 had exceeded *1 billion devices* per chip (e.g., the 1 GB DRAM). This increasing device count has been accompanied by a shrinking minimum feature size (Fig. 1b), which is expected to be smaller than 0.18 μm by 2000. Progress in ULSI manufacturing technology seems likely to continue to proceed in this manner. Further reductions in the unit cost per function, and in the power-delay product of ULSI devices, are projected. The entire saga of IC fabrication represents a remarkable application of scientific knowledge to the requirements of technology. This book represents an enthusiastic report on the state-of-the-art of ULSI silicon processing, as practiced at the time of its publication.

Figure 2 illustrates the sequence of steps that occurs in the course of manufacturing an integrated circuit. These steps can be grouped into two phases: 1) the *design phase*; and 2) the *fabrication phase* (Fig. 3). While this book is concerned only with the fabrication phase of this undertaking, it is also useful to briefly outline the steps of the design phase here. This provides the context which allows readers to perceive the role of silicon processing within the totality of integrated circuit manufacturing. Readers wishing to explore steps of the design phase in detail are referred to other technical literature, including the texts listed in References. 1, 2 and 3.

The desired functions and necessary operating specifications of the circuit are initially decided upon in the *design phase*. The chip is designed from the "top down." That is, the required large *functional blocks* are first identified. Next, their *sub-blocks* are defined, and finally

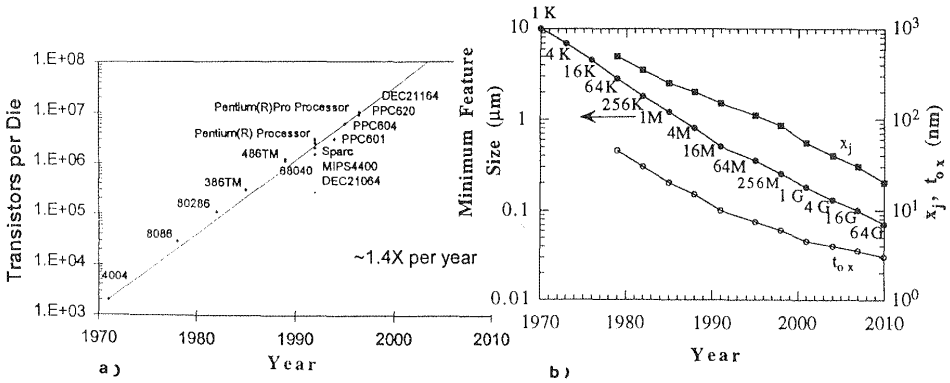


Fig. 1 (a) Increase in the number of transistors per microprocessor chip versus year of introduction, for a variety of microprocessor designs.[4] (© IEEE 1998) (b) Predictions of the decrease in minimum device feature size, junction depth, and gate oxide thickness versus time on integrated circuits, according to the SIA National Technology Roadmap for Semiconductors.

the *logic gates* needed to implement the sub-blocks are chosen. Each logic gate is designed by appropriately connecting devices that are ultimately slated for fabrication on the Si wafers. Upon completion of these various levels of design, each level is rechecked to insure that correct functionality has been achieved. When all aspects of the circuit design are correct to the designer's satisfaction, *test vectors* (that will be used to test the manufactured circuits), are generated from the schematic of the logic gates.

The circuit is then layed out. The *layout* consists of sets of patterns that will be transferred to the silicon wafer. These patterns correspond to device regions or interconnect structures, and such patterns are sequentially transferred to the wafers as part of the wafer fabrication sequence (That is, through the use of photolithographic processes and a set of masks, see Fig. 4a). The

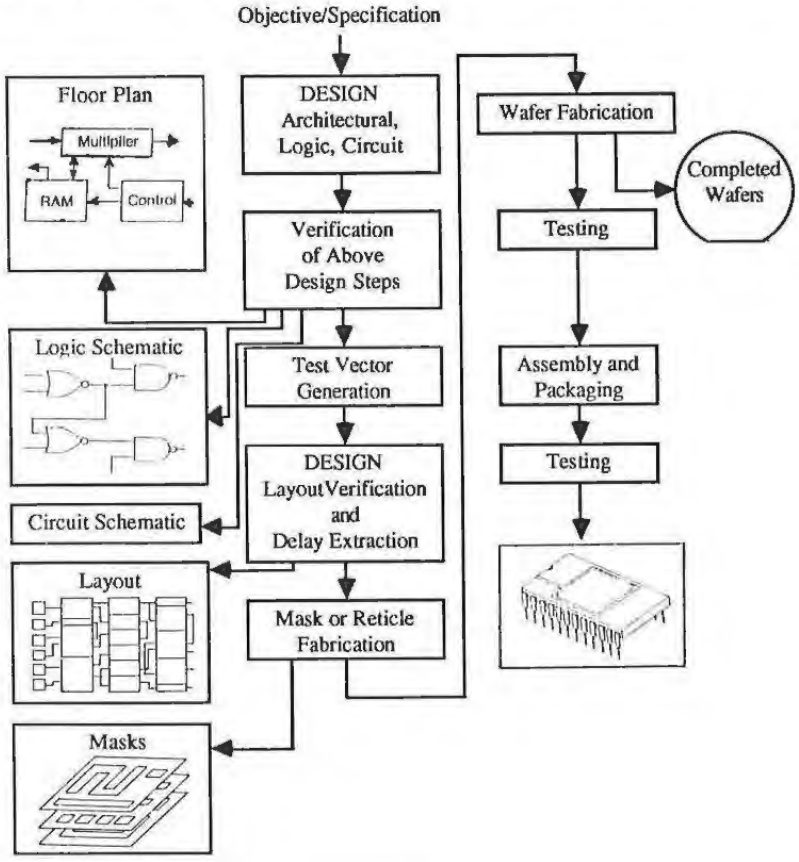


Fig. 2 Steps required for the manufacture of integrated circuits.

result of each pattern-transfer step is a set of features created on the wafer surface. These features are generally either in the form of: a) an etched opening in a film (or region of the substrate), or b) a patterned feature of a film present on the surface (e.g., an interconnect line or pad). After the openings (or *windows*) are created by the pattern transfer step, either controlled quantities of dopant are added to the silicon substrate through the openings, another layer is deposited that makes contact to the underlying layer through the opening. In either case, device regions and structures that interconnect device regions, are produced by the patterning processes and associated fabrication steps. A cross-section of a completed device is shown in Fig. 4b.

While the circuit is designed from the "top down," creation of the layout proceeds from the "bottom up." A variety of typical devices (e.g., transistors and resistors) are first layed out. Then, a set of cells representing the required primitive logic gates are created by interconnecting appropriate devices. Next, sub-blocks are generated by connecting these logic gates, and finally the functional blocks are layed out by connecting the sub-blocks. Additional items required by the circuit design are also incorporated during the layout process (e.g., power busses, clock-lines, and input-output pads). The completed layout is then subjected to a set of *design rule checks* and *propagation delay simulations* to verify that correct implementation of the circuit

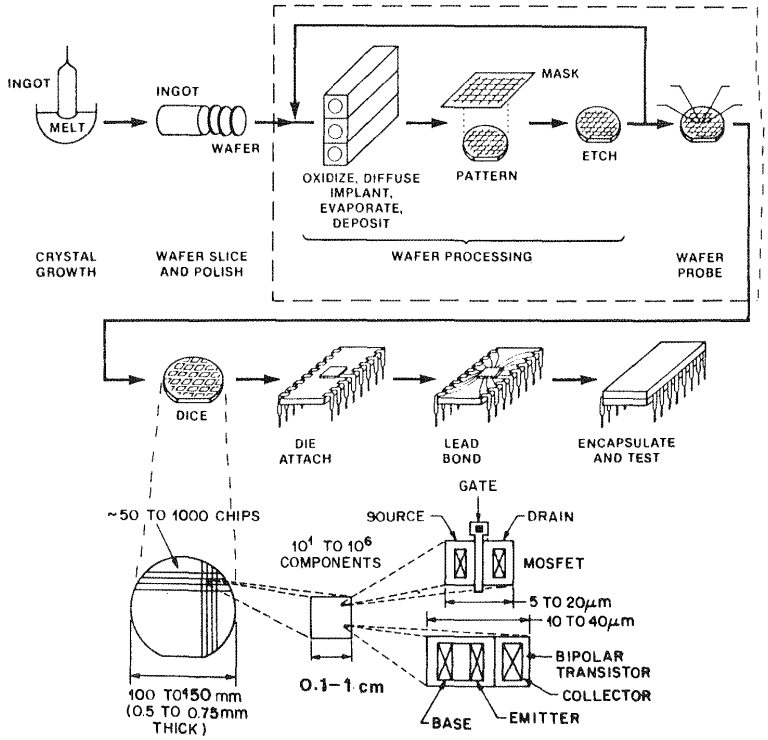


Fig. 3 The *wafer fabrication* process sequence of integrated circuits.

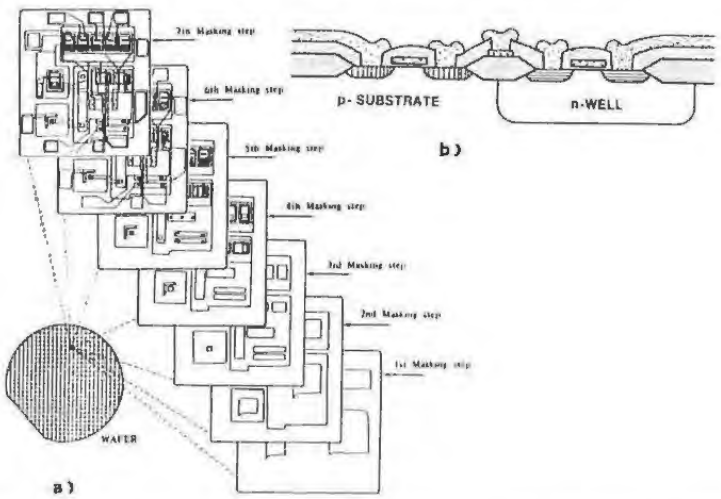


Fig. 4 (a) Example of the patterns transferred to a wafer during a seven-mask process sequence, and (b) Cross section of completed devices in a basic CMOS process.

has been achieved in layout form. Upon completion of this checking procedure, the layout information is ready to be used to generate a set of masks that will serve as tools for specifying the circuit patterns on silicon wafers. This layout information is stored on a computer.

In the three volumes of *Silicon Processing for the VLSI Era* integrated circuit manufacturing steps are described (Fig. 3). These steps start at the point where the layout information has been finalized. At that point procedures are utilized to convert the layout information stored on the computer, into a set of masks or reticles. This is described in Chap. 13. The individual fabrication process modules associated with creating patterns, introducing dopants, and depositing films on silicon substrates (to form integrated circuit features) are also subjects of this volume.

In Volume 2, information about how such individual process modules are combined to create complete ICs are covered in detail (namely the tasks of *process integration*). A brief discussion of process integration issues is also provided in Chaps. 15 and 16 of Volume 1. Volume 3 concentrates on the device physics of the submicron MOSFET and how device characteristics and the device processing steps are interrelated.

In Vol. 1, background information on science and technology common to many IC fabrication steps is also provided (e.g., vacuum technology [Chap. 3], material properties of thin films [Chap. 5], and optical science for microlithography [Chap. 13]). A short description of the so-called *back end* of IC processing (i.e., assembly and packaging) is given in Chap. 17. These three books can be read selectively to glean specific information. References are made to other chapters of the texts to clarify points for readers who wish to use the text in this manner.

REFERENCES

1. C. Mead & L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, Reading, MA 1980.
2. S. Muroga, *VLSI Systems Design*, John Wiley & Sons, New York, 1983.
3. A. Mukherjee, *Introduction to nMOS and CMOS VLSI Systems Design*, Prentice-Hall, 1986.
4. R.A. Glasser, *Tech. Dig. IEDM*, 1998, p. 317.

Chapter 7

SILICON EPITAXIAL GROWTH and SILICON ON INSULATOR

The *epitaxial* growth process provides an excellent means of growing a thin film of single crystal material upon the surface of a single crystal substrate. The term *epitaxy* is derived from two Greek words meaning "arranged upon." If the grown film is the same material as the substrate, the process is termed *homoepitaxy*. Homoepitaxy is generally referred to as *epitaxy* or simply *epi* and these terms will be used interchangeably throughout this text. Homoepitaxy of single crystalline silicon is the most important technological use of epitaxy, and is the primary subject of this chapter.

On the other hand, if the epitaxial film is grown on a substrate that is chemically different, the process is then termed *heteroepitaxy*. The principal applications of heteroepitaxy are in the deposition of silicon on sapphire (SOS) and in the fabrication of heterojunction semiconductor devices. When a heteroepitaxial deposition is made on a substrate that has a lattice constant very different than that of the film, a strained layer can form near the film/substrate interface. If the deposited film is thin enough, the strain can be accommodated elastically, but if the film is too thick the strain becomes relaxed through the generation of defects (e.g., misfit dislocations). These defects are known to propagate from the interface into the device region where they can adversely affect device properties. Conversely, if the lattice mismatch is small (as in the case of the growth of silicon-germanium alloys on silicon substrates), the strain is accommodated elastically. Under these conditions the epitaxial growth is termed *pseudomorphic* growth. Pseudomorphic growth is the method used in the formation of silicon-heterojunction-bipolar transistors (HBTs). Figures 7-1a, 7-1b, and 7-1c show a schematic representation of homoepitaxy, heteroepitaxy (strain relieved), and pseudomorphic epitaxial growth, respectively.

Epitaxial growth can be achieved by depositing the film directly from either the vapor phase,

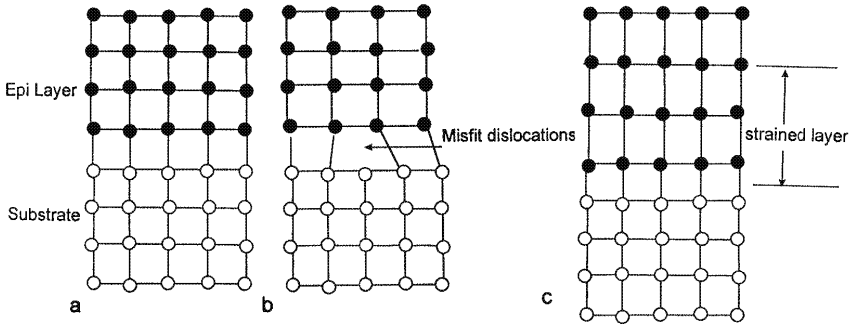


Fig. 7-1 Schematic of epitaxial structures: a) homoepitaxy; b) strain relieved heteroepitaxy (e.g., silicon-on-sapphire [SOS]); and c) psuedomorphic epitaxy.

the liquid phase, or the solid phase. These techniques are termed *vapor-phase epitaxy* (VPE), *liquid-phase epitaxy* (LPE), and *solid-phase epitaxy* (SPE), respectively. VPE is the predominant method used in silicon processing, since it provides a means of achieving excellent control of the film thickness, impurity concentration, and crystalline perfection. A disadvantage of VPE is that it is requires relatively high temperatures (800–1150°C) in order to realize the necessary crystalline perfection. Such high-temperature processing can lead to the occurrence of *out-diffusion* and *autodoping*. These phenomena place a limitation on the ability to control the doping level in the epi layer and are discussed in detail in Section 7.4.4 of this chapter.

LPE has found its widest use producing epitaxial layers of III-V compounds (e.g., GaAs and InP). LPE, however, occurs during the re-freezing of molten layers of silicon on a silicon surface during the laser melting GILD process (see Chap. 9). SPE occurs during the crystalline re-growth that takes place during the low-temperature annealing of amorphous layers formed as a result of damage during high-dose ion implantation (see Chap. 10).

This chapter covers: 1) device applications of epitaxy; 2) fundamentals of epitaxial film growth; 3) epitaxial process considerations, including silicon precursors and doping effects; 4) defects in epitaxial layers; 5) pattern effects; 6) low-temperature epitaxy; 7) selective epitaxy; 8) epitaxial film characterization; and 9) silicon on insulators (SOI). The fundamentals of the kinetics and gas dynamics of CVD (epitaxial growth is a CVD process) are covered in detail in Chapter 6. In this chapter only the relevant results are repeated.

7.1 THE DEVICE APPLICATIONS OF EPITAXY

In the early days of semiconductor fabrication, epitaxial deposition was used to improve the performance of bipolar transistors that were then commonly used. By growing a lightly doped n-type epitaxial layer over a heavily doped n^+ substrate the bipolar device could be optimized for high breakdown voltage of the collector-substrate junction, while preserving a low collector resistance. The low collector resistance was required to improve the device operating speeds at moderate currents. Figure 7-2 depicts a cross-section of a high-speed oxide-isolated bipolar transistor showing the presence of the epitaxial layer. A key feature of bipolar epitaxy is that the film is deposited over patterned, n^+-doped diffused layers called *buried layers* or *sub-collectors*. BiCMOS devices are also be fabricated in epitaxial films deposited over buried layers.

More recently, epitaxial processes have been used to fabricate CMOS integrated circuits. In CMOS circuits the complete device is built in a thin (2–4 μm), lightly-doped p-type (or in some cases, intrinsic) epitaxial layer that is deposited over a uniformly heavily p^+-doped substrate. Figure 7-3 shows a cross section of a twin-well CMOS device depicting the epi layer.

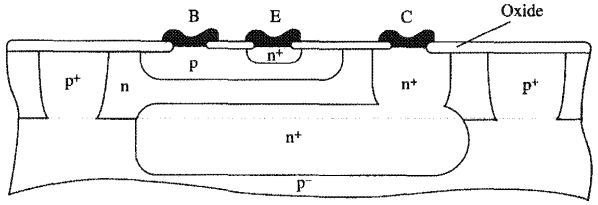


Fig. 7-2 Cross-section of a bipolar transistor built in an epitaxial layer. The light line indicates the boundary of the deposited n-type epi layer. Note that the n^+-buried layer extends into the epi layer as a result of autodoping and out-diffusion.

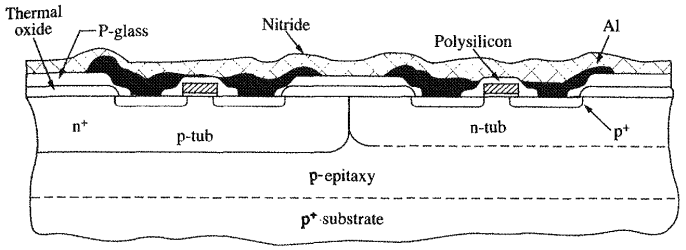


Fig. 7-3 Cross-section of a twin-well CMOS structure built in a p^- epitaxial layer and a p^+ substrate. Note that the p^+-region extends into the epi layer as a result of autodoping and out-diffusion.

In the next section we discuss the reasons for using epitaxial silicon rather than *polished bulk wafers* (*PBW*) for constructing advanced CMOS devices. It is interesting to note that IC manufacturers usually purchase CMOS epitaxial wafers directly from the wafer suppliers, while for bipolar ICs they process the epitaxy themselves. The reason is that a masking layer is required to produce the buried layer prior to the epi in the bipolar process.

7.1.1. Why Are CMOS Epitaxial Wafers Used?

CMOS epitaxial wafers, when compared to polished bulk wafers provide: a) improved latch-up immunity; b) layers of silicon free of SiO_x precipitates, COPs, or vacancy condensates (all of which are formed during crystal growth); and c) a silicon surface that is smoother than that found on PWBs, with minimal surface damage.

Latch-up in CMOS circuits is a phenomenon that causes the formation of a low resistance path between the supply voltage and ground. As a result, it is possible for the circuit to draw an unusually high amount of current. Such current draw can cause the circuit to cease functioning or even to "burn-up." The device physics underlying latch-up is beyond the scope of this text, but is covered in detail in Vol. 2 of this series. Although there are process and design techniques that can minimize latch-up, the use of epitaxial silicon has become the "standard" approach to fabricating ULSI CMOS microprocessor circuits because it provides latch-up protection.

Polished bulk wafers are known to include SiO_x precipitates, COPs, and vacancy condensate defects at or near the wafer surface. Moreover, despite the best polishing techniques, the wafer surface can retain polishing defects such as micro-damage and surface roughness. Surface defects are known to diminish the gate-oxide integrity and increase junction leakage. Epitaxial layers, on the other hand, do not contain oxygen (hence no SiO_x precipitates), COPs, or vacancy condensates, since these defects only form during crystal growth (during which oxygen is dissolved in the as-grown Si crystal, see Chaps 1 and 2). Since epi is deposited on a polished wafer, the new surface that forms is generally free of polishing micro-damage and surface roughness. As a result, CMOS devices produced in epitaxial material have superior dielectric integrity and junction leakage currents, compared to similar devices made on polished wafers.

The use of epitaxial wafers in the fabrication of microprocessors is now standard, while its use in DRAM devices is rather limited. There are two reasons for this. First, the gate oxide thickness required for DRAM devices is generally thicker than that required for microprocessor devices, even for the same "generation" of technology. These thicker gate oxides are less susceptible to the presence of surface defects and do not benefit as much from the advantages of the epi layer. Second, the cost to purchase a 200-mm epitaxial wafer is about 1.5–1.8 times that of a polished wafer. Since low manufacturing costs for DRAMs are vital for profitability, the

Table 7-1 APPLICATIONS OF EPITAXY

Application	Thickness (μm)	Resistivity Range (ohm-cm)
Bipolar devices	0.8–1.6	0.5–4.0
CMOS devices	2–4	2–10
BiCMOS devices	3–5	2–20
High voltage devices	5–50	50–80

added cost per wafer is often too high to justify its use. However, the use of epi wafers for DRAM manufacturing is under serious consideration as thinner gate oxides become necessary. When the yields on devices fabricated in bulk wafers no longer meets the requirements for DRAM manufacturing, the industry may be compelled to change to epitaxial wafers. On the other hand, the DRAM manufacturers will try to find alternative processes to extend the life of bulk wafers, rather than pay the added cost associated with the use of epitaxial wafers.

Table 7-1 shows some of the applications of silicon epitaxy along with the epi thickness and resistivity ranges that are commonly used. Table 7-2 (which is extracted from the SIA roadmap) shows the epi layer requirements for the most advanced CMOS devices.[1] The two key points illustrated in that table are: 1) thinner epi layers are required (i.e., down to 1 μm); and 2) the structural defect density must be reduced by more than a factor of 2 from its current level. These two discordant requirements represent a significant challenge, since thinner films require lower growth temperatures, and lower temperatures can lead to increased defect density.

7.2 GROWTH OF EPITAXIAL LAYERS

In this section an atomic mechanism that has been proposed to explain epi film growth is first discussed. Then the various chemicals precursors used in conventional epitaxial growth are described. The kinetics of CVD were covered in detail in Chap. 6, and only the results pertinent to epitaxial growth will be repeated here.

7.2.1 Atomistic Model of Film Growth

Whatever *precursor* (source gas) is used and whatever chain of reactions occur (see next section) the final reactant species must be adsorbed on the growing surface where the reaction takes place. The reaction produces silicon and some by-product. In order for film growth to occur, the silicon atoms must remain adsorbed on the surface and the by-products desorbed. The adsorbed silicon atoms are referred to as *adatoms*. If an adatom is in a position such as A in Fig. 7-4, one of several events can take place. If the adatom A remains in its position other silicon atoms can become attached to it and a small silicon cluster or island can form. If numerous clusters form and begin to grow, it is likely that as they merge, a highly defected or even polycrystalline film will form. However, as it turns out, adatoms in position A are quite mobile at the deposition

Table 7-2 EPITAXIAL LAYER REQUIREMENTS vs. TECHNOLOGY NODE
Extracted from NTRS, SIA (1997)[1]

First year product shipment	1997	1999	2001	2003
Feature size (nm)	250	180	150	130
Layer thickness (μm)	2–5	2–5	2–4	2–4
Thickness tolerance (±%)	5	4	4	4
Layer structural defect density (m^2)	≤3.3	≤2.9	≤2.6	≤2.3

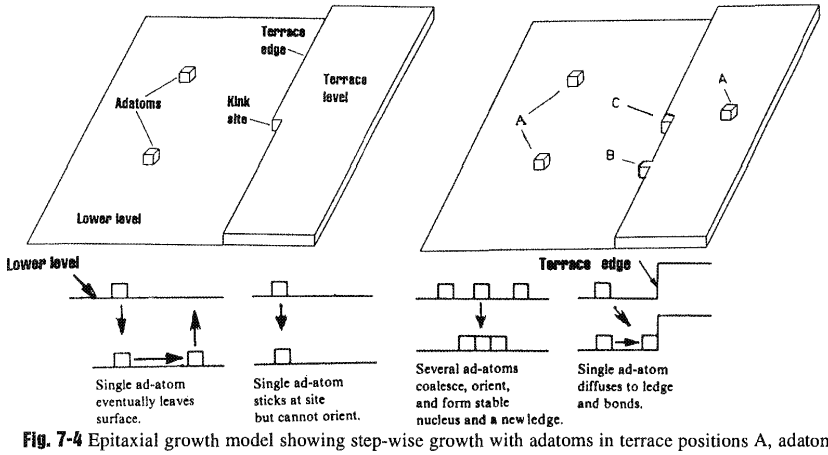


Fig. 7-4 Epitaxial growth model showing step-wise growth with adatoms in terrace positions A, adatoms at steps (or terrace edges) B, and adatoms at stable kink positions C.

temperatures and are likely to migrate along the surface. If a migrating adatom reaches a position along a growing *step* (or *terrace edge*), such as position B in Fig. 7-4, the adatom is more likely to remain in place (since atoms at position B are more energetically stable than at positions A). The stability arises from the fact that more Si-Si bonds have already formed when the adatom is at position B. The most energetically favored site for the adatom, however, is the so-called *kink position* (position-C in Fig. 7-4). When the adatom arrives at such a kink position, one-half of the Si-Si bonds can be established and further migration is thereafter unlikely to occur. For growth to continue, additional adatoms must migrate to kink positions and become incorporated in the growing film. In this respect, film growth proceeds by lateral movement (2-dimensional) of the steps on the lattice surface. After one layer is completed a second layer grows. Since epi proceeds by this lateral growth, an impurity sitting on the wafer surface can impede the motion and produce a stacking fault or dislocation defect in the film. For this reason, extremely clean silicon surfaces are required to achieve high-quality epi growth.

If adatom migration is suppressed during growth, a polycrystalline film is more likely to be deposited. Atom mobility is restricted by high deposition rates and low temperatures. At any particular deposition temperature there exists a maximum deposition rate above which polycrystalline films are deposited, but below which single crystal epitaxial films are formed.[2] Figure 7-5, shows a plot of the effect of growth rate and temperature on the formation of single (monocrystalline) or polycrystalline films. From that curve we observe that monocrystalline growth is favored at high temperatures and low growth rates, and polycrystalline deposition is favored under conditions of high growth rates and low temperatures. The maximum deposition rate at which single-crystal growth can be maintained is found to increase exponentially with temperature. When these results are plotted on an Arrhenius plot (Fig. 7-5) a straight-line dependence is obtained (indicative of a thermally activated process). The value of the activation energy extracted from this curve is ~5 eV. These results may be explained as follows. At high growth rates there is not sufficient time for the adatoms to migrate to the kink sites, thereby polycrystalline growth is favored. As the temperatures increase, however, the surface migration rate also increases, allowing time for silicon to reach the kink sites before additional adatoms form other clusters. Once at the kink site, the film continues to grow by lateral motion. The

Chapter 9
DIFFUSION in SILICON

Diffusion is the phenomenon by which one chemical constituent moves within another as a result of the presence of a chemical gradient (more accurately, a chemical potential gradient). The diffusion of controlled impurities or dopants in silicon is the foundation of *pn* junction formation and device fabrication. In the early days of transistor and IC processing, dopants were supplied to the silicon by *chemical sources*. These dopants were then thermally diffused to the desired depth by subjecting the wafers to an elevated temperature treatment (900–1200°C) in a furnace. However, with the onset of submicron device fabrication, *ion implantation* became the standard method for introducing dopants into Si, and chemical-source diffusion was essentially abandoned. However, interest in this method was revived to some degree, primarily for forming ultra-shallow junctions (<100 nm). The chemical techniques being investigated for this application include rapid vapor doping (RVD) and P-GILD. They are described later in this chapter.

Nevertheless, after the dopants are introduced into silicon wafers by ion implantation they still need to be electrically activated. Such electrical activation requires a high-temperature operation called an *anneal*. Dopant diffusion also occurs during these activation anneals, as well as during any other high-temperature processes that may take place after the anneal (e.g., thermal oxidation, BPSG reflow). This implies that a high-temperature treatment will affect the distribution of all dopants already in the silicon. Therefore, it must be possible to trace the doping profile of all of the previous implants up to the last high-temperature thermal process. The control of the doping profile becomes ever more difficult as the devices shrink in size, and for that reason a fundamental understanding of diffusion remains important for ULSI fabrication.

Unfortunately, diffusion in silicon is not accounted for simply by solving the mathematical equations of diffusion. In this regard, the diffusion of dopants in silicon differs from that of diffusion in metals. Since silicon is a semiconductor, and the diffusing dopants are electrically charged, one must account for the interaction of the dopants with silicon point defects. These interactions are controlled by equilibrium reactions relating the balance of charge from all sources in the silicon crystal (i.e., intrinsic electron-hole pairs, *n*-type and *p*-type dopants, and the various charged defects). Silicon defects are discussed in Chap. 2.

In the early days of silicon technology, the junction depths in devices were typically as deep as 2000–4000 nm and their fabrication was relatively uncomplicated. With current generations of technology, the required junction depth has been scaled to 50–100 nm and future requirements will dictate junctions as shallow as 20 nm. This reduction in junction depth has made the fabrication of diffused junctions a major challenge. Notable advances in ion implantation (see Chap. 10) and thermal processing technology (RTP) have gone a long way to enabling the formation of such shallow junctions. Accompanying the fabrication of ultra-shallow junctions is the challenge of developing reliable analytical techniques for determining junction depths and concentration profiles. These techniques are also discussed in this chapter. Table 9-1 shows the National Technology Roadmap for Semiconductors[1] requirements for the range of minimum junction depth necessary for each successive technology generation from 1997 to 2012.

In this chapter the following topics are covered: a) mathematics governing the mass transport phenomena of diffusion (Fick's equations and their solutions); b) the temperature dependence

Table 9-1 MINIMUM JUNCTION DEPTH VERSUS TECHNOLOGY NODE
(Extracted from NTRS, SIA [1997]).

First year product was shipped	1997	1999	2001	2003	2006	2009	2012
Linewidth in nm	250	180	150	130	100	70	50
Minimum junction depth, nm	50–100	36–72	30–60	26–52	20–40	15–30	10–20

of diffusion coefficients; c) the diffusion coefficients of the substitutional impurities; d) the simulation of one-dimensional doping profiles with SUPREM; e) atomistic models of diffusion in Si; f) diffusion in polycrystalline-Si; g) diffusion in SiO_2; h) anomalous diffusion effects; i) transient enhanced diffusion (TED); and j) measurement techniques used in diffusion studies.

9.1 THE MATHEMATICS OF DIFFUSION

In this section the basic differential equations of simple diffusion and their solutions for some special cases of diffusion in Si are described. The meaning of the *diffusion coefficient* D is defined and a method to experimentally determine its value is discussed. (The terms *diffusion coefficient, diffusivity,* and *diffusion constant* are used interchangably.) The atomic nature of the Si matrix and the interaction of the dopants with Si defects is reviewed in a subsequent section.

9.1.1 Fick's First Law

The basic mathematical tools for treating diffusion were elucidated in 1855 when Fick proposed that equations analogous to Fourier's heat-flow equations, should also apply to the diffusive flow of matter.[2] The first law posited by Fick states that if a concentration gradient, $\partial C/\partial x$, of an impurity exists in a finite volume of a matrix substance, then there exists a tendency for the impurity material to move in such a direction as to decrease the gradient. When diffusion occurs, the impurity concentration at any location becomes a function of both distance and time. One can represent the concentration profile by $C(x,t)$. If the flow persists for a sufficiently long time, the material will become homogeneous and the net flow of matter will cease. Fick went on to state that the flow of the impurity, represented by the *flux* J is proportional to the concentration gradient, $\partial C/\partial x$ and that the constant of proportionality is defined as the *diffusivity* D of the impurity in that particular matrix. Fick's statements can be represented mathematically as:

$$J = -D\frac{\partial C(x,t)}{\partial x} \tag{9.1}$$

Equation 9-1 implies that as the concentration gradient decreases, the flux or diffusion decreases. The flux, J, is defined as the mass of material moving per unit area, per unit time. The units used in Eq. 9-1 are shown in Eq. 9-1a:

$$J\ (\mathrm{gm/cm^2 sec}) = -D(\mathrm{cm^2/sec})\frac{\partial C(\mathrm{gm/cm^3})}{\partial x(\mathrm{cm})} \tag{9.1a}$$

Equation 9-1a shows the units for D as cm^2/sec, but D can also be given in other units (e.g., μm^2/hr, or m^2/sec). The constant D depends on the diffusion temperature, the diffusing species (also called the *diffusant*), and the concentration of the diffusant. Other factors such as the ambient conditions during diffusion and the presence of other dopants also affect the measured value of D, and they will be discussed in ensuing sections of this chapter.

9.1.2 Fick's Second Law

Fick's Second Law in one-dimension is now derived from the first law. In order to obtain useful results one needs to determine what happens to the concentration gradient with the progression

of time. To do this refer to Fig. 9-1, where the amount of dopant entering and leaving a finite volume of matrix is illustrated. Consider a finite volume of matrix material ΔV, having a thickness Δx and a unit cross-sectional area (i.e., $\Delta V = 1 \cdot 1 \cdot \Delta x$). Material enters or leaves this volume only in one dimension (i.e., only in the $\pm x$ direction). We define J_1 as the flux of material *entering* the volume Δx, and J_2 as the flux *leaving* the same volume element. If the volume element is very thin (i.e., $\Delta x \rightarrow 0$), then the difference in the fluxes is:

$$J_2 - J_1 = -\Delta x\left(\frac{\partial J}{\partial x}\right) \tag{9.2}$$

Since the mass of material that enters the element in unit time (J_1) is different than the amount which leaves (J_2), the concentration within the element must also be changing with time. The change of the concentration with respect to time is given by:

$$\frac{\partial C}{\partial t} = \frac{J_2 - J_1}{\Delta x} \tag{9.3}$$

By rearranging Eq. 9-2 to get $(J_2 - J_1)/\Delta x$ on one side, and substituting the other side into Eq. 9-3 one gets:

$$\frac{\partial C(x,t)}{\partial t} = -\frac{\partial J}{\partial x} \tag{9.4}$$

Now substitute J from Eq. 9-1 into Eq. 9-4, to obtain *Fick's Second Law*, which can be written:

$$\frac{\partial C(x,t)}{\partial t} = \frac{\partial}{\partial x}\left(D\frac{\partial C}{\partial x}\right) \tag{9.5}$$

Equation 9-5 is the most general representation of Fick's second law. If the diffusion coefficient D is independent of position (concentration) then D can be brought outside the partial differential and Eq. 9-5 can be rewritten in a simpler form as:

$$\frac{\partial C(x,t)}{\partial t} = D\left[\frac{\partial^2 C(x,t)}{\partial x^2}\right] \tag{9.6}$$

The fact that D in Eq. 9-6 is assumed to be independent of position implies it is also independent of concentration (which varies with position as a result of the concentration gradient). As it turns out, for low concentrations of diffusant the assumption that D is independent of position is valid, and Eq. 9-6 can be used. But for high diffusant concentrations one must use Eq. 9-5.

9.1.3 Solutions to Fick's Second Law

Two solutions to Fick's Second Law when D is independent of concentration (i.e., Eq. 9-6 prevails), will be examined. The first solution relates to the case where dopants are introduced into the silicon from a vapor source. The source is considered to bring a constant supply of dopant atoms to the silicon surface where they diffuse into the silicon at the process temperature.

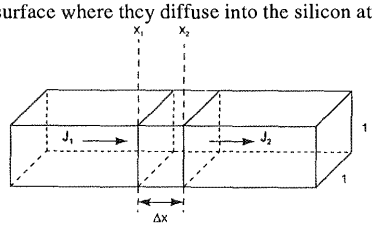


Fig. 9-1 Schematic showing an element of volume with the flux J_1 entering and J_2 leaving.

It is assumed that the source maintains a constant value of surface concentration during this entire diffusion process. In diffusion technology, this type of dopant incorporation is termed *predeposition*, and is used to introduce a known quantity of impurities into the silicon. As noted earlier, this step in the past was carried out by diffusion from a chemical source, but now has been largely replaced by ion implantation. The pre-deposited dopant (whether chemically or ion implanted) is then redistributed in the silicon by a further thermal treatment called a *drive-in* or *redistribution diffusion*. Despite the reliance on ion implantation for the pre-deposition step, it is still illustrative to solve Fick's differential equation for the pre-deposition boundary conditions. Moreover, chemical source pre-deposition is still being used for some deep junction applications and is now under consideration again for some ultra-shallow junction applications.

9.1.3.1 Chemical Pre-Deposition: In order to solve the second-order differential equation (Eq. 9-6) it is necessary to establish initial and boundary conditions. For the chemical predeposition step the initial condition is that the concentration of dopant in the silicon at t = 0 for all values of x is equal to 0. This condition is written as:

$$C(x, 0) = 0 \tag{9.7}$$

The two boundary conditions for this process step are established next. The first boundary condition (Eq. 9-8) states that the concentration of dopant at x = 0 (i.e., the concentration at the surface of the Si), for any time during the diffusion, is fixed at a value C_s for the entire diffusion cycle. The units of C_S are atoms/cm^3 and the condition is described mathematically as:

$$C(0, t) = C_S \tag{9.8}$$

The second boundary condition states that the concentration of dopant at $x = \infty$ is equal to 0 for any time, or:

$$C(\infty, t) = 0 \tag{9.9}$$

The solution to the partial differential equation (Eq. 9-6) with the application of the initial and boundary conditions is given by Eq. 9-10:

$$C(x,t) = C_S \text{erfc}\left(\frac{x}{2\sqrt{Dt}}\right) \tag{9.10}$$

where D is the diffusion coefficient, x is the distance coordinate, t is the time of the diffusion, and *erfc* is the *complementary error function*. The complementary error function is tabulated[3] for values of the dummy variable $z = x/2(Dt)^{1/2}$. The denominator of Eq. 9.10, $2(Dt)^{1/2}$, is termed the *diffusion length*, which is representative of how far atoms move during a given thermal step. The term is also referred to as *"root Dt"*. The units of "root Dt" are length (cm) and its value is used when comparing the thermal exposure or thermal budget of a process (see Chap. 8).

Figure 9-2 shows a plot of the concentration profile for a predeposition for several different times at a constant temperature. D is assumed to be constant and the profile only depends on the time of the diffusion. As the diffusion time is increased the surface concentration remains constant at C_s, while the dopants move deeper into the silicon. The total amount of dopant introduced into the silicon also increases with time. The *junction depth*, x_j, can be defined as the point at which the concentration of the diffusant equals the background doping in the Si. The total amount of dopant that accumulates in the silicon (Q_O) during the predeposition can now be calculated by integrating Eq. 9-10 with respect to x over all of space (in one-dimension) and obtain Eq. 9-11.

$$Q_O = \int_{-\infty}^{+\infty} C(x,t) = \int_{-\infty}^{+\infty} C_s \text{erfc}\left(\frac{x}{2\sqrt{Dt}}\right) dx = \frac{2}{\sqrt{\pi}} C_s \sqrt{Dt} \tag{9.11}$$

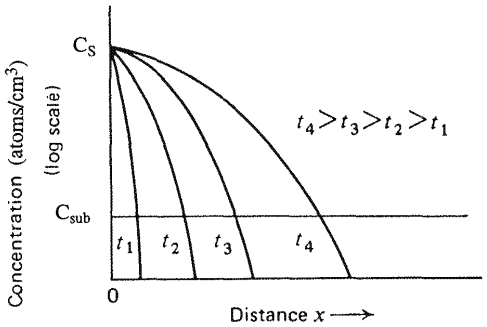


Fig. 9-2 Schematic representation of the diffusion profile for constant source diffusion (pre-deposition). Note the surface concentration remains constant with increasing diffusion time.[72] From S.K. Ghandi, *VLSI Fabrication Principles*. Reprinted with permission of John Wiley & Sons.

The junction depth, x_j, can be calculated by determining the depth (value of x) at which the diffusing dopant concentration just equals the background concentration in the silicon, C_{sub}. This is accomplished by solving Eq. 9-10 for x when $C = C_{sub}$:

$$x_j = (2\sqrt{Dt})\ \text{erfc}^{-1}\left(\frac{C_{sub}}{C_s}\right) \tag{9.12}$$

where erfc^{-1} is the inverse of the complementary error function. Both Eqs. 9-11 and 9-12 contain the "root Dt" term.

9.1.3.2 Drive-In Diffusion: Next consider what happens to the dopant atoms that have been introduced during the predeposition if the silicon is subjected to additional thermal treatment. We would expect that the total dose, Q_o, in the silicon should remain constant and the diffusant would move deeper into the silicon. This process is termed the *redistribution* or *drive-in* diffusion. The drive-in is used to move the impurities to the desired junction depth following the predeposition. In order to obtain a closed-form solution of Fick's Second Law, it is necessary to assume that the initial dopant is confined to a thin layer at the surface. As a matter of fact, the solution assumes that the dopant Q_o, exists as a δ-function in a thin rectangular region at the surface. The total quantity of impurity present is fixed at Q_o (atoms/cm^3). For the drive-in case, the initial condition is given by Eq. 9-13:

$$C(x, 0) = 0 \quad \text{for } x > \delta \tag{9.13}$$

and the two boundary conditions are given by Eqs. 9-14 and 9-15:

$$\int_0^{+\infty} C(x,t)dx = Q_o \tag{9.14}$$

and

$$C(\infty,t) = 0 \tag{9.15}$$

The boundary condition in Eq. 9-14, states that the total amount of dopant is fixed at Q_o. The solution to Fick's Second Law, under these conditions is:

$$C(x,t) = \frac{Q_o}{\sqrt{\pi Dt}} \exp\left(\frac{-x^2}{4\sqrt{Dt}}\right) \tag{9.16}$$

This solution has the form of a *Gaussian distribution*. The surface concentration, C_S (which in this case decreases as a function of time) is determined from Eq. 9-17 when $x = 0$:

$$C_s = C(0,t) = \frac{Q_o}{\sqrt{\pi Dt}} \quad \textbf{(9.17)}$$

From Eq. 9-17 one can see that the surface concentration decreases as the diffusion time increases. The junction depth for the drive-in diffusion case is calculated in a manner similar to that used to obtain Eq. 9-12 and is given by:

$$x_j = 2\sqrt{Dt}\left(\ln\frac{Q_o}{C_{sub}\sqrt{\pi Dt}}\right)^{1/2} \quad \textbf{(9.18)}$$

Figure 9-3 shows a plot of the concentration profile $C(x)$ for the Gaussian case for several drive-in diffusion times at a constant temperature. The key points to note are that the value of C_S decreases and the junction depth moves deeper into the silicon as the diffusion time increases.

An example of how to use these two diffusion models is now provided. In the first example the diffusion profile, the junction depth, and total dopant concentration after a pre-deposition step are each calculated. In the second example the information from the previous example is used to establish the diffusion profile and junction depth after a drive-in diffusion step.

EXAMPLE 9-1: Calculate: a) diffusion profile; b) junction depth; & c) total amount of dopant introduced after boron predeposition performed at 950°C for 30 min, in a neutral ambient. Assume the substrate is *n*-type with a background doping level of 1.5×10^{16} atoms/cm^3 and the B surface concentration reaches solid solubility ($C_s = 1.8 \times 10^{20}$ atoms/cm^3).

SOLUTION: a) Under the conditions of the predeposition, the diffusion profile is controlled by the erfc of Eq. 9-10. The profile can be found by using the erfc curve of Fig. 9-4. In order to use this curve, values of $x/(2\sqrt{Dt})$ must be calculated for values of x. Corresponding values of $C(x)/C(0)$ [where $C(0)$ is the B surface concentration C_s], are found from Fig. 9-4, and finally values of $C(x)$ are calculated. First, the value of the B diffusion coefficient at 950°C (1223 K) is found from (i.e., calculated using Eq. 9-38):

$$D_B(1223\text{ K}) = 0.76 \exp(-3.46\text{ eV}/k\ 1223\text{ K})\ [\text{cm}^2/\text{sec}] = 4.19 \times 10^{-15}\ \text{cm}^2/\text{sec}.$$

Next calculate the diffusion length, $z = 2\sqrt{Dt}$, for this process:

$$z = 2\sqrt{Dt} = 2\,(4.19 \times 10^{-15}\ \text{cm}^2/\text{sec} \cdot 1.8 \times 10^3\ \text{sec})^{1/2} = 5.49 \times 10^{-6}\ \text{cm}.$$

Now set up the following table to assist in the calculation of the diffusion profile:

x (cm x 10^{-4})	$x/(2\sqrt{Dt})$	$C(x)/C(0)$	$C(x)$ (atoms/cm^3)
0	0	1	1.8×10^{20}
0.05	0.91	0.198	3.5×10^{19}
0.075	1.36	0.054	9.7×10^{18}
0.10	1.82	0.010	1.8×10^{18}
0.15	2.73	0.00011	2.0×10^{16}

b) The junction depth, x_j, is defined as the value of x where the concentration of the diffusion profile equals the substrate doping (1.8×10^{16} atoms/cm^3). When $C(x) = 1.8 \times 10^{16}$ atoms/cm^3, then $C(x)/C(0) = 1.0 \times 10^{-4}$. Next obtain $x_j/(2\sqrt{Dt})$ from Fig. 9-4 (i.e., $x/\sqrt{Dt} = 2.76$)

Thus: $x_j/2\sqrt{Dt} = 2.76$; and $x_j = 2.76 \cdot 5.49 \times 10^{-6}$ cm $= 1.51 \times 10^{-5}$ cm $= 0.151\ \mu$m

c) The total amount of dopant is calculated from Eq. 9-11:

$$Q_o = C_s(2\sqrt{Dt})/\sqrt{\pi} = (1.8 \times 10^{20} \cdot 5.49 \times 10^{-6})/1.77 = 5.58 \times 10^{14}\ \text{atoms/cm}^2.$$

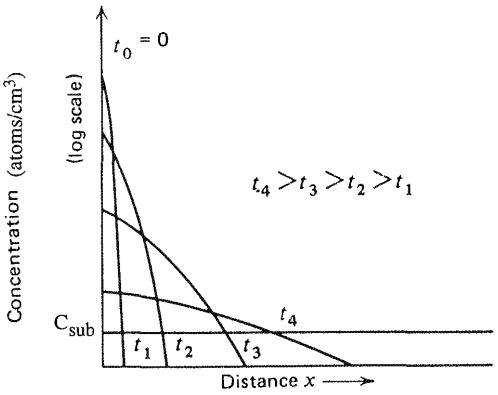


Fig. 9-3 Schematic representation of the diffusion profile for a limited source diffusion (drive-in). Note the surface concentration decreases with increasing diffusion time and the total dopant concentration remains constant.[72] From *VLSI Fabrication Principles*. Reprinted with permission of John Wiley & Sons.

EXAMPLE 9-2: Calculate the diffusion profile and junction depth after the predeposition of Example 9-1 is subjected to a neutral-ambient drive-in at 1050°C (1323 K) for 60 min.

SOLUTION: For the drive-in diffusion conditions of this example the Gaussian distribution of Eq. 9-16 (the curve labeled *Gaussian* in Fig. 9-4) controls the profile. First calculate the diffusion coefficient, D_B, and the diffusion length ($z = 2\sqrt{Dt}$) for the new diffusion conditions:

D_B (1323 K) = 0.76 exp(-3.46 eV/k 1323 K) = 5.0×10^{-14} cm^2/sec; and $z = 2\sqrt{Dt} = 2.68\times10^{-5}$ cm

Next calculate the value of the surface concentration C_s from Eq. 9-17:

$$C_s = Q_o/\sqrt{\pi Dt} = 5.58\times10^{14} \text{ atoms/cm}^2/(\pi \cdot 5.0\times10^{-14} \text{ cm}^2/\text{sec} \cdot 3600 \text{ sec})^{1/2}$$
$$= 2.34\times10^{19} \text{ atoms/cm}^3$$

Set up the following table to assist in the calculation, using Eq. 9-16, and $4Dt = 7.2\times10^{-10}$:

x (cm x 10^{-4})	$x^2/4Dt$	exp[$-x^2/4Dt$]	$C(x)$ (atoms/cm^3)
0	0	1	2.34×10^{19}
0.1	0.139	0.87	2.03×10^{19}
0.2	0.55	0.57	1.33×10^{19}
0.3	1.25	0.286	6.69×10^{18}
0.4	2.22	0.109	2.55×10^{18}
0.5	3.47	0.031	7.25×10^{17}
0.6	5.00	0.0067	1.69×10^{18}
0.7	6.81	0.0011	2.57×10^{16}
0.8	8.88	0.00014	3.27×10^{15}

Now calculate x_j directly from Eq. 9-18. First, we estimate x_j from the table above. Since C_{sub} is 1.8×10^{16} atoms/cm^3, x_j must lie between 0.7 μm and 0.8 μm.

$x_j = 2\sqrt{Dt}\,[\ln (C_s/C_{sub})]^{1/2} = 2.68\times10^{-5}\,[\ln(2.34\times10^{19}/1.8\times10^{16})]^{1/2} = (2.68\times10^{-5} \times 2.67)$ cm

or:

$x_j = 0.715\ \mu$m.

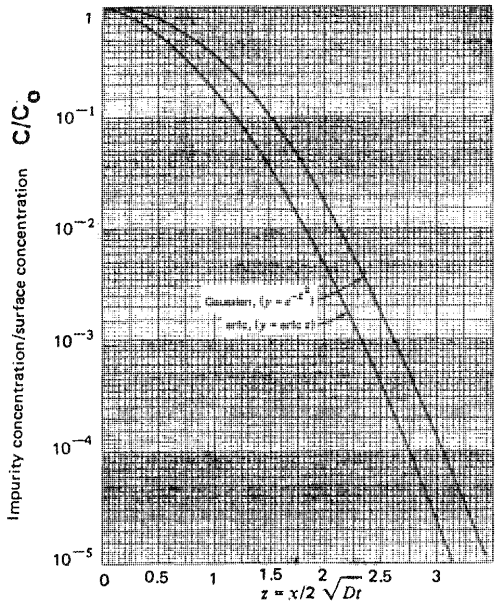


Fig. 9-4 Normalized concentrations [C/C(0)] as a function of normalized distance ($x/2\sqrt{Dt}$) for the erfc and Gaussian curves.

9.1.3.3 Drive-In From An Ion Implantation Predeposition: If the predeposition is from an ion implanted source (which produces a doping concentration with a near-Gaussian distribution close to the surface), the Gaussian profile will move when subjected to a high temperature anneal (see Chap. 10 on Ion Implantation). The solution to Eq. 9-6 (concentration-independent diffusion) with an initially implanted Gaussian profile has been reported for the cases of drive-ins performed in a neutral ambient,[4] and an oxidizing ambient.[5] For drive-ins in an oxidizing ambient, the solution to Eq. 9-6 is difficult to obtain in closed form, since it involves a moving boundary problem. Consequently, numerical methods must be employed to obtain the solutions (see Chap. 2, Vol. 3 of this series).

9.1.4 Concentration Dependence of the Diffusion Coefficient

The diffusion profiles that were calculated in the previous section are valid for the case when the diffusion constant is not a function of the doping concentration. That is why solutions of Eq. 9-6 were used, rather than those for Eq. 9-5. Constant-value D prevails when the doping concentration is lower than the intrinsic-carrier concentration n_i at the diffusion temperature, and conversely, concentration dependent diffusion occurs when the doping level is greater than the intrinsic carrier concentration. The first case is termed *intrinsic diffusion* while the latter is called *extrinsic diffusion*. To determine where extrinsic diffusion starts one needs to know the intrinsic carrier concentration at the diffusion temperature. Equation 9-19 is the expression used to calculate the intrinsic carrier concentration in silicon:

$$n_i = p_i = \sqrt{N_V N_C}\,\exp\left(-\frac{E_g(T)}{2kT}\right) \quad (cm^{-3}) \tag{9.19}$$

Here, N_V and N_C are the density of states at the valence and the conduction band edges, respect-

ively, E_g is the energy gap of Si, k is Boltzmann's constant, and T is the temperature in kelvins (K). The value of E_g depends somewhat upon temperature and the doping concentration and these factors must be taken into account when determining the intrinsic carrier concentration. The bandgap dependence on temperature in the range of 800–1100°C, and is given by:

$$E_g(T) = E_g(0) - bT \tag{9.20}$$

where $E_g(0) = 1.46$ eV and $b = 2.97 \times 10^{-4}$ eV/K.

There is a "bandgap narrowing" effect, δE_g, that occurs for heavily doped silicon. A "lumped" model has been developed to provide a value for the narrowing effect.[6] This model is not based upon first principles, since no simple physical model has been found to explain the effect. One proposal is that the lattice strain caused by the presence of heavy doping induces the band gap to change. When bandgap narrowing, $\delta E_g/2kT$ is included, the *effective intrinsic carrier concentration* n_i^{eff} is given by:

$$n_i^{eff} = n_i (\delta E_g/2kT) \tag{9.21}$$

Under extrinsic conditions, D depends on the doping concentration, and since the concentration changes with distance (i.e., a concentration gradient exists), D also changes with distance. To obtain a solution to Fick's Law one can expand Eq. 9-5 to read:

$$\frac{\partial C(x,t)}{\partial t} = \frac{\partial}{\partial x}\left(D\frac{\partial C}{\partial x}\right) = \frac{\partial D}{\partial x}\frac{\partial C}{\partial x} + D\frac{\partial^2 C}{\partial x^2} \tag{9.22}$$

The $\partial D/\partial x$ term makes Eq. 9-22 an inhomogeneous differential equation making it difficult to solve. Fortunately, a graphical procedure termed the *Boltzmann-Matano analysis* has been developed, which renders a solution for the diffusion coefficient, D, as a function of concentration.[7] It does not, however, explicitly provide the diffusion profile C(x,t), as was obtained for the intrinsic diffusion case. The Boltzmann-Matano method allows D(C) to be calculated from the experimentally determined concentration profile (C vs. x plot). There are several methods that are commonly used to obtain the concentration profile and they are discussed later in this chapter.

9.2 DEFECTS AND DOPANT DIFFUSION

It has been realized for some time that the diffusion of dopant atoms in Si is strongly coupled to the presence of point defects. This early realization came about when it was found that the diffusion coefficient of the dopant atoms did not follow the behavior predicted by Fick's Laws, but rather exhibited certain anomalous effects. These effects will be described in detail in later sections but are briefly mentioned here. The first anomaly was that the diffusivity of boron was greatly enhanced when phosphorous was subsequently diffused into the Si (the so-called *emitter push effect*). The second was that some dopants exhibited a higher diffusivity when the drive-in diffusion was performed in an oxidizing ambient (*oxidation-enhanced diffusion* or *OED*). Since these effects were initially observed, other anomalous effects have been found, and as a group their behavior can be explained by the presence of point defects in the silicon. Consequently, the types and interaction of point defects and dopants in silicon are now discussed.

9.2.1 Point Defects in Silicon

Chapter 2 described some of the point defects that can exist in silicon, particularly at the temperatures normally associated with diffusion. Some of these defects are now briefly reviewed and how they relate to the diffusion of dopant atoms in silicon is discussed. The nomenclature used in the excellent treatise on diffusion by Fahey, Griffin, and Plummer is employed.[8]

The first type of point defect to consider is the defect which arises from the introduction of impurity atoms into silicon. An impurity atom, such as a *dopant* atom that sits on a lattice site, is termed a *substitutional defect.* Even though the atom resides on a lattice site it is either larger or smaller than the silicon atom, and thereby causes a local perturbation in the periodicity of the lattice. Any perturbation to the lattice periodicity is termed a *defect.* It turns out that most of the common dopant atoms (P, As, Sb, B, In, Ga) dissolve into silicon as substitutional atoms. That is, they reside on silicon lattice sites in place of silicon atoms. These dopant atoms are also remarkable for their high solubility in silicon when compared to other elements that also dissolve in silicon. The solubilities of these dopants are discussed in Chap. 1.

A dopant atom on a substitutional site is given the designation A and can be either a donor or an acceptor. When a dopant atom occupies a substitutional site it contributes either an electron to the conduction band, or a hole to the valence band. In either case, the dopant atom then becomes an ion. It is the ion (a charged atom) that must diffuse in the silicon. It is not surprising that since the dopant atom (or ion) is charged it would interact with other charged species (e.g., crystalline point defects and free carriers).

As discussed in Chap. 2, there are other point defects that may be present at the diffusion temperature, namely: a) *vacancies*; b) *interstitials*; and c) *interstitialcies*. Such *lattice defects* are designated as X, where X can be either a vacancy or interstitial-type defect. The vacancy defect is defined as a missing Si atom or an empty lattice site and is designated as V. The presence of a vacancy requires the covalent bonds around that site to reconfigure to accommodate the defect. The V defect itself can become charged by capturing one or two electrons (or a hole), or can remain neutral. As discussed in Chap. 2, four vacancy defects have been identified:

1. The neutral vacancy, V^0
2. The singly-charged negative vacancy (in which one electron is captured), V^{-1}
3. The doubly-charged negative vacancy (in which two electrons are captured), V^{-2}
4. The singly-charged positive vacancy (in which one hole is captured), V^{+}

Figures 9-5a1, 9-5a2, and 9-5a3 show how the unsatisfied bonds reconfigure to accommodate the various defects in the silicon lattice.

The next type of defect discussed is the *self-interstitial*, which can be described as a Si atom on an *interstitial* site. Self-interstitials are given the designation I. The term self-interstitial is used to distinguish these defects from those involving extrinsic or foreign atoms (which can also exist in the lattice as interstitials). Figures 9-5b1 and 9-5b2 show the bonding around a tetrahedral and a hexagonal interstitial site, respectively. The last type of defect (not often recognized), is called the *interstitialcy*. The interstitialcy is also designated by I. The interstitialcy defect differs from the interstitial defect in that it has two "associated" atoms in non-substitutional sites, around a substitutional site. The two atoms constituting the interstitialcy can consist of: a) both Si atoms (an I-defect); or b) a Si atom and a dopant atom (an AI-defect). The interstitialcy defects can remain neutral or become charged, and have the following configurations:

1. A neutral interstitialcy with two silicon atoms, I^0.
2. A singly-charged positive interstitialcy comprised of two silicon atoms, I^{+1}.
3. An associated dopant and silicon interstitial that is neutral, $(AI)^0$. This type of defect behaves as an acceptor site.
4. An associated dopant and silicon interstitial that is singly-negatively charged, $(AI)^{-1}$. This type of defect also behaves as an acceptor site.
5. An associated dopant and silicon interstitials that is singly-positively charged, $(AI)^{+1}$. This type of defect behaves as a donor site.

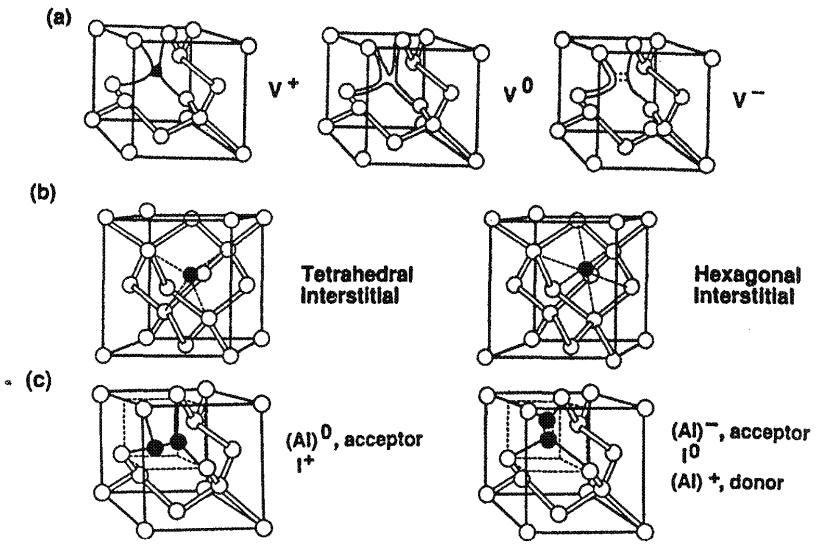


Fig. 9-5 Schematic of how the bonds may reconfigure to accommodate various defects in the silicon lattice: (a) vacancy defects; (b) interstitial defects; and (c) interstitialcy defects. If both dark spheres in (c) are silicon, the defect is the (I) defect. But if one is silicon and one is a dopant atom, the defect is (AI).[8] Reprinted with permission of the American Physical Society.

Figures 9-5c1 and 9-5c2 show the atomic arrangement of interstitialcy defects. The distinction between the interstitial and the interstitialcy is often not emphasized since it is difficult to discern between the two defect types by experimentation. As a result, both of these defects are referred to as *interstitial-type* defects.

Now that the three defects (i.e., the vacancy V, the interstitial-type I, and the dopant impurity A) have been described, it is appropriate to investigate the interaction among these defects, as this will help in understanding how such interactions may affect diffusion.

The following associated do-pant-atom/defect configurations may exist in the lattice: a) a dopant atom is on a lattice site (A); b) a doping atom is adjacent to a vacancy (AV); and c) a dopant atom is one of the atoms of the interstitialcy (AI). If the dopant atom actually occupies the interstitial site, it is designated A_i. It is the dopant-atom/defect pair that migrates during diffusion in one or more of the following configurations: AV, AI, or A_i. The coordinated dopant-atom/defects can be formed by the following reversible equilibrium reactions. The "Law of Mass Action" applies to these reactions.

$$I + A \leftrightarrow AI \tag{9.23a}$$

or:

$$I + A \leftrightarrow A_i \tag{9.23b}$$

$$V + A \leftrightarrow AV \tag{9.24}$$

$$I + AV \leftrightarrow A_i \tag{9.25}$$

$$V + AI \leftrightarrow A \quad \text{or} \quad V + A_i \leftrightarrow A \tag{9.26}$$

There is a further equilibrium reaction that should also be mentioned, and that is the creation and annihilation of vacancies and interstitials via the Frenkel defect:

$$0 \leftrightarrow V + I \quad \textbf{(9.27)}$$

The $\leftrightarrow$ symbolizes that the reaction is reversible, and since the concentrations of I and V can be affected by external forces (i.e., point defects arising from other sources), these reactions may be favored in one or the other direction. For example, if a large concentration of interstitials is created by oxidation, then the forward direction of Eq. 9-23 is favored, and the concentration of AI increases. The increase in AI concentration results in increased dopant diffusion.

A brief description of Eqs. 9-23 to 9-27 is given next. The reaction in Eq. 9-23 represents an exchange between a substitutional atom A and the interstitial(cy) I. It is termed the *kick-out* reaction, since the Si-I kicks out a dopant atom from a substitutional site and forms an AI pair. The reaction in Eq. 9-24 describes the dopant-vacancy reaction and governs vacancy-controlled diffusion. The reactions given in Eqs. 9-25 to 9-27, describe recombination reactions between I and V defects. The reaction in Eq. 9-26 describes a dissociative reaction termed the *Frank-Turnbull* mechanism. This mechanism is not considered to be important for Si diffusion since it is believed to have a low probability of occurrence. Finally, the forward reaction of Eq. 9-27, describes the formation of a Frenkel defect (vacancy and interstitial) as is described in Chap. 2.

In the past it was considered that dopant diffusion in silicon was controlled solely by the vacancy mechanism, but now it is well established through numerous investigations that both I- and V-type mechanisms contribute to dopant diffusion. As it turns out, most of the common dopants have a dominant I-diffusion-controlled mechanism. The evidence for this conclusion is discussed in detail in a later section.

The reactions dealt with above do not give the origin of the defects. There are two possibilities of defect formation that were described in Chap. 2, namely: a) an atom leaving a lattice site and creating a V and I, and b) a silicon atom diffusing from the surface of the silicon into the bulk to create an interstitial or a silicon atom that was on a lattice site moving to the surface (and vanishing by becoming part of the disorder at the surface) and creating a vacancy in the bulk. Another important source of defect production is lattice damage resulting from ion implantation. As a result, the diffusion of dopants from an ion implanted source can substantially differ from that of a chemical source.

9.2.2 Temperature Dependence of the Diffusion Coefficient Under Intrinsic Conditions

The temperature dependence of the diffusion coefficient, D, for intrinsic diffusion for the common dopant atoms will now be examined. However, before we proceed, it is instructive to first inspect the temperature dependence of the self-diffusion of silicon. Most self-diffusion studies used radio-active isotopes of silicon to determine the diffusion profile. Even to this day, however, the basic mechanism of Si self-diffusion is not well understood.

Some results from a study that examined the effect of hydrostatic pressure on Si self-diffusion show the dominant mechanism is interstitial-type-dominated diffusion.[9] The most interesting aspect about silicon self-diffusion is that the activation energy for the process is about 1 eV greater than that of dopant diffusion. For some reason the movement of silicon-defect pairs takes more energy than that of dopant-defect pairs. The diffusion coefficient for Si self-diffusion D_{self} is given by:[10]

$$D_{self} = 1400 \exp(-5.01 \text{ eV}/kT) \text{ cm}^2/\text{sec} \quad \textbf{(9.28)}$$

Table 9-2 COMPILED VALUES OF APPARENT ACTIVATION ENERGIES FOR DIFFUSION OF COMMON DOPANTS[8]

Donors	Q_A(eV)	Acceptors	Q_A(eV)
P	3.51–3.67	B	3.25–3.87
As	4.05–4.34	Al	3.36
Sb	3.8–4.05	Ga	3.75
		In	3.60

9.2.3 Intrinsic Diffusion Coefficients

Numerous experimental measurements have been made on the effect of temperature on the diffusion coefficients (D_A) of the common dopants during intrinsic diffusion. In general the resulting data can be represented by an Arrhenius equation.

$$D_A = D_{Ao} \exp(-Q_A/kT) \quad (9.29)$$

Here D_A is the diffusion coefficient of the dopant atom, D_{Ao} is the pre-exponential term or *frequency factor*, which is related to the *frequency* of the lattice vibrations (i.e., the frequency at which atoms strike the potential barrier that they must overcome to move in the lattice), and Q_A is the *apparent activation energy* of the diffusion process (which is related to the *height* of the energy barrier that the impurity must overcome in order to move within the lattice).

The value of Q_A for the common dopants has been compiled by Wohlbier.[11] This data is reproduced in Table 9-2, which was extracted from the summary of Wohlbier's work, given by Fahey, Griffin, and Plummer.[8] Readers interested in additional details are referred to that article, which lists the original references, or to the Wohlbier monograph. Although the values of apparent activation energy vary somewhat, depending on the particular study and dopant, all of the values range between 3.3 to 4.3 eV.

9.2.4 Fast Diffusers in Silicon

There is another class of impurities that diffuse in silicon much more rapidly than the electrically active dopant atoms. These impurities are called the *fast diffusers* and their diffusion coefficients at 900°C are listed in Table 9-3. The fast diffusers, as evidenced by their name, diffuse much faster in silicon than do the dopants atoms. The fast diffusers (e.g., Fe, Cu, Pt) move great distances in silicon during high temperature processing, thereby increasing their probability of being trapped at an extended defect and causing increased leakage current in a device. The effect of the fast diffusers on device properties is discussed in Chap. 2. The apparent activation energies (0.2–2 eV) for the fast diffusers are much lower than for the substitutional impurities. At one time it was believed that the low values of activation energy implied interstitial diffusion, while the high activation energy values of the dopant atoms were due to vacancy diffusion. This analysis is not correct, and it is now well established that both vacancies and interstitials play a role in the movement of fast diffusers in silicon.

9.3 ATOMISTIC MODELS OF DIFFUSION

Some of the possible ways a dopant atom may move in the crystalline lattice are next explored from an atomistic perspective. The most favorable configuration for a dopant atom to diffuse in silicon is when it is associated with a nearest-neighbor defect. Only a fraction of the dopant atoms are associated with defects in this manner, but it is this fraction that contributes to the

Table 9-3 DIFFUSION COEFFICIENTS of FAST DIFFUSANTS in SILICON @ 900°C

Element	D (cm^2/sec)	Element	D (cm^2/sec)
Na	1.6×10^{-3}	Ag	2.0×10^{-3}
K	1.1×10^{-3}	Pt	1.6×10^{-2}
Cu	4.7×10^{-2}	Fe	6.2×10^{-3}
Au	1.1×10^{-3}	Ni	1.0×10^{-1}
Au_i	2.4×10^{-4}	O_2	7.0×10^{-2}
Au_S	2.8×10^{-3}	H_2	9.4×10^{-3}

diffusion. This fraction has been calculated as:[12]

$$C_{AX} = \theta_{AX} \frac{C_A C_X}{C_S} \exp\left(\frac{E_{AX}^b}{kT}\right) \quad \textbf{(9.30)}$$

where C_A is the dopant concentration; C_X is the unassociated defect concentration; C_S is the concentration of lattice sites ($5 \times 10^{22}/cm^3$), E_{AX}^b is the binding energy of the AX defect. The θ_{AX} factor accounts for the number of equivalent ways of forming the defect AX at a particular site (e.g., $\theta_{AV} = 4$). There have been numerous investigations to determine the value of E_{AX}^b and the reader is referred to the Fahey treatise for further detail.[8]

There are several potential ways dopant atoms may diffuse in the lattice, namely: a) direct dopant exchange (no point-defect interaction); b) vacancy-dopant interaction; c) interstitialcy-dopant interaction; and d) interstitial-dopant interaction. The interstitial and interstitialcy mechanisms are very similar and it is difficult to discern which is actually operative during diffusion.

First consider the possibility that mechanism (a) is responsible for controlling diffusion in Si. That is, what is the likelihood that diffusion proceeds by substitutional dopant atoms exchanging positions with adjacent silicon atoms (the *direct exchange* mechanism, shown schematically in Fig. 9-6a)? For diffusion to continue after the first interchange, the dopant atom must exchange positions again with the next Si atom. If the dopant atom instead exchanges positions with the original Si atom it will not make progress diffusing within the lattice. The continuation of the direct interchange process is unlikely because the activation energy associated with the breaking of Si bonds is high, and at least six Si bonds must be broken for each exchange to occur.

On the other hand, it is much more energetically favorable if the substitutional impurity is adjacent to a silicon vacancy and the exchange happens between the vacancy and the dopant atom. When this exchange occurs, it is termed the *vacancy mechanism* ($A + V \leftrightarrow AV$), as schematically shown in Fig. 9-6b. The uncomplicated motion involved in the vacancy mechanism led early investigators to believe that the vacancy interchange should dominate dopant diffusion in silicon. This belief was supported by the fact that vacancies dominate diffusion in metals. Consequently, detailed models were developed to explain diffusion in silicon according to the vacancy mechanism.[12,13]

Diffusion via the vacancy mechanism, however, does not occur by a single interchange of a vacancy with a dopant atom. If that was indeed the case, there would be no net diffusion. That is, in order for the dopant to diffuse over a long range, the vacancy must diffuse away from the dopant atom at least to the third nearest neighbor. Only in this way can it return to the dopant atom via a different path, thus allowing diffusion to continue further along. It can be shown that

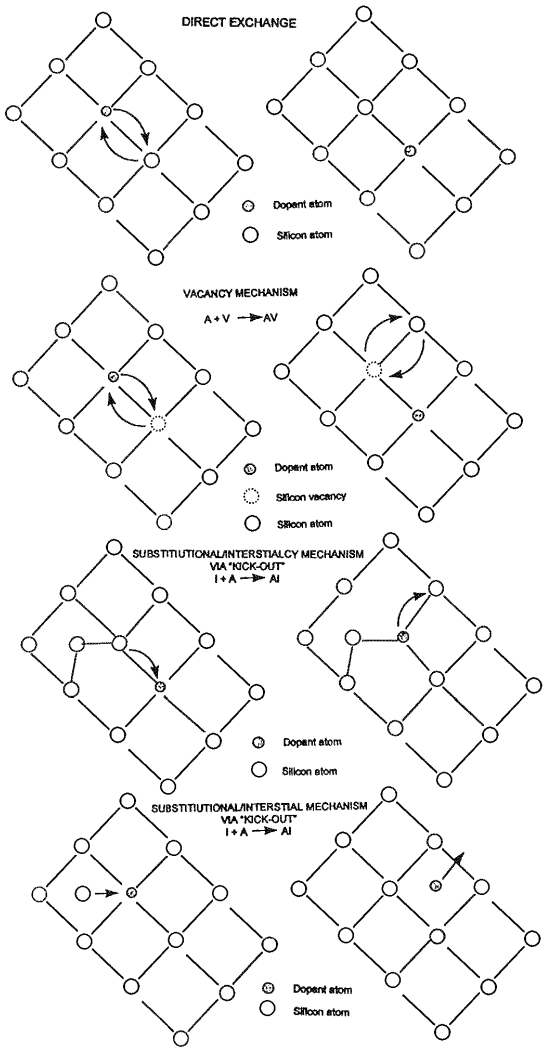


Fig. 9-6 Dopant movement at the atomic level: a) direct exchange; b) vacancy mechanism; c) substitutional-interstitialcy mechanism via "kick-out"; d) substitutional-interstitial mechanism via "kick-out".[8] Reprinted with permission of the American Physical Society.

the activation energy for the diffusion of the AV associated pair is less than that of A diffusing by itself, and as a result the vacancy mechanism is more energetically favorable than the dopant atom diffusing by direct interchange.

The *interstitialcy mechanism* (I + A ↔ AI) can be represented on an atomic level as a dopant atom in a substitutional site that is approached by a silicon interstitialcy (Fig. 9-6c).[14] Recall that the interstitialcy defect has two associated atoms in non-substitutional sites around a substitutional site. In the case of the silicon interstitialcy both atoms are silicon. One of the silicon interstitial atoms knocks a dopant atom out of a substitutional site and forms a dopant interstitialcy. The dopant atom then moves toward a silicon atom on a lattice site and *kicks out* the silicon atom and forms a reconfigured dopant-interstitialcy. Unlike the vacancy mechanism, which requires that the AV pair partially dissociate, the interstitialcy mechanism does not have the same restrictions. The energy for this type of motion has been determined to be less than for self-diffusion by about ΔQ_{AI} (the attractive potential between A and I), where ΔQ_{AI} is ~1 eV.

The final atomistic mechanism used to explain dopant diffusion is the "pure" *interstitial mechanism* ($I + A_s \leftrightarrow A$). In this mechanism (see Fig. 9-6d), a silicon interstitial replaces a dopant atom on a substitutional site to form a dopant interstitial. The dopant interstitial moves through the interstices of the silicon lattice until it finds a substitutional site to occupy. When the dopant atom occupies the substitutional site it forms another silicon interstitial. This mechanism is very similar to the interstitialcy mechanism discussed above and as a result it is difficult to separate it from the interstitialcy mechanism.

The general consensus among investigators in the field regarding the atomistic mechanism controlling the movement of the common dopants in silicon is as follows: Phosphorus and boron have the largest amount of interstitial-type diffusion, while arsenic exhibits both vacancy and interstitial diffusion, and antimony seems to be dominated by vacancy-controlled diffusion. How these conclusions were reached is discussed in a subsequent section.

9.4 DIFFUSION MODELLING

It is important to be able to predict the movement of dopant atoms in Si during thermal processing with mathematical simulations. Simulation of the diffusion and redistribution of the dopant allows process development and integration engineers to substantially reduce the number of process experiments required during technology development. There are several simulation programs that are commonly used, with the most common being SUPREM (*Stanford University PRocess Engineering Model*). The modeling of diffusion in one-dimension is covered by SUPREM III, while the extension to two-dimensions is covered by SUPREM IV. SUPREM IV, of course, is also capable of treating the one-dimensional case. Some of the models that are used in SUPREM III and their extension to SUPREM IV will now be discussed.

9.4.1 SUPREM III

The models in SUPREM III are empirically determined and are derived from the observed dopant-diffusion interactions in silicon. The basis of the diffusion model in SUPREM III is that diffusion is governed by the one-dimensional continuity equation, which accounts for the atom flux from the concentration gradient plus the enhanced flux that results from the built-in electric field formed from ionized impurities in a concentration gradient, and is given by Eq. 9-31:

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left(D\frac{\partial C}{\partial x}\right) \pm \frac{q}{kT}\left[\frac{\partial}{\partial x}\left(DC_i\frac{\partial \phi}{\partial x}\right)\right] \qquad (9.31)$$

mechanism; c) mechanism via

Where D is the dopant diffusion constant, and C and C_i are the total and electrically charged dopant concentrations, respectively. The potential ϕ that results from the concentration gradient is given by:

$$\phi = kT/q\ [\ln(n/n_i)] \tag{9.32}$$

where n is the carrier concentration resulting from dopant ionization and n_i is the intrinsic carrier concentration at the diffusion temperature. In many real problems the local charge neutrality is maintained and the second term on the right hand side of Eq. 9-31 can be dropped, resulting in Fick's Second Law. More simply, if D does not depend on the concentration of the dopant then the simplified Fick's Second Law holds as expressed in Eq. 9-33, below, which is the same as Eq. 9-6.

$$\frac{\partial C(x,t)}{\partial t} = D\left[\frac{\partial^2 C(x,t)}{\partial x^2}\right] \tag{9.33}$$

If the values of D are known, the solutions to Eqs. 9-31 and 9-33 will provide accurate simulations of the dopant concentrations for many applications.

There are two approaches used to obtain the values of D. In the first approach the value of D is determined from measured diffusion profiles and then plotted for various temperatures and doping concentrations. Solutions to Eq. 9-31 and 9-33 are then fitted to the experimental plots by finding values of D that give solutions that match the D values extracted from the observed data. This type of model is known as a *phenomenological diffusion model*, because it is based on observation.

The second type of model assumes that the correct values of D can be obtained directly from the physics of the dopant-defect interaction. This type of model is referred to as a *point-defect-based diffusion model.* One such advanced model uses a kinetic Monte Carlo simulation to obtain the formation and diffusion energies associated with the defects, dopants, and impurities. The simulations, however, are number-crunching intensive and require a high-end workstation or a supercomputer to be able to get results in a reasonable time. The attraction of these models, however, is that they do not require diffusion profiles and empirical curve-fitting to predict the outcome. The description of such models is beyond the scope of this text and the interested reader is referred to the literature.[15,16]

Phenomenological models have been implemented into SUPREM III and other 1-dimensional process simulators, such as PREDICT and RECIPE. These models furnish an adequate fit to 1-dimensional diffusions, but are not successful at predicting the two-dimensional diffusion so critical for advanced devices with small feature sizes. The two-dimensional case requires the use of SUPREM IV. Also the phenomenological models are not adequate at predicting diffusion behavior of such defect-dominated phenomena as oxidation-enhanced diffusion (OED).

9.4.2 SUPREM III Models for Boron, Arsenic, Phosphorus, and Antimony Diffusion

The models used in SUPREM III are based on the *vacancy model under non-oxidizing conditions*, proposed by Fair and Tsai.[17] These models, however, do not accurately reflect what is occurring on an atomic scale (since diffusion is related to both dopant/interstitialcy and dopant/vacancy interactions). However, the Fair-Tsai models are nonetheless very useful, since they provide an accurate representation of the diffusion profile for the common dopants.

The Fair-Tsai Model assumes that the diffusivity of an ionized dopant atom is based on the sum of the diffusivities of neutral vacancies and ionized vacancies, weighted by the probability

of their existence. According to the model there are four possible states of a vacancy: 1) the neutral vacancy V^0; 2) the single-negatively-charged vacancy V^{-1}; 3) the double-negatively-charged vacancy V^{-2}; and 4) the single-positively-charged vacancy V^{+1}. During *extrinsic diffusion* each contribution must be modified by the ratio of the doping level to the intrinsic carrier concentration raised to the power (including the correct sign) of the charge state. For example the contribution of the *double-negatively-charged* vacancy is modified by $(n/n_i)^2$, while that of the *single-positively-charged vacancy* by (n_i/n). Thus, the effective diffusion coefficient under non-oxidizing conditions can be calculated from the sum of all the individual vacancy components. The effective D is given by:

$$D = D^0 + D^-\left(\frac{n}{n_i}\right) + D^{2-}\left(\frac{n}{n_i}\right)^2 + D^+\left(\frac{n_i}{n}\right) \quad \textbf{(9.34)}$$

where the individual diffusivities on the right hand side of the equation correspond to the interaction between dopant atoms and neutral or charged vacancies.

9.4.3 Modeling Intrinsic Diffusion

Intrinsic diffusion is dominant when $n \leq n_i$ at the diffusion temperature. In such cases the effective D is independent of dopant concentration and depends only on the diffusivities of the individual defects. Thus, when intrinsic diffusion is being modeled, Eq. 9-34 is re-written as:

$$D^i = D^0 + D^{-1} + D^{-2} + D^{+1} \quad \textbf{(9.35)}$$

and D^i is referred to as the *intrinsic effective diffusion coefficient*. The models used to determine D^i for boron, arsenic, phosphorus, and antimony are examined next.

9.4.3.1 Boron: Since boron diffuses as a negatively charged "atom" the model assumes that diffusion occurs primarily by the interaction of the boron with the *neutral* and *positively* charged vacancies. In that case the intrinsic diffusion coefficient of boron $D_B{}^i$ can be represented by:

$$D_B{}^i = D_{BV}{}^0 + D_{BV}{}^{+1} \quad \textbf{(9.36)}$$

where:

$$D_{BV}{}^0 = 0.037 \exp(-3.46\ \text{eV}/kT) \quad \text{cm}^2/\text{sec} \quad \textbf{(9.37a)}$$

and:

$$D_{BV}{}^{+1} = 0.72 \exp(-3.46\ \text{eV}/kT) \quad \text{cm}^2/\text{sec} \quad \textbf{(9.37b)}$$

resulting in Eq. 9-38 for the total intrinsic diffusivity for boron $D_B{}^i$:

$$D_B{}^i = 0.76 \exp(-3.46\ \text{eV}/kT) \quad \text{cm}^2/\text{sec} \quad \textbf{(9.38)}$$

9.4.3.2 Arsenic: The intrinsic diffusion of arsenic is assumed to be controlled by neutral and negative vacancies and the intrinsic diffusion coefficient for As ($D_{As}{}^i$) is given by:

$$D_{As}{}^i = D_{As}{}^0 + D_{As}{}^{-1} = 0.066 \exp(-3.44\ \text{eV}/kT) + 12.0 \exp(-4.05\ \text{eV}/kT) \quad \text{cm}^2/\text{sec} \quad \textbf{(9.39)}$$

9.4.3.3 Phosphorus: The intrinsic diffusion of phosphorus is assumed to be controlled by the interaction of the impurity ions with neutral vacancies only. Thus, $D_P{}^i$ is given as:

$$D_P{}^i = D_P{}^0 = 3.85 \exp(-3.66\ \text{eV}/kT) \quad \text{cm}^2/\text{sec} \quad \textbf{(9.40)}$$

9.4.3.4 Antimony: The intrinsic diffusion of antimony is assumed to be dominated by interaction between neutral and single negatively charged vacancies. Thus, $D_{Sb}{}^i$ is given by:

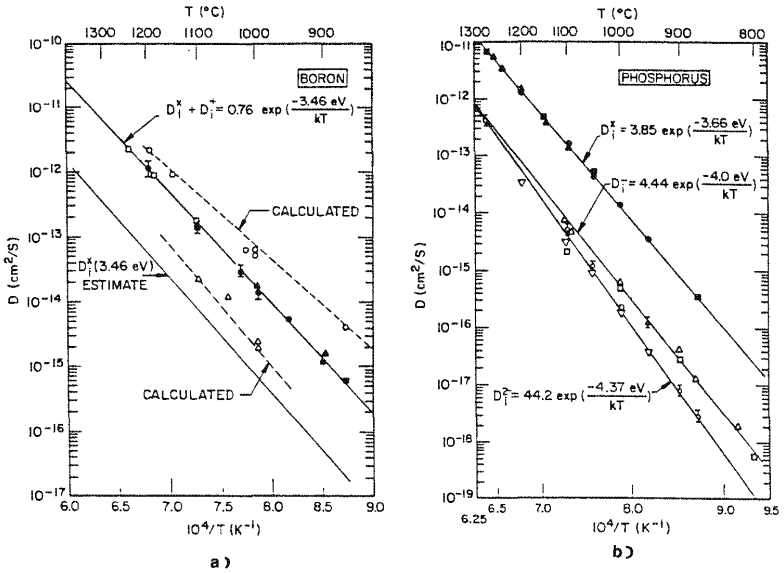


Fig. 9-7 Intrinsic diffusion coefficients vs. temperature for: a) boron; b) phosphorus.[13] Reprinted with permission of North Holland Publishing Co.

$$D_{Sb}^{i} = D_{Sb}^{o} + D_{Sb}^{-1} = 0.214 \exp(-3.65\ eV/kT) + 13 \exp(-4.08\ eV/kT)\ cm^2/sec \quad (9.41)$$

Figures 9-7 (a–d) show the Arrhenius plots of the intrinsic diffusion coefficients of B, P, As, and Sb. As can be seen from the plots, the agreement between the measured data and the calculated curves are quite good.

9.4.4 Extrinsic Diffusion Coefficients for the Common Dopants (As, P, and B)

An extensive amount of work has been done relating to the modeling of extrinsic diffusion of the common dopants As, P, and B. The earliest models developed by Fair were phenomenological expressions based on the result of careful determination of diffusion profiles.[20] However, these models did not account for the effects of both I and V on dopant diffusion. Uematsu developed an integrated diffusion model based on measurable parameters that can predict the diffusion profiles of these dopants. The model considers contributions from the vacancy mechanism (V), the kick-out mechanism (I), and the Frank-Turnbull mechanism (F-T).[68] The details of the Uematsu model are beyond the scope of this text and the reader is referred to the original paper. Here only the contributing defects for each of the common dopants are noted.

9.4.4.1 Arsenic: Figure 9-8 shows a typical diffusion profile for arsenic. There are several features that are common to As diffusion profiles. First, arsenic is a relatively slow diffuser and shows a relatively abrupt profile. Such features are desirable when fabricating shallow junctions for advanced CMOS and bipolar devices. Consequently arsenic is the most widely used *n*-type dopant in ULSI processing and is the mainstay for the *n*-channel shallow source/drain diffusion

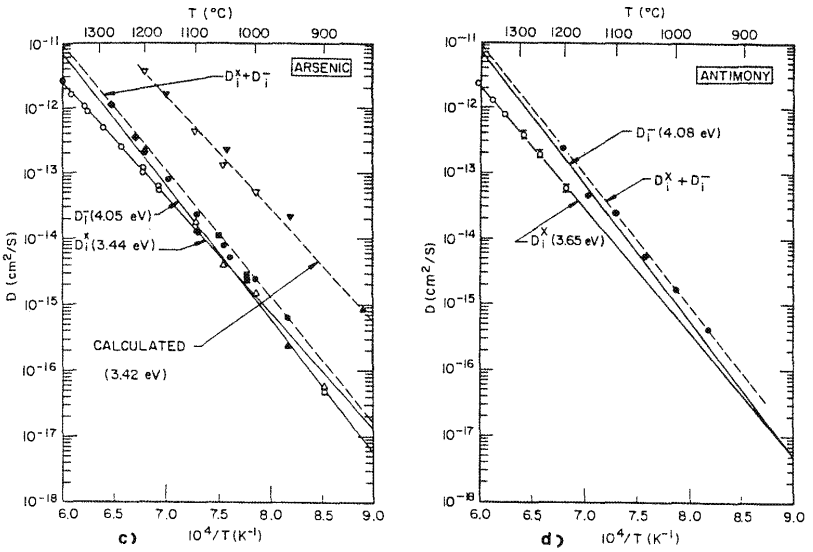


Fig. 9-7 Intrinsic diffusion coefficients vs. temperature for: c) arsenic; d) antimony.[13] Reprinted with permission of North Holland Publishing Co.

in advanced CMOS devices and the emitter diffusion in bipolar and BiCMOS devices. The second observation is that the total As concentration (C_{As}) is frequently greater than the electrically active (C_T) concentration, particularly at high doping levels. It has been found that the amount of inactive As increases with the fourth power of the active concentration (i.e., $C_{As} - C_T \propto C_s^4$).[69] Various models have been proposed to explain this phenomenon, including As-vacancy complex, small As clusters, and As-precipitate formation and dissolution. The arsenic-

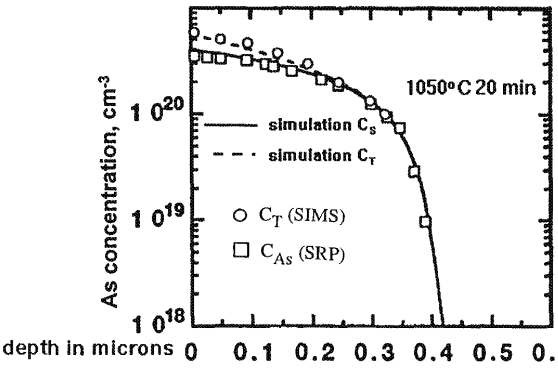


Fig. 9-8 Arsenic diffusion profile. Note the steep diffusion front and the difference between the total As concentration C_T and the active As concentration, C_{As}, for high As concentrations. Excellent agreement exists between simulated and measured profiles.[68] Courtesy of American Physical Society.

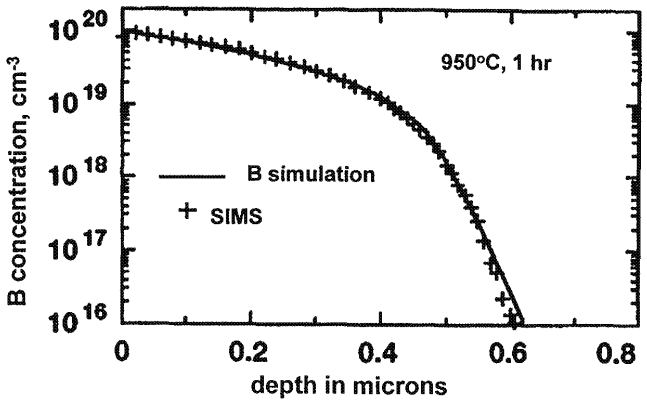


Fig. 9-9 Boron diffusion profile showing agreement between simulated and measured (SIMS) profiles.[68] Courtesy of American Physical Society.

vacancy cluster model appears to have the most support. Regardless of the mechanism, the fact that the amount of electrically-active As may change during processing must be taken into account. To obtain meaningful As profiles both SIMS (for C_T) and SRP (for C_{As}) should be used.

As discussed previously As diffusion has nearly equal contributions from both I and V. Unlike boron diffusion, which is controlled by the presence of silicon self-interstitials, the I-component in As is controlled by neutral As-interstitials (As_i^o). The vacancy contribution is dominated by the presence of V^{-2} and V^{-1}. The complex formation (As complexes) is dominated by the neutral vacancy complex, $(As_2V)^o$. Excellent agreement is observed between this model and experimental results as shown in Fig. 9-8.

9.4.4.2 Boron Diffusion: Figure 9-9 shows a typical profile for B diffusion. B is a relatively fast diffusant, and consequently great care must be exercised when attempting to fabricate shallow *p*-channel source/drain regions. The diffusion of boron from an implanted source is further increased by transient enhanced diffusion (discussed below). The formation of shallow boron diffusions is a fertile area of research and several techniques are candidates to achieve such shallow junctions. The most prominent approach is the use of *ultra-low-energy* (*ULE*) ion implantation (<1 keV), followed by high-ramp-rate rapid thermal annealing (see Chap. 10). Other approaches such as P-GILD, solid-source diffusion, plasma doping, and RTP gas doping are also under consideration.

Boron diffusion is known to be dominated by the "kick-out" mechanism. The diffusion profile can be modeled when three parameters are used, namely, neutral B interstitial $(B_i)^o$, a neutral self-interstitial (I^o), and a positive self-interstitial (I^+). As shown in Fig. 9-9, excellent agreement between the model and experimental results have been obtained.

9.4.4.3 Phosphorus Diffusion: Although phosphorus is less commonly used in advanced devices, it still finds some application for the source/drain contacts in CMOS and for emitter diffusion in bipolar devices. Phosphorus diffuses quite normally in the intrinsic regime and is a relatively fast substitutional diffuser. When the dopant concentration exceeds the intrinsic carrier concentration, however, the in-diffusion behaves anomalously. By examining the P-diffusion profile in Fig. 9-10 one observes that at high concentrations the diffusion profile becomes quite complex and has a shape with three distinct regions: a) the *high concentration* or plateau region;

b) the *transition* or *kink* region; and c) the low concentration, or *tail* region. Figure 9-10 provides a schematic representation of these regions. No simple equation has been found to fit the profile and numerous models have been expounded to explain this strange behavior.

The most well known model was proposed by Fair and Tsai.[70] A key element of that model is that a singly-charged *phosphorus ion-vacancy pair*, $(PV)^{-1}$ controls phosphorus diffusion. This defect is believed to have a negative charge, and is present in the high concentration region. The defect pair becomes important when it dissociates as it diffuses into the kink and tail regions. The dissociation is believed to create excess vacancies, which in turn enhances the phosphorus diffusion in those regions of the profile. The problem with this model is that I-defects are known to dominate P diffusion. Hu *et al.*, described a model for the kink and tail regions that results from a two-stream (interstitial and vacancy) diffusion process.[71]

More recently, Uematsu applied the unified model to explain anomalous P diffusion.[68] In his model the diffusion is described in terms of an interstitial and a vacancy mechanism. For high surface concentration the vacancy mechanism (by way of a V^{-2}) governs the diffusion in the plateau region, while the kick-out mechanism through I and P_i^o dominates in the kink and tail regions. The changeover from the vacancy to the kick-out mechanism is what is believed to be responsible for the kink and tail profiles observed in P diffusion. For low phosphorus surface concentrations only the kick-out mechanism is operative and no kink and tail are observed.

9.4.5 Modeling Diffusion with SUPREM IV

The most advanced diffusion model in SUPREM IV is a physics-based model that depends on the interaction of point defects (silicon self-interstitials and vacancies) and the diffusant dopants. This model is capable of predicting diffusion in an oxidizing ambient. SUPREM IV is currently supported by the commercial suppliers Avant!/TMA (under the name TSUPREM4) and Silvaco, Inc.[18] The advanced model in SUPREM IV was first developed to solve the coupled oxidation and diffusion problem while using the same grid. Readers interested in the equations used in this model are referred to the suppliers of these programs or the literature.

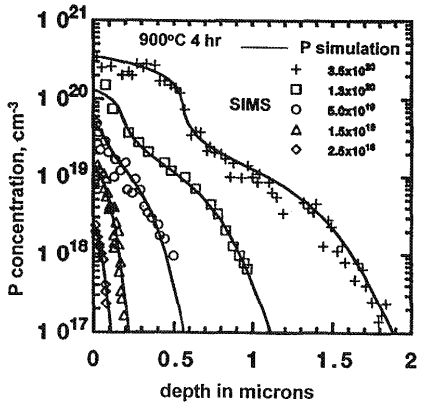


Fig. 9-10 Phosphorus diffusion profile for several surface concentrations. Note the presence of the "kink" and "tail" regions only for the high concentration diffusions. Excellent agreement exists between simulated and measured (SIMS) profiles.[68] Courtesy of American Physical Society.

Two-dimensional modeling of diffusion is simulated by calculating the local diffusion constants based on the point defect and impurity concentrations at the location of interest.[21] The more accurately the local defect concentration is known, the more accurately the local value of diffusion can be determined. Since the concentration of the point defects may depend on the geometry at a local region, it may be necessary to simulate the movement of the oxide interface and redistribution of the impurities if diffusion takes place in an oxygen-bearing environment.

To accurately model diffusion in two dimensions the following information is usually required:

1. The nature of the point defects,
2. How far the point defects move into the bulk,
3. How the point defects interact and recombine at the surface of the silicon and in the bulk regions,
4. The predominant mechanism that drives the diffusion (i.e, is it interstitial[cy] or vacancy).

The diffusion equations used in SUPREM IV are non-linear, since the value of D and built-in electric field depend upon the point defects and the carrier concentration (ionized dopants). Hence, a full solution requires the simultaneous solution of a set of non-linear coupled differential equations for the dopants and the defects.[22] SUPREM IV solves these equations using numerical methods. The solutions produce several models that are useful for obtaining diffusion profiles in one- and two-dimensions. Models from SUPREM IV are briefly discussed below.

The simplest fastest model to simulate is the *FERMI Model.* It assumes that the point-defect concentrations depend only on the position of the Fermi level in the silicon. The actual point-defect concentrations are not calculated in this model. The FERMI model is most useful when the speed of the simulation is more important than its accuracy. The model does not work for oxygen-enhanced diffusion (OED), high-concentration diffusion, or implant-damage effects.

The *TRANS Model* accounts for the two-dimensional simulation of point defects. This model includes such events as the generation of point defects at interfaces, the diffusion of point defects into the bulk, and the recombination of defects at interfaces and in the bulk. The model is capable of simulating OED, but not high-concentration effects (such as the phosphorus kink and tail). It is also capable of simulating implant damage but not as well as the so-called FULL model. The TRANS model is typically used for routine simulations.

The *FULL Model* is the most accurate and requires significant computer CPU time to perform simulations. The FULL model does everything the TRANS model does, plus it simulates the effects of dopant diffusion on point-defect concentrations. It includes such refinements as pair saturation and dopant-assisted recombination effects. This model simulates high-concentration effects, OED, and implant-damage effects. Since it requires significant computer time, use of the FULL model is only recommended when maximum accuracy is necessary, or when high-concentration effects or implant-damage effects are pertinent.

9.5 DIFFUSION IN POLYCRYSTALLINE SILICON

Polycrystalline silicon is used as the gate electrode (conductor) in CMOS devices, as the emitter in bipolar devices, and as high-value load resistor in SRAMs (see Chap. 6 for more information on polysilicon). CMOS devices require that the polysilicon be heavily doped (degenerate) for maximum conductivity and to minimize the so-called "poly depletion" effect. The doping level

of the polysilicon at the polysilicon/oxide interface has a strong effect on the device threshold voltage. In deep-submicron CMOS devices it is necessary to dope some of the polysilicon with boron and some with arsenic on the same circuit. In advanced bipolar applications the polysilicon serves as the emitter contact diffusion source. Most often the emitter is doped with arsenic, although phosphorus may be used in less advanced applications. High-value resistors require controlled and repeatable resistivity of the polycrystalline-Si films.

Impurity diffusion in polycrystalline thin films, in general (and in polysilicon in particular), is quantitatively different from diffusion in single crystal material. The difference arises as a result of the grain boundaries present in polycrystalline films. That is, bulk diffusion is the controlling mechanism in single crystal material, while grain-boundary diffusion predominates in polycrystalline films. The diffusivity along grain boundaries can be up to 100 times faster than the diffusivity in bulk material.

Polycrystalline-silicon films are made up of grains with sizes ranging from 50 to 300 nm. The diffusion of the dopants *within* the grains is comparable to that of single-crystal silicon. But, the dopant atoms also diffuse much more rapidly along the grain boundaries, and then diffuse into the grains. Since the grains are small, only a short time is required for the dopant (which is entering from all sides of the grain) to fully diffuse within the grain. As a result, the overall diffusion is controlled by the grain boundaries, the grain structure, and the preferred orientation of the film. These quantities, in turn, depend upon such deposition conditions of the film as temperature, deposition rate, thickness, and post-deposition annealing treatments.

Each measurement of the diffusion constant D in polysilicon depends very much on the film's history, and published results cannot be taken universally. Data has been obtained for the diffusivity of phosphorus, boron, and arsenic in polysilicon, and these diffusivity values are listed in Table 9-4.[23,24,25,26,27] It has been found that phosphorus and arsenic can precipitate at grain boundaries, resulting in reversible changes of resistivity upon annealing. That is, the precipitates can dissolve at elevated processing temperatures and go into solution in the grains, thereby lowering the resistivity of the film. The final resistivity of the film thus depends not only on the doping level in the grains, but upon the grain size, any precipitates at the grain boundaries, and the presence of other defects that can reduce the mobility of the carriers.

9.6 DIFFUSION IN SILICON DIOXIDE

Silicon dioxide serves as a diffusion mask for both chemical and ion-implanted predeposition. Its sole purpose, in these cases, is to keep the diffusant atoms away from regions where they are not wanted. Silicon nitride (Si_3N_4) can be used for similar purposes. The diffusivities of the

Table 9-4 SOME REPORTED DIFFUSION COEFFICIENTS in POLYSI FILMS[33-27]

Dopant	D_0(cm^2/sec)	E (eV)	D (cm^2/sec)	T (°C)
As	8.6×10^{4}	3.9	2.4×10^{-14}	800
As	0.63	3.2	3.2×10^{-14}	950
B	$1.5–6.0 \times 10^{-3}$	2.4–2.5	9.0×10^{-14}	900
B			4.0×10^{-14}	925
P			6.9×10^{-13}	1000
P			7.0×10^{-13}	1000

common dopants in SiO_2 can be calculated from measured profiles assuming solutions to Fick's second law, with the appropriate boundary conditions.

The Group III and V elements are known to form glassy networks with SiO_2 and accordingly their diffusivity strongly depends on their concentration. The diffusivities of these elements are very low for concentrations less than 1%, and generally do not need to be considered in detail.
A recent paper reports on the diffusion of phosphorus from a phosphorus vapor source into thermal oxides.[28] The diffusivity of phosphorus is given by:

$$D_p = 3.79 \times 10^{-9} \exp\left(-\frac{2.3}{kT}\right) \quad (cm^2/sec) \tag{9.42}$$

and it was found to be independent of the phosphorus concentration. The solubility of phosphorus in the glass, in the temperature range of 1000°C, was found to vary between 3×10^{20} and 2×10^{22} cm^{-3}. The diffusion mechanism is described as phosphorus dissolving in the interstitial sites as P_2, where it is incorporated into the network of SiO_2. The P_2 exchanges sites with silicon atoms and continues diffusing through the silicon sites.

9.6.1 Boron Penetration of Thin Gate Oxides

In the case of thin gate oxides, the diffusion of boron through the gate oxide must be considered. This effect is termed *boron penetration*. During boron penetration, dopant atoms from the heavily-boron-doped polysilicon can diffuse through the thin gate oxide layer and into the device channel, where its presence can then change the device threshold. Also, boron incorporated in the gate oxide during the diffusion can degrade the oxide breakdown characteristics and charge trapping rate.[29] It is further noted that the diffusion of boron through oxide is enhanced by the presence of hydrogen and fluorine in the oxide. Fluorine can be incorporated into the gate oxide if the boron is introduced into the polysilicon using BF_2 as the implantation source.

There are two main ways to reduce boron penetration or diffusion through a thin oxide. The most common approach is to incorporate nitrogen into the gate oxide. This technique is known to reduce boron diffusion by the nitrogen bonding in the glassy network, which in turn impedes the boron flux through the glass. A model has been developed to explain why the presence of nitrogen in the oxide hinders the boron diffusion.[30] The model explains that the value of D_0 decreases and that of Q_A increases as the nitrogen concentration increases.

The diffusivities (at 900°C) of the common dopants in SiO_2 are listed in Chap. 8, Table 8-5. There is a class of materials that are fast diffusants in SiO_2, including H_2, He, OH^{-1}, Na, O_2, and Ga. Values of D greater than 10^{-13} cm^2/sec have been determined for these elements.

9.7 ANOMALOUS DIFFUSION EFFECTS

Any deviation from in the diffusion behavior predicted by Fick's Law may be considered as *anomalous behavior*. The most common anomalous diffusion effects include: a) the effect of high dopant concentration on diffusion behavior (extrinsic diffusion), as discussed earlier; b) the enhancement of diffusion as a result of the built-in electric field; c) the effect of sequential diffusions including the "emitter push effect"; d) lateral diffusion under a window; e) oxidation-enhanced (and retarded) diffusion; and f) transient enhanced diffusion. To a large extent these phenomena are dependent upon the interaction of the dopant atoms with silicon point defects. Their study helps better understand the interactions between dopants and defects.

9.7.1 Electric Field Enhancement

During diffusion at elevated temperatures, the diffusing species are usually ionized. As a result, a gradient of charge (or more formally, a gradient in the Fermi level) will develop in the crystal. The gradient of the Fermi level will result in a *built-in electric field*. The electric field always operates in the direction to enhance the diffusion of the ionized impurities.

For the case of diffusion of a dopant-defect (AX) pair in the presence of an electric field, the total diffusion flux, J, will have both a classical diffusion component and an electric-field drift component. The flux of a dopant-defect pair (AX) in the presence of an electric field is then given by:

$$J_{AX} = -D_{AX}\frac{\partial C}{\partial x} + Z_{AX}\mu_{AX}C_{AX}E \qquad (9.43)$$

where Z_{AX} defines the charge state of AX dopant/defect pair and μ_{AX} is the mobility of AX. Assuming Boltzmann statistics pertain and that the Einstein relation between the mobility and the diffusivity also applies, the calculation of the enhancement of the diffusivity due to an electric field is rather straightforward, but lengthy. The reader is referred to the original reference for the complete solution.[31]

The enhanced diffusivity in an electric field is given by:

$$D_A = h\left[D_{A^+X^o} + D_{A^+X^-}\left(\frac{n}{n_i}\right) + D_{A^+X^{-2}}\left(\frac{n}{n_i}\right)^2\right] \qquad (9.44)$$

where h is defined as:

$$h = 1 + \frac{C_A}{2n_i}\left[\left(\frac{C_A}{2n_i}\right)^2 + 1\right]^{-1/2} \qquad (9.45)$$

The value of h varies from 1 for the case of $n << n_i$, and reaches 2 for the case of $n >> n_i$. The field enhancement may not only result from the presence of the diffusing dopant but also from the presence of another dopant in the crystal from a previous diffusion step.

9.7.2 Emitter Push Effect

Since the performance of an *npn* bipolar transistor depends precisely upon the width of the base region, the diffusion of the emitter must be accurately controlled. In the early days of bipolar device fabrication it was found that the measured base width, after the emitter drive, was larger than predicted from the diffusion coefficients of both the phosphorous and the boron. During the emitter drive, the boron (under the phosphorus) diffused deeper into the silicon than expected. This effect became known as the *emitter push* (or *emitter dip* effect). The emitter push effect is depicted schematically in Fig. 9-11.

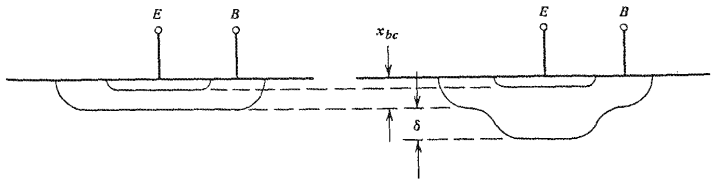


Fig. 9-11 Enhanced base diffusion under the emitter resulting from a heavy phosphorus diffusion. (Emitter push [or dip] effect).

Fair proposed that the enhanced boron diffusion under the emitter results from a combination of effects, related to the creation and dissociation of PV pairs.[32] The dissociation of the pairs near the boron increases the vacancy concentration and thereby the boron diffusion. It is now known that interstitials play a key role in both boron and phosphorus diffusion. Other experiments have shown that there is indeed a net interstitial supersaturation below the phosphorus diffused layer and these excess interstitials enhance the boron diffusion.[33,34,35,36]

Fortunately the importance of the phosphorus emitter push effect has diminished, since arsenic has replaced phosphorus as the emitter dopant. This change was made because arsenic allows shallower and better-controlled junctions to be formed. The use of arsenic, however, is not without its problems, since arsenic tends to form clusters and also tends to interact with the boron. In some cases arsenic was found to retard the diffusion of the boron (emitter pull).

9.7.3 Lateral Diffusion Under Oxide Windows

Since the majority of diffusion processes are carried out in masked areas one must be concerned not only with vertical diffusion, but also with lateral diffusion under the mask. This is not truly an anomalous effect since a concentration gradient exists in the lateral direction, as well as in the vertical direction. Thus, diffusion should proceed in both directions (two-dimensions).

One of the key processing issues related to lateral diffusion occurs when boron is diffused into the source-drain regions of a CMOS device, and the boron diffusion proceeds laterally under the edge of the gate electrode. There are two harmful effects that arise from lateral diffusion under the gate edge. First, the gate/drain overlap capacitance increases, and this can cause a decrease in the circuit performance. Second, the length of the active channel region (the distance between the source and drain) will decrease. If this decrease is excessive, it is possible to reach what is termed a "punchthrough" condition. When punchthrough occurs either the leakage current of the device substantially increases or the device will cease to function. Figure 9-12 shows the effect of boron lateral diffusion in the source/drain region of an MOS device.

For intrinsic diffusion under both constant-source (i.e., pre-deposition) conditions and limited-source (i.e., drive-in) conditions, the lateral penetration is found to be about 75–85% of the vertical diffusion depth (Figs. 9-13a and 9-13b).[37] At high-concentration (i.e., extrinsic) diffusion, as well as for shallow diffusions, lateral penetration is found to extend only 65–70% of the vertical diffusion distance.[38] There is an additional anomalous effect caused by the stress generated at the edge of an oxide or nitride window. Both enhanced and retarded diffusions

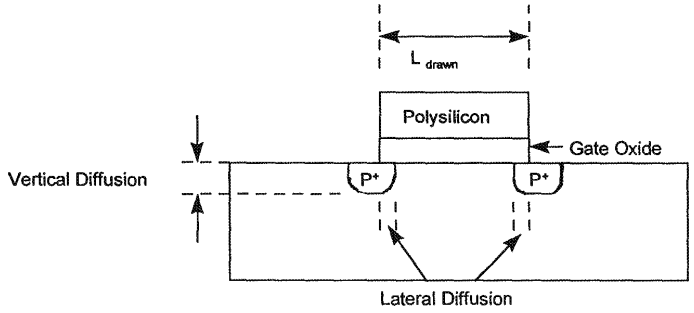


Fig. 9-12 The effect of lateral diffusion in an MOS device. The channel length is reduced to less than the polysilicon line size by the amount of lateral diffusion.

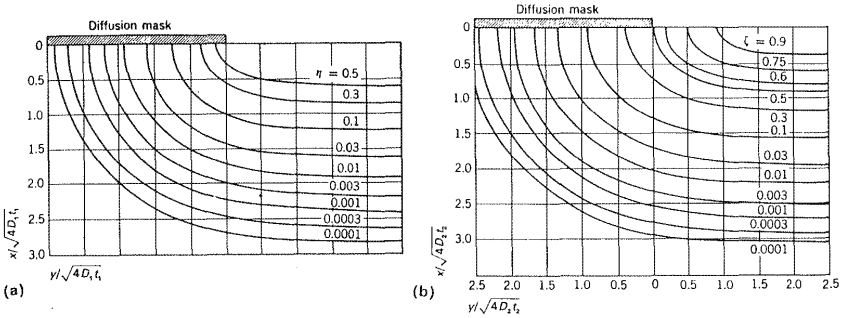


Fig. 9-13 Diffusion contours at the edge of an oxide window: a) constant source diffusion; b) limited source diffusion.[37]

sions have been observed as a result of elastic strain near the window edge.[39] It is known that strain can induce defects which affect the diffusivity of the dopants in silicon.

It is extremely difficult to directly measure the lateral diffusion in shallow junctions and the electrical performance of a MOS device may be used as an indication of lateral diffusion. TEM has been employed for this task, but it is difficult to discern the precise junction location. Some new techniques are beginning to be used for such measurements and they will be discussed in a later section.

9.7.4 Oxidation-Enhanced Diffusion (OED)

It has been known for some time that when diffusion takes place under *oxidizing* conditions the motion of certain dopants is greatly enhanced, as compared to the same diffusion in a neutral ambient. To be more precise, the diffusion of both boron and phosphorus are enhanced during oxidation. While arsenic diffusion is only slightly enhanced under oxidizing conditions, the diffusion of antimony (Sb) is retarded under the same oxidation conditions. The general name for this phenomenon is *oxidation-enhanced diffusion* (OED). There also exists a subsidiary effect called *nitridation retarded diffusion* (NRD), which will be discussed subsequently.

It is important to examine OED from two different perspectives. First of all, the amount of OED that occurs during a drive-in step must be known to accurately model diffusion. From a practical point of view one needs to take OED into account when defining a diffusion schedule. Failure to do so could lead to a much deeper junction than anticipated. It is also important to consider the possibility that different regions on the device may exhibit differing amounts of OED, since the amount of OED depends on the oxidation rate and different parts of the device may have different oxidation rates.

The second perspective is that the study of OED provides a window on how silicon vacancies and interstitials interact with the common dopants during diffusion. There have been a myriad of investigations of OED in silicon, and the treatise by Fahey, Griffin, and Plummer provides an excellent critical review of the OED literature through 1989.[9] Some of their conclusions are briefly described here, but readers interested in the details of the analysis should refer to the Fahey, *et al.*, paper.

As far back as 1974, S.M. Hu recognized that the growth of oxidation-induced stacking faults, OISF's (see Chap. 2) and OED had a common origin.[40] Direct observation of stacking faults in the TEM revealed that they consisted of silicon interstitial type defects, bounded by Frank dislocations. It was also known that the stacking faults grew during oxidation and the

growth was controlled by silicon atoms attaching themselves to the ends of the dislocation. The silicon atoms were injected into the bulk of the silicon from the surface during the oxidation process. Armed with the fact that interstitials enter the silicon during oxidation it was postulated that these same interstitials also affect OED.

Based on various OED experiments over the past 20 years a convincing and repeatable correlation between defects and dopants emerged. The conclusions are:

1. Since the diffusion of boron and phosphorus are enhanced by oxidation, the dominant (>50%) mechanism controlling their diffusion is the presence of self-interstitials,
2. Since the diffusion of antimony is retarded during oxidation the dominant mechanism controlling its diffusion is the presence of vacancies,
3. Since the diffusion of arsenic is not greatly enhanced or retarded during oxidation, there is no dominant mechanism, and both vacancies and interstitials play an important role.

The literature is replete with studies on how OED is affected by temperature and time. To perform these studies it is necessary to measure and quantify OED. The measurement technique used is shown in Fig. 9-14. Here the measurement is performed by comparing the diffusion of a dopant during oxidation with the same dopant when the surface is not being oxidized (at the same time and in the same sample). This is accomplished by masking part of the wafer with silicon nitride. It is well known that oxide will not grow under the nitride. The OED is then determined from the difference in the junction depth between the masked and unmasked regions.

The results of these studies are briefly summarized here. The major conclusion is that OED depends on the oxidation rate and the dependence follows the power law:

$$\text{OED} \propto (dx_{ox}/dt)^n \tag{9.46}$$

where (dx_{ox}/dt) is the oxidation rate and n is an experimentally determined fitting parameter, with a value less than 1, and typically in the range of 0.2–0.3.

When the oxidation rate is increased (for instance, during wet oxidation or during high temperature growth), it is expected that the relative amount of OED should decrease. The

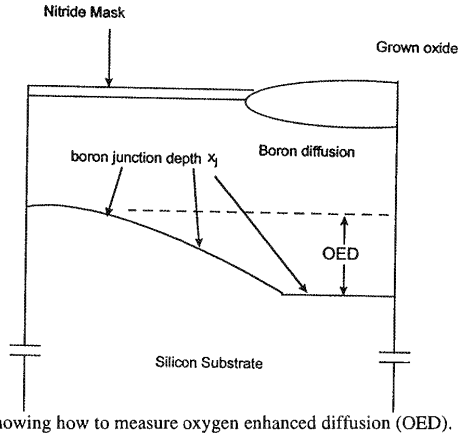


Fig. 9-14 A schematic showing how to measure oxygen enhanced diffusion (OED).

relative reduction of the OED with higher temperatures has been confirmed. There were also studies performed on the effect of silicon crystal orientation dependence on OED.[41,42] In these studies it was found that the relative amount of OED decreased in the order (111), (110), and (100) during dry oxidation, in line with the expected oxidation rates. It was also found that for oxidation temperatures in excess of 1150°C, the diffusion of boron was actually retarded, rather than enhanced, implying the injection of vacancies at this high temperature. At the same time antimony exhibited enhanced diffusion, confirming the presence of excess vacancies.

There have also been a series of experiments that studied OISF shrinkage and diffusion during the direct nitridation of silicon by ammonia (NH_3). It was found that during nitridation, OISF actually shrink in size. This shrinkage could result from a vacancy reaching the OISF and changing places with a silicon atom, thus making the stacking fault smaller. It is believed the vacancies are injected from the surface of the silicon during the nitridation. The mechanism for the creation of the vacancies is not yet well understood. One idea is that the stress-relief mechanism prevalent during the high-temperature nitridation is the formation of Frenkel pairs. The silicon interstitials from the Frenkel pairs are consumed by the nitride, while the vacancies enter the silicon where they react with the interstitials from the OISF.

If one considers the results of diffusion studies on the common dopants during *nitridation* one discovers that the diffusion of boron and phosphorus are retarded, while that of antimony is enhanced. This effect is parallel, but opposite, to that observed during oxidation. Since the diffusion of boron and phosphorus is dominated by interstitial defects, their undersaturation in the bulk will result in a reduced diffusivity. On the other hand, since the diffusion of antimony is dominated by vacancies it is not surprising to find enhanced diffusivity along with supersaturation of vacancies. The study of both OED and NRD provide a set of self consistent results which show a clear linkage between the diffusion of the dopants and the presence of defects in the silicon.

9.7.5 Transient Enhanced Diffusion (TED)

After an ion implantation step it is necessary to thermally treat (anneal) the silicon to activate the implanted ions (i.e., to move the dopant atoms to substitutional sites), and to remove residual implantation damage. During this anneal a new type of anomalous diffusion is observed, called *transient enhanced diffusion* or *TED*. The study of TED has become an area of fertile research, and by understanding TED the knowledge of the interaction between dopant atoms and defects has also been increased.

Before the discussion of the models that have been posited to explain TED is undertaken, some of the features that are common to this phenomenon will be described. First, TED is associated with the ion implantation damage and does not occur if the dopant is introduced by chemical means. Second, the diffusion enhancement is transient. That is, it exists for a finite amount of time after which "normal" diffusion returns. Finally, TED takes place in the temperature range where little or no atomic movement is expected (670–900°C). As a matter of fact, as much as 100 nm of boron dopant movement has been measured in this low temperature range.

These observations will now be discussed in terms of the proposed mechanism for TED. Briefly, TED is explained by assuming that silicon interstitials are injected from the implant damaged region into the bulk of the material. Furthermore, this interstitial injection occurs at relatively low temperatures for some finite time. The interstitials, once in the bulk, enhance the diffusion of boron and phosphorus. The sources of these interstitials are {311} defects, which

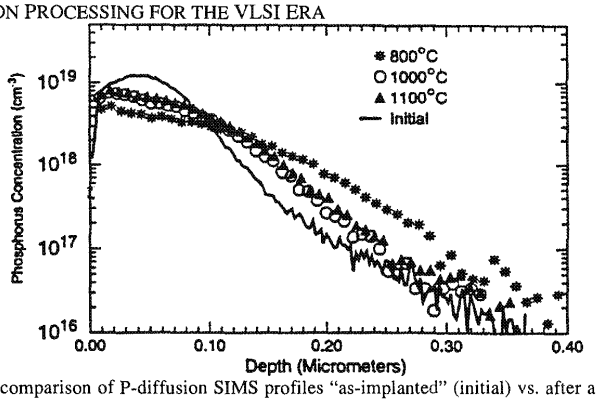


Fig. 9-15 A comparison of P-diffusion SIMS profiles "as-implanted" (initial) vs. after a furnace anneal at 800, 1000, and 1100°C. There is more diffusion occurring at 800°C than at 1100°C as a result of TED.[73]

coalesce after ion implantation damage. During the anneal cycle, these {311} defects begin to dissolve, and during their dissolution they emit interstitials into the silicon.

Figure 9-15 shows a SIMS plot of the phosphorus concentration as a function of distance in both the "as-damage-implanted" case and after anneal at several temperatures. The astonishing result is that diffusion proceeds farther at 800°C than at 1100°C. The fact that so much atomic movement can occur at such low temperatures implies that TED can take place during the loading and unloading of the wafers in a conventional furnace. Before a furnace annealing process even starts, the atoms may have moved too far to make useful junctions. This is one of the reasons why RTA has rapidly replaced furnaces for annealing the implanted atoms for shallow junctions. Fig. 9-16 displays a plot of the enhanced diffusion of phosphorus as a function of time at temperature for two different levels of implant damage. The conclusions that can be drawn from this curve are: a) the enhancement decreases rapidly with time, and b) for short times the enhancement is independent of the amount of damage. A phenomenological equation can be written for the time dependence of the enhanced diffusivity as:[43,44]

$$D_E = D_i + D_{0TED} \exp(-t/\tau) \qquad (9.47)$$

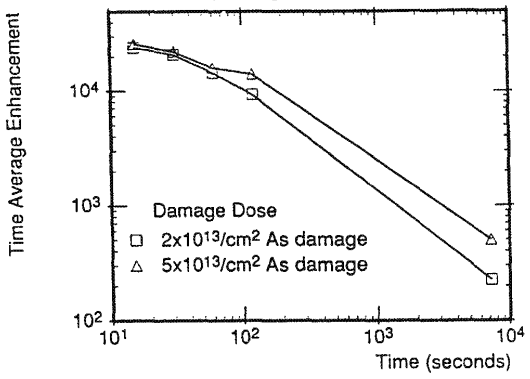


Fig. 9-16 Amount of enhanced diffusion as a function of time at temperature for two different levels of implantation damage.[74] Reprinted with permission of the American Physical Society.

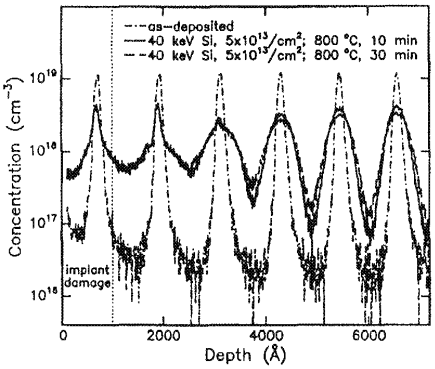


Fig. 9-17 Diffusion of boron (δ-doped using MBE) at 815°C for 10 and 30 min, showing enhanced broadening due to TED arising from a Si-damage implant.[45]

where D_E is the enhanced diffusivity; D_i is the intrinsic diffusivity; D_{0TED} is the enhancement factor at the onset of TED; t is the time into the TED event; and τ is the time constant for the enhanced diffusion. For short times and low temperatures, the total diffusion is dominated by the D_{0TED} term.

It is difficult to measure TED directly from the SIMS profile in the implanted silicon since the measurement is confounded by too many variables, including: a) the TED itself; b) the damage profile; c) the stationary component of the profile; and d) the variation of the interstitial profile in the implant region. A novel technique has been developed to directly measure TED. The technique uses CVD or MBE epitaxially-doped "superlattice" regions on float zone (FZ) silicon. These samples consist of regularly spaced spikes of dopant which are used as "marker" layers to monitor the atom movement during the TED event. A damage implant (either silicon or germanium) is then made into the sample silicon to create the required defects. The distance of the damage layer has been well characterized.

Figure 9-17 shows a SIMS profile of the as-deposited and diffused profiles of these spikes (boron doped in this case), as a function of depth after the introduction of ion implantation damage and anneal. The damage was introduced by a silicon implantation and the furnace anneal was performed at 815°C for 10 and 30 min.[45] Also shown on that curve is the depth (about 100 nm) of the implant damage. The boron profile before diffusion shows uniform sharp peaks as a result of the doping spikes that were grown-in during the low-temperature MBE. The results after the damage implant and anneal show that some of the boron spikes significantly broaden as a result of the thermal treatment. The amount of broadening is a direct measure of the diffusivity of the boron. From Fig. 9-17 one sees that the boron markers located within and adjacent to the implant damage remain immobile during annealing, while those farther away show broadening (indicating diffusion). The amount of broadening increases, and then decreases, as it moves farther away from the implant damage region. The broadening that is observed at the 350-nm marker corresponds to a 1000 times increase over the normal diffusivity of boron. The fact that no additional broadening takes place after 10 min testifies to the fact that the TED is a transient phenomenon. The duration of the transient, on the other hand, does strongly depend on the annealing temperature and is found to decrease from over 5 hours at 670°C to less than 10 minutes at 815°C. These times are related to the constant τ in Eq. 9-47.

The observation that the concentration peaks of the boron in or near the implanted region remain stationary, suggests that near the implant damage the boron atoms form immobile clusters (BI_n where n = 1, 2, or 3) with the excess silicon interstitials. The results of the marker experiments (no broadening) along with the fact that the electrical concentration of B is generally much less than the chemical concentration in that region, provides additional evidence of B-I cluster formation.

The source of the interstitials and their time dependence is now discussed. Eaglesham *et al.*, performed high resolution TEM studies of silicon in the implanted region after very short time anneals in order to observe the defect structure.[45] The appearance of an extended defect is quite obvious, and the defect has been identified as the *rodlike* or *{311} defect.* These defects are known to be present during oxidation and ion implantation of silicon. It is generally agreed that the {311} defect consists of a condensation of interstitials which form five- and seven-member rings. Another way of looking at these defects is that they consist of an extra monolayer of hexagonal silicon in the lattice that has no dangling bonds. As a result, these defects are relatively stable, except at high temperatures (> 950°C).

Using TEM, Eaglesham *et al.*, were able to measure the time dependence of the density and size of the {311} defects during short-time anneals and then to calculate the density of silicon interstitials.[19] Figure 9-18 shows the time dependence of the silicon interstitial density (the so-called *evaporation rate*) for several anneal temperatures. The evaporation rate was found to have an exponential dependence on time, with time constants that correspond very well to the duration of the TED as described in Eq. 9-47. Recall that the duration of TED decreased as the temperature increased. This correspondence strongly suggests that TED results from the growth and dissolution of {311} defects that are formed as a result of ion implantation. During the dissolution there are a large number of silicon interstitials formed. They greatly enhance the diffusion rate of boron, (and phosphorus) through the "kick-out" or interstitialcy mechanism.

It has been found that the presence of substitutional carbon in silicon in the range of $5 \times 10^{18}/cm^3$ effectively suppresses TED without affecting the electrical activity of the boron.[46] It is believed that the carbon atoms absorb the injected interstitials and therefore the interstitials are not available to contribute to TED.

Although TED has been studied most extensively with boron diffusion, both phosphorus[47] and antimony[48] also demonstrate the effect. In order to get TED to occur in antimony it is

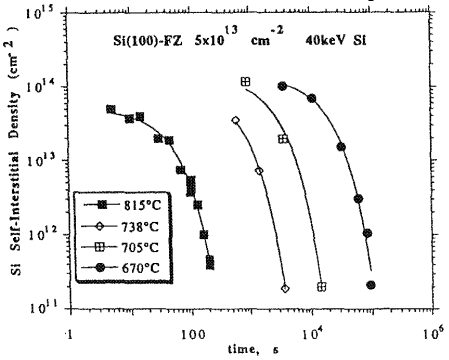


Fig. 9-18 Evaporation of interstitials from {311} defects as a function of time at several temperatures.[19]

necessary to achieve vacancy supersaturation in the region of the diffusion. This can be accomplished using MeV silicon implantations.

9.8 DIFFUSION SYSTEMS AND DIFFUSION SOURCES

Regardless of the predeposition process (chemical or ion implantation) a drive-in (or redistribution) step is usually required to move the dopants in the silicon (i.e., to attain electrical activation, and reduce the implantation damage). This drive-in step is performed either in a diffusion furnace or a rapid thermal annealing (RTA) system. (These technologies are both discussed in detail in Chap. 8.) However, RTA allows dopant activation to occur using a smaller thermal exposure than can be obtained with a furnace. As a result, the use of RTA for annealing ion-implanted shallow source/drain junctions in CMOS device processing has become more popular. There are some advanced device flows that use only RTA for drive-in diffusions.

Not all fabrication processes, however, use ion implantation for dopant introduction. Some still use chemical procedures. For example, backside phosphorus gettering (Chap. 2) uses a $POCl_3$ source and a diffusion furnace, while bipolar base diffusions may be predeposited in diffusion furnaces sourced with boron nitride (BN) discs. Nevertheless, ion implantation is the standard technology for introducing dopants into Si wafers, and this topic is covered in Chap. 10. New technologies are still being developed for creating shallow junctions in ULSI devices, and a discussion of these follows. Older technologies for chemical diffusion of dopants into Si are covered in the 1st edition of this text (© 1986); interested readers are referred there for this information.

9.8.1 Advanced Diffusion Technologies

Advanced technologies have been studied for introducing dopants into a silicon device. These new techniques include: a) *plasma immersion ion implantation doping* (PII); b) *rapid vapor phase doping* (RVD); and c) *gas immersion laser doping* (GILD). In this section, RVD and GILD will be briefly discussed. The subject of PII is covered in Chap. 10.

9.8.1.1 Rapid Vapor Doping (RVD): Rapid vapor doping (RVD) is emerging as an alternative to ion implantation doping for some advanced applications, such as capacitor doping for DRAMs, side-wall doping in deep trenches, and even for shallow source/drain junctions in CMOS applications. Unlike vapor doping in a furnace, RVD does not depend on the formation of a doped glass layer on the surface of the wafer. For RVD, the doping proceeds directly from the gas phase into the solid. For that to occur the dopant must first be adsorbed on the silicon surface, followed by solid-state diffusion. The potential advantages of the RVD technique over ion implantation (particularly for shallow-junction applications) are that RVD does not suffer from the effects bedeviling implantation, namely: channeling, lattice damage, or wafer charging.

Rapid vapor doping uses rapid thermal processing (RTP) to quickly and uniformly heat the wafer to the desired temperature while the wafer is exposed to a gaseous dopant source. The precise control of the gas flow and dilution are necessary to ensure a uniform dopant concentration across the wafer surface. The final dopant surface concentration (C_s) and junction depth (x_j) depend on the gas phase concentration of dopant and the duration and temperature of the RTP treatment. The gas phase concentration is controlled by diluting the source gas (e.g., phosphine) with a hydrogen carrier gas. There are some disadvantages of RVD, such as: a) a relatively high thermal budget is required for the surface reaction and solid-state diffusion to get the dopants into the silicon; and b) "hardmasks" are required for selective doping.

There have been reports describing the RVD fabrication of shallow boron junctions from vapor sources. Kiyota *et al.*, reported forming ultra-shallow boron junctions by flowing B_2H_6 in H_2 at 900°C for 30 sec.[49] This process resulted in a junction depth of 30 nm and a surface concentration of 5.8×10^{19} cm^{-3}. This level of doping is adequate for fabricating base regions in bipolar devices or the source/drain extensions in CMOS devices. Short-channel CMOS devices produced using RVD showed improved performance over devices produced by conventional ion implantation.[50] Similar studies using phosphine (PH_3) diluted in hydrogen at 900°C reported junction depths <100 nm and surface concentrations between 3.0×10^{18} and 1.3×10^{19}/cm^3.[51]

9.8.1.2 Gas Immersion Laser Doping (GILD): The origins of GILD technology extend back more than 15 years, when lasers were first used to locally melt silicon immersed in a doping gas (such as PF_5 or BF_3).[52,53] The technology has been refined since that time, and it is now possible to fabricate devices using GILD technology. The basis of GILD technology is to shine high energy density (0.5–2.0 J/cm^2) *eximer laser* light (308 nm) for a short pulse time (20–100 ns) onto a silicon wafer immersed in a dopant gas phase. The energy is absorbed into the first 10 nm of the silicon surface, where it dissipates by melting the surface layer. The thin melted layer absorbs and dissolves the dopant atoms directly from the gas phase. Since the diffusivity of the dopants in the liquid phase is about 8 orders of magnitude faster than in the solid phase, the dopants rapidly and uniformly diffuse throughout the molten layer. When the laser power is turned off, the thin molten layer epitaxially recrystallizes (freezes). During recrystallization the dopant is incorporated into the silicon lattice on substitutional sites. Consequently, the majority of the dopant is activated immediately after cool down and no further anneal is required. The liquid-to-solid transformation occurs at an extremely high rate (> 3m/s) and it is possible for the amount of active dopant incorporated in the silicon to exceed the equilibrium concentration (supersaturation). As long as the wafer is not subjected to subsequent elevated temperature processing, the excess dopant will stay in solution.

Since the time that the wafer is exposed to the high-energy laser beam is short, the bulk of the wafer remains cool and little or no solid-state diffusion occurs. As a result the dopant is confined only to the molten region and any dopants that were previously introduced into the silicon remain unperturbed. The depth of the junction is controlled by the depth of the molten silicon, which depends on the energy density and pulse time of the laser treatment.

Since the concentration of the dopant is uniform throughout the "diffused" region the concentration profile tends to be more box-like rather than erfc or Gaussian. Consequently, the profile gives an abrupt junction. Figure 9-19 shows the diffusion profile after a GILD of phosphorus in a *p*-type substrate. The box-type profile is apparent. A 30 ns pulse results in a junction depth of 30 nm with a surface concentration 4×10^{20}/cm^3. As the pulse time increases the junction depth gets deeper.

A major innovation towards making GILD a production-worthy technology was the introduction of *projection-GILD or P-GILD*. This technology utilizes a step-and-repeat camera, similar to those described for projection lithography in Chap. 13. Eximer laser light is shined through a dielectric mask. The light from the mask is demagnified (e.g., a 4x reduction), focused, and projected onto the wafer. The wafer is stepped after each field of the mask is exposed (or melted, in this case). The amount of doping incorporated into the layer increases as the number of laser exposures increase, while the depth of the junction increases with the increase in the energy of each pulse. The dopant is incorporated only in the regions that are melted (i.e., only beneath the regions of the mask that transmit the laser light).

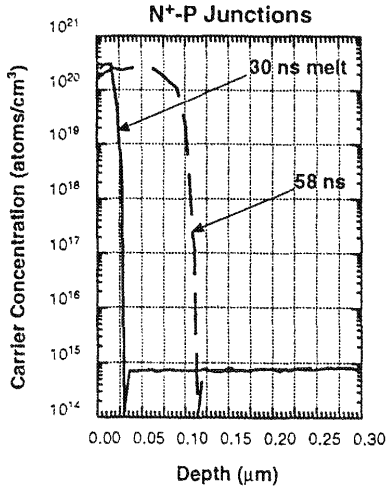


Fig. 9-19 Spreading resistance profile (SRP) for a phosphorus doped P-GILD formed in a *p*-type doped substrate. Note that x_j increases with surface melt time and the diffusion profile forms a "box" type profile with no anomalous tail.[53]

9.9 MEASUREMENT TECHNIQUES FOR DIFFUSED (AND ION-IMPLANTED) LAYERS

The diffusion coefficient D is experimentally determined from the concentration profile in the silicon subsequent to chemical predeposition (or ion implantation) and drive-in. In order to determine accurate values of D, reliable measurements of the junction depth, sheet resistance, and diffusion profile must be feasible. Sheet-resistance measurements have become the standard method for "in-line" monitoring of diffusion processes in IC production facilities. These measurements are performed immediately after a diffusion step. In this section some of the important measurement techniques for determining sheet resistance and doping profiles are described.

9.9.1 Sheet Resistance Measurements

The *sheet resistance*, R_s of a diffused layer is the resistance exhibited in a square (i.e., a region of equal length and width) of that layer, which has a thickness x_j (junction depth). The value of R_s is expressed in units of ohms/sq (Ω/sq), and is related to the *average resistivity* ρ of a diffused layer by:

$$R_s\ (\Omega/\text{sq}) = \rho\ (\text{ohm-cm})/x_j\ (\text{cm}) \tag{9.48}$$

The layer (or film) resistivity is a material property that depends on the number of free carriers and the mobility of these carriers. The average resistivity ρ is found from:

$$\rho = \left[q \int_0^{x_j} \mu(C_A) C_A(x)\,dx\right]^{-1} \tag{9.49}$$

where q is the charge on a carrier $\mu\ (C_A)$ is the carrier mobility (which is a function of the carrier concentration), and $C_A(x)$ is the diffusion profile.

The value of R_S is readily obtained by measuring the resistance of the diffused layer using the *4-point probe technique*.[54] Figure 9-20 shows the co-linear probe arrangement that is the basis for the ASTM standard,[55] which is the standard for measuring R_S. To make this measurement a current, I, is forced between the outer two probes, and the voltage drop, V, is measured between the two inner probes. When all of the probes are equally spaced, the resistivity is given by:

$$R_S = 2\pi s F(V/I) \tag{9.50}$$

where s is the probe spacing, and F is a correction factor (that arises from the fact that the sample is not semi-infinite, but has a finite thickness and diameter).[64] When the sample thickness t is small, relative to the probe spacing (as in the case of a diffusion, where $t = x_j$), then Eq. 9-50 can be rewritten as:

$$\rho = 4.532t\left(\frac{V}{I}\right) \quad \text{or} \quad R_s = \frac{\rho}{t} = 4.532\left(\frac{V}{I}\right) \tag{9.51}$$

There is an additional set of correction factors related to the finite diameter of the wafer and the effect of the probes being close to the edge of the wafer. The correction factor is 1 if the probes are at least 3x their spacing from the wafer edge. When the probes are closer to the edge, the correction factor decreases to a value of 0.5–0.7, depending on the orientation of the probes to the wafer edge. The factors are listed in Ref. 64. For a standard probe spacing of 0.0625 inches, the probes need at least 4.7 mm from the wafer edge in order not to require a correction factor. To prevent erroneous readings the probe spacings must be accurately known, and if they are not equal, that must also be taken into account in the correction factor.

The sheet resistance measurement is first performed with the current in the forward direction, and then in the reverse direction (in order to minimize errors due to thermoelectric heating and cooling effects). The two (V/I) readings are averaged. The measurements should also be made at several current levels, until the proper level is found. That is, if the current is too low, the values of the forward and reverse readings will differ, while if it is too high, I^2R heating will cause the measured reading to drift with time. The ASTM standard also recommends best current levels, which depend on the resistivity range of the diffused layer. The amount of pressure applied to the probes during measurement is also important, since excessive pressure can drive the probes past the junction, while too little pressure results in poor electrical contact to the sample.

Automated 4-point probe equipment is currently available from several vendors. The equipment is capable of measuring the sheet resistance in numerous places on the wafer. The

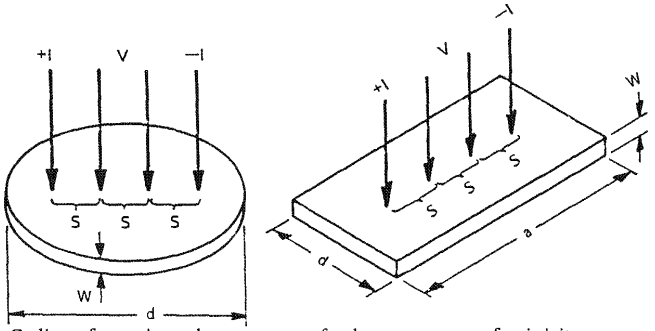


Fig. 9-20 Co-linear four-point probe arrangement for the measurement of resistivity.

system has a built-in algorithm to calculate the sheet resistance at these points and to use the appropriate correction factor for each position. The data can be presented in a graphical format using a *contour map* of resistivity. In this way the process engineer can look for process variations that are wafer dependent. The program can also calculate and depict statistical information about the sheet resistance variation. Such tools are used for in-line process control in many wafer fabs.

It is possible to estimate the doping profile of a diffusion using the *differential sheet resistivity* technique. In this technique, the sheet resistivity is measured initially and then again after a small amount of the silicon surface is removed. Anodic oxidation, followed by an etch of the anodic film, is used to control the amount of silicon removed in each step. The sheet resistivity will increase as dopant is removed, since the amount of doping is decreasing and the thickness of the diffused layer is decreasing with each step. The rate of the increase in R_S depends on the resistivity profile shape. The resistivity depends on the dopant concentration and the carrier mobility. If the mobility is known for each value of resistivity, the carrier concentration profile can be obtained with this technique. Note that the differential-sheet-resistivity method is tedious, and is only used in special cases where other profiling techniques are not available.

9.9.2 Capacitance-Voltage (C-V) Measurements

The carrier concentration profile can be measured directly within a diffused region by using the C-V technique. The use of C-V measurements to determine the diffusion profile is predicated on the measurement of the reverse bias capacitance of an n^+/p or p^+/n junction as a function of voltage. The capacitance of a junction is given by:

$$C(V) = \left[\frac{q\varepsilon_{Si} C_A}{2}\right]^{1/2}\left[V_{bi} \pm V_R - \frac{2kT}{q}\right]^{1/2} \tag{9.52}$$

where ε_{Si} is the permittivity of silicon, C_A is the substrate doping concentration, V_{bi} is the built-in potential of the junction, and V_R is the applied reverse bias voltage.

The technique requires the formation of a very shallow n^+ or p^+ diffusion over the surface of the diffused layer of interest. The doping in the heavily doped region must be more than 100 times that in the lightly doped region (where the profile is measured) for the technique to work. The capacitance is measured while applying a variable reverse dc voltage to the junction. The instantaneous value of C_A is then calculated from Eq. 9-52. The depth at which C_A is measured is obtained from the value of the applied reverse voltage, assuming the distance equals the depletion width of a one-sided junction. If those conditions hold, the depth of the measurement is given by:

$$x = \sqrt{\frac{2\varepsilon_{Si}(V_{bi} + V_R)}{qC_A}} \tag{9.53}$$

The major limitation of this technique is that the initial values of C_A that are measured are at least as deep as the value of x_j for the n^+ or p^+ diffusion plus the zero-bias depletion width (which extends into the substrate). The depth of the profiling is limited to the reverse breakdown voltage of the junction. Therefore, the C-V technique has little value for determining the doping profile in shallow junctions. The technique also falls short when abrupt profiles are measured, since fictitious tails are often observed near the steep edge.[56]

9.9.3 Spreading Resistance Profiling (SRP)

The use of *spreading resistance* SR (i.e., the resistance associated with the divergence of the current lines emanating from a small-tipped electrical probe that is placed onto the silicon), was first proposed by Mazur and Dickey for measuring the thickness of diffused or epitaxial layers, and establishing the impurity profile for these structures.[57] The technique is referred to as *spreading resistance profiling* (SRP). In this technique, the SR of a reproducibly formed point-contact is measured. This can be accomplished by using one, two, or three probe configurations. The two-probe method has met with the most success on commercially available equipment. SR measures only the electrically active dopant concentration, and other techniques, such as SIMS, are needed if the total (active + inactive) concentration profile is required.

To make spreading resistance measurements, a known current is applied between the probes, and the voltage drop is measured across these probes to obtain a spreading resistance (R_{SR} = V/I). Most of the voltage drop occurs at a distance that is only a few times the probe radius, and there-fore the value of R_{SR} is measured in a very small volume of silicon immediately under the probe. In the two probe method, the Si resistivity ρ is related to the R_{SR} by the relationship:

$$\rho = 2 R_{SR} \mathbf{a} \tag{9.54}$$

where **a** is an empirical quantity which is related to the effective electrical contact radius (it is not however the probe tip radius). The value of **a** is determined by measuring R_{SR} on a well-characterized material of known resistivity. An ASTM Standard has been developed for conducting SR measurements, and the reader is referred to this reference for details.[58]

In order to use spreading resistance the probe tips must be well conditioned. This is achieved by stepping the probes at least 500 times on a silicon substrate that has been ground with fine grit. The probe tips must also be calibrated to obtain the empirical value of **a.** This is best done on a wafer that has a well-documented resistivity (determined using a more direct technique). It is important that the surface of the calibration wafer and the surface of the unknown sample be identically prepared.

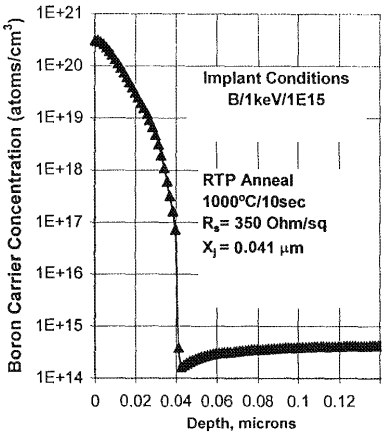


Fig. 9-21 Spreading resistance profile measurement showing the capability of measuring very shallow junctions. Courtesy of Applied Materials

The main use of SR is to determine doping profiles of diffused or epitaxial layers. This is accomplished by angle lapping the wafer and then making SR measurements along the length of the lapped surface. Knowing the angle of the taper gives the depth as a function of distance from the surface. The same precautions used to lap a wafer for junction depth measurements (see next section) must be obeyed when beveling a specimen for SR measurements, including the use of the laser measurement to determine the angle. Elaborate correction techniques and correction factors (CF) have been developed to convert the values of R_{SR} into carrier concentration values, particularly in multilayer samples. The values of the CFs depend on the nature of the layer and the proximity of the measurement to the junctions.

By beveling the sample to extremely low angles (<0.05 degrees), it is now possible to perform SRP on very shallow junctions. The skill required to accurately make these measurements is restricted to a few commercial laboratories.[59] Fig. 9-21 shows a SR profile for a shallow junction boron diffusion, with an x_j of only 30 nm. The most difficult part of the measurement is determining where the top surface ends and the bevel starts. Depositing a layer of polysilicon on the top surface can aid in finding where the bevel begins.

9.10 JUNCTION DEPTH MEASUREMENTS: PHYSICAL TECHNIQUES

9.10.1 Angle Lap and Stain

The *angle-lap and stain method* is effective for determining the depth of some *pn* junctions, and the test method is available in ASTM Standard Test Method F-110-84.[60] The method works well on relatively deep, highly-doped junctions, but is not accurate for shallow junctions. Wu *et al.* have reported that the accuracy of this technique can be significantly improved (to ± 20 nm of the true metallurgical junction), if the proper lapping and staining precautions are taken.[61]

The angle-lapping technique requires the grinding of a small angular bevel on the silicon wafer. This is accomplished by mounting the wafer on an angle block, and moving the assembly over a flat glass onto which a lapping-compound slurry has been poured. To obtain small angles ($\leq 1^{\circ}$) of good quality, the following practices should be followed:

1. The angle-lap assembly should be massive, to eliminate rocking during the lapping process. Small pieces of silicon affixed to the bottom of the holder, where they act as runners, are found to be helpful in minimizing the rocking, as well.
2. A slurry consisting of the lapping compound (alumina of ≤ 0.3 μm particle size, mixed with water) is used. The slurry should be changed often to keep it free from dirt and Si chips.
3. Any metallization on the wafer surface should be removed prior to lapping, since it smears when lapped, and will interfere with the subsequent staining process.
4. If the region of interest does not have a reasonably thick (>300 nm) protective oxide or nitride layer, edge rounding can occur. This can make it difficult to locate the position of the surface with certainty. Thus, it is good practice to chemically deposit ~500 nm of low temperature silicon dioxide onto such surfaces prior to lapping.

After lapping has been performed (which brings the junction to the surface), a staining solution is used to delineate either the *n*-type area or the *p*-type area. One such stain consists of a copper sulfate ($CuSO_4$) solution for staining the *n*-region. The staining occurs as a result of the following events: a) a drop of $CuSO_4$ is dispensed onto the junction; b) at the same time the junction is illuminated by an intense light source which causes it to become forward-biased;

Chapter 10

ION IMPLANTATION FOR ULSI

Ion implantation is a process in which energetic, charged atoms (or molecules) are directly introduced into a substrate. In ULSI fabrication, ion implantation is primarily used to add dopant ions (most often selectively) into the surface of silicon wafers. Acceleration energies for CMOS doping applications range between 0.2 keV (for shallow junctions such as source/drain extensions) to ≈2 MeV (for multiple-well doping). The superiority of ion implantation over chemical (diffusion) doping methods has enabled implantation to replace diffusion doping for almost all planar doped structures (i.e., one exception involves the doping of deep trench sidewalls, in which impurities are diffused into the Si from doped CVD SiO_2 films).

Ion implantation is superior to chemical diffusion for introducing dopants for the following reasons:

1. Ion species can be implanted with high accuracy over many orders of magnitude of doping levels (whereas dopant control by diffusion is far less exact).
2. Depth profiles of the implanted species can be established by controlling ion energy and channeling effects. This capability can be extended to produce the depth profiles needed in complex applications by using multiple-energy implants.
3. By using polymer-based photoresist patterns as the masking material, dopants can be introduced into selected regions of the wafer surface at near-room temperatures.
4. Both *p*- and *n*-type dopants (with a variety of activation and diffusion characteristics) can be implanted into Si.
5. The crystalline structure of implant-damaged Si can be recovered through subsequent thermal annealing cycles.

In this presentation, the physical mechanisms which determine implanted atom profiles will first be discussed. A comparison between theoretical models and experimental observations (i.e., to illustrate how closely actual implantation profiles match the profiles predicted by the models) will then be provided. Next, a description of ion implantation systems is undertaken (including safety aspects and operational limitations). This is followed by a presentation of the methods used to monitor and evaluate various aspects of ion implantation, including: a) implantation doses; b) implantation dose uniformity across a wafer; c) implantation profiles; d) implantation damage; and e) the degree to which damage is removed (and/or remains) after annealing. Finally a group of topics relating to practical aspects and specific applications of ion implantation processes for CMOS transistors is discussed. Table 10-1 lists some of the important applications of ion implantation in ULSI.

Note that this chapter was revised by Dr. Michael Current. 371

Table 10-1 ION IMPLANTATION APPLICATIONS IN ULSI CMOS

Junction Formation
(Used to fabricate both MOS and bipolar devices)

CMOS Fabrication (see Vol. 2 of this series)
- Threshold Voltage Control/Adjustment
- Channel-Stop Implantation
- Source/Drain Formation
- Well Formation
- Punchthrough Stopper Implantation
- Graded Source and Drain Formation

Bipolar Fabrication (see Vol. 2 of this series)
- Predeposition
- Base Formation Implantation
- Arsenic-Implanted Poly-Si Emitter
- High Value Resistor Formation Implantation

Formation of Silicon-on-Insulator Materials
- High Dose Oxygen Implantation (SIMOX - see Ch. 7)
- High Dose Hydrogen Implantation (Smart-Cut- see Ch. 7))

Miscellaneous Process Applications
- Backside Damage Layer Formation for Gettering (see Chap. 2)
- High-Energy Implantation to Form Buried Layers for Gettering
- Ion Beam Mixing to Promote Silicidation Reactions
- Buried Insulator Layer Formation (Chap. 7 and Vol. 2 of this series)
- Nitrogen Implantation into Poly-Si Gates to Reduce Boron Diffusion
- Polysilicon Resistor Formation
- High Energy Implantation to Form Buried Collectors
- Emitter Formation Implantation

It is useful to begin the discussion with a definition of *implantation dose*. The ion beam current in implanters ranges between about 1 μA and 30 mA, depending on the implant species, energy, and type of implanter. (Implanters used for oxygen implantation for the formation of SIMOX materials have beam currents in the range of 50–100 mA.) The number of implanted ions per unit area is termed the *dose*, ϕ. Typical doses range from 10^{11}–10^{16} atoms/cm^2. The dose (in atoms/cm^2) is related to beam current I (in amperes), beam area A (in cm^2), and implantation duration, t (in sec) by:

$$\phi = \frac{I\,t}{q_i\,A} \tag{10.1}$$

where q_i is the charge per ion (normally equal to one electronic charge = 1.6×10^{-19} coulomb).

10.1 ADVANTAGES (AND PROBLEMS) OF ION IMPLANTATION

10.1.1 Advantages

1. The *most important advantage* is the ability to more precisely control the number of implanted dopant atoms into substrates (e.g., with a repeatability and uniformity of better than ±1%). For dopant control in the concentration range of 10^{14}–10^{18} atoms/cm^3, ion implantation is clearly superior to conventional chemical deposition techniques.

2. Implanted impurities are introduced into the substrate with less *lateral* distribution than from diffusion doping processes. The reduced lateral impurity penetration allows devices to be fabricated with smaller feature sizes.

3. Mass separation by the ion implanter allows elemental selection for dopant beams, even if the source materials contain a mixture of elements.

4. A single ion implanter can be used for a variety of implant species. There is a risk of cross contamination by other species used in the implanter through sputtering and vapor transport, but newer generation implanters minimize this problem through innovative equipment designs.

5. Ion implantation can inject dopant atoms into a semiconductor by implantation through a thin surface layer (e.g., SiO_2). The surface layer can also serve as a protective screen against contamination by metals and other dopants during the introduction of the desired impurities.

6. The distribution of implanted ions in silicon can be modeled and predicted to a high degree of accuracy for most dopants and transistor materials. This is a tremendous advantage in new device and process development.

7. Complex-doping profiles can be produced by superimposing multiple implants having various ion energies and doses.

8. Under some implantation conditions either highly abrupt or graded junctions can be formed.

9. A variety of materials are suitable for creating masks to keep ions from being implanted into unwanted regions. The masking layer prevents ions from being introduced into regions under the mask (except near the mask edge). Lateral scattering effects do occur as a result of the ion implantation process, but are smaller than lateral diffusion distances.

10. Ion implantation is a near-room-temperature process, allowing materials such as polymer-based photoresist to routinely be used as implant-mask layers.

11. The inherent purity of processing under high vacuum reduces problems from a variety of contamination and defect causing mechanisms.

12. The parameters that control an ion implantation process are amenable to automatic control for beam setup, implantation, and cassette-to-cassette wafer motion.

10.1.2 Problems/Limitations of Ion Implantation

1. Ion implantation causes damage to the material structure of the target (i.e., the wafer). In crystalline targets (e.g., single-crystal Si wafers), crystal defects and even amorphous layers are formed. To restore the wafer to its pre-implantation condition, thermal processing after implantation must be performed. In some cases, significant implantation damage cannot be removed.

2. The maximum implantation depth achievable with conventional implanters is relatively shallow ($\approx 1\ \mu m$).

3. The lateral distribution of implanted species (although smaller than lateral diffusion effects), is not zero. This is a fundamental limiting factor in fabricating some minimum sized device structures, such as the electrical channel length between source and drain in self-aligned MOSFETs.

4. Throughput is typically lower than diffusion doping processes. High implantation doses may require undesirably long implantation times, depending on the implanted species and equipment design and operation.

5. Ion implanters are complex machines, among the most sophisticated systems in the wafer fab. In order to be effectively utilized they must be conscientiously operated, monitored, and maintained by well-trained personnel.

6. Ion implantation equipment contains many potential safety hazards to the personnel who operate or service the machines (e.g., high voltage, radiation) and toxic gases). To minimize the

likelihood of accidents from operating and especially maintaining such equipment, careful safety procedures must be established and strictly followed.

7. Some older-style implanters use diffusion pumps in the beam column, and this will lead to organic chemical contamination due to oil backstreaming. The effects of such contamination are described in Chaps. 3 and 5.

10.2 IMPURITY PROFILES OF IMPLANTED IONS

In order to benefit from the ability to control the number of impurities implanted into a substrate, it is necessary to know where the implanted atoms are located after implantation (i.e., it must be possible to predict the *depth distribution*, or *profile* of the as-implanted atoms). For example, this information is necessary for selecting appropriate doses and energies when designing a fabrication process sequence for new or modified integrated circuit devices. What is needed to make accurate predictions of implantation profiles is a theoretical model (or models) based on the energy interaction mechanisms between the impinging ions and the substrate. In this section the topic of how such theoretical models have been developed, and under what conditions they provide accurate predictions of implantation profiles, will be addressed. Figure 10-1 outlines the evolution of the models developed for determining implantation profiles and indicates the conditions under which they can be used to provide useful predictions.

Despite the fact that the derivation of the models is quite mathematical and complex, the scope of this text is limited to a more qualitative discussion. Even on a largely qualitative level such a presentation is valuable. That is, it provides the reader with an appreciation of the intellectual underpinning of ion implantation profile prediction, and also serves as an introduction to other physical mechanisms associated with ion implantation. These include channeling effects during implantation, substrate damage from implantation, and recoil effects that occur when implantations are done through thin layers present on the substrate surface. Readers interested in gaining a deeper and more quantitative understanding of implantation profile models can refer to references given at the end of the chapter. Some are comprehensive surveys,[1,2] while others are papers in which the models were originally published (and which thus discuss more fully the assumptions underlying the model, and details of the derivations and associated calculations).[3,4,5]

10.2.1 Definitions Associated with Ion Implantation Profiles

As energetic ions penetrate a solid target material they lose energy due to collisions with atomic nuclei and electrons in the target, and eventually come to rest. The total distance that an ion travels in the target before coming to rest is termed the *range*, R. As a result of the collisions between the ions and the target material nuclei, this trajectory is not a straight line (and in fact, the value of the total distance traveled is not even the quantity that is of highest interest). What is of greater interest than R is the projection of this range on the direction parallel with the incident beam, since this represents the penetration depth of the implanted ions along the implantation direction. This quantity is called the *projected range*, R_p (Fig. 10-2a). As the number of collisions and the energy lost per collision experienced by the penetrating ion are random variables, ions having the same initial energy and mass will end up spatially distributed in the target. An average ion will stop at a depth below the surface given by R_p. Some ions, however, undergo fewer scattering events than the average, and come to rest more deeply into the target. Others suffer more collisions, and come to rest closer to the surface than R_p. As a

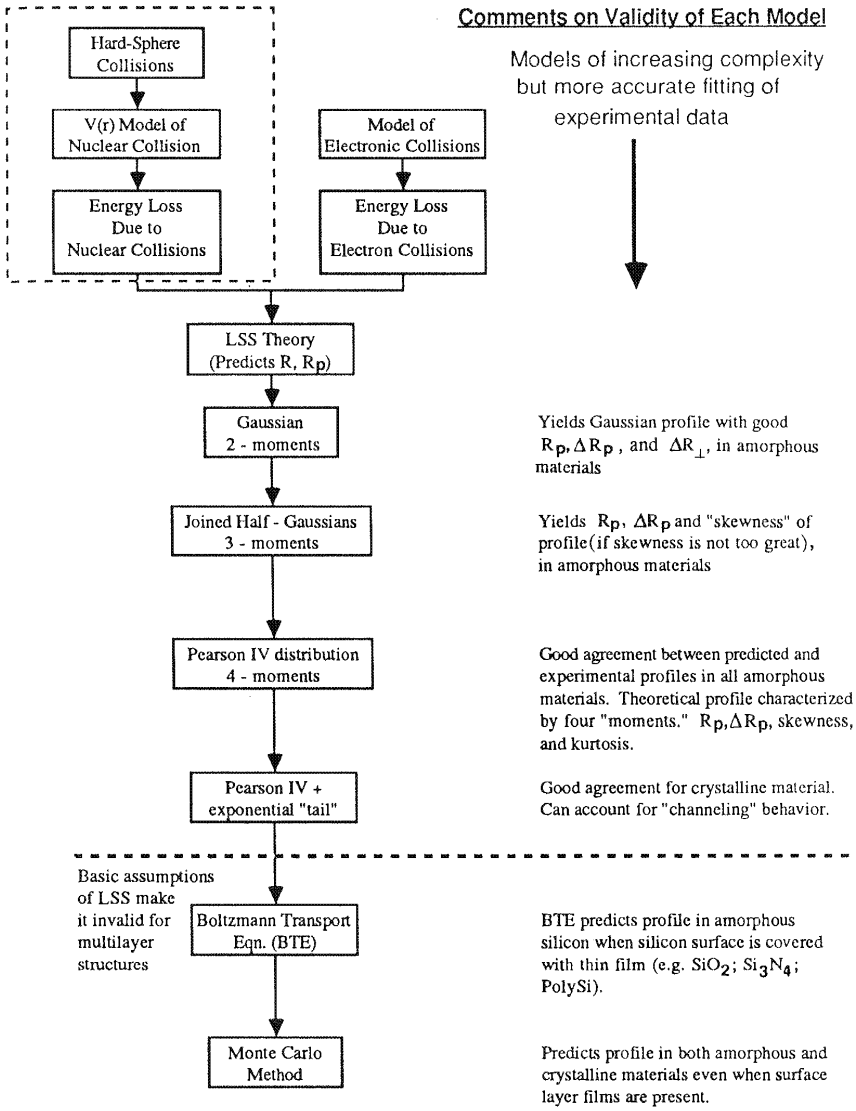


Fig. 10.1 Mathematical models for predicting ion implantation concentration profiles.

result, when large numbers of ions are implanted, R_p corresponds to the depth at which most ions stop, and this is the distance at which the profile has maximum value. The statistical fluctuation along the direction of the projected range is given by the quantity known as the *projected straggle*, ΔR_p (Fig. 10-2b).

The ions are also scattered to some degree along the direction perpendicular to the incident direction, and the statistical fluctuation along this direction is called the *projected lateral straggle*, $\Delta R_\perp$(Fig. 10-2b). Information about the projected lateral straggle is important because it describes the extent of lateral penetration of ions under the edges of implant masks. Such lateral penetration may represent a limiting dimension in some integrated circuit device structures. In general, values of ΔR_p and $\Delta R_\perp$ are within ±20% of one another. Figure 10-3a shows an opening in a "thick" mask material and resultant ion concentrations in the lateral direction in the target. It is seen that lateral straggle effects cause implanted ions to be distributed under edges of the mask. In Fig. 10-3b contours of equal-ion concentration for 70-keV B atoms implanted into a 1-μm slit are shown.[6]

10.2.2 Theory of Ion Stopping

As an ion moves through a solid target, it transfers energy by collisions with the target nuclei (*nuclear collisions*) and by coulombic interaction with the electrons in the target material. In the latter mechanism, the energy transferred to the electrons can lead to exciting the electrons to higher energy levels (*excitation*), or to the ejection of electrons from their atomic orbits (*ionization*). The energy loss due to such target interactions gradually slows the ion, eventually bringing it to a stop. If the energy of the ion at any point along its trajectory in the target is given by E, the process of *energy loss through nuclear collisions* can be characterized by an *energy loss per unit length due to nuclear stopping*, $S_n(E)$, and *energy loss from interactions with target electrons* by an *energy loss per unit length due to electronic stopping*, $S_e(E)$. The total rate of energy loss $(dE/dx)_{total}$ is given by the sum of these stopping mechanisms:

$$\left[\frac{dE}{dx}\right]_{total} = S_n(E) + S_e(E) \qquad (10.2)$$

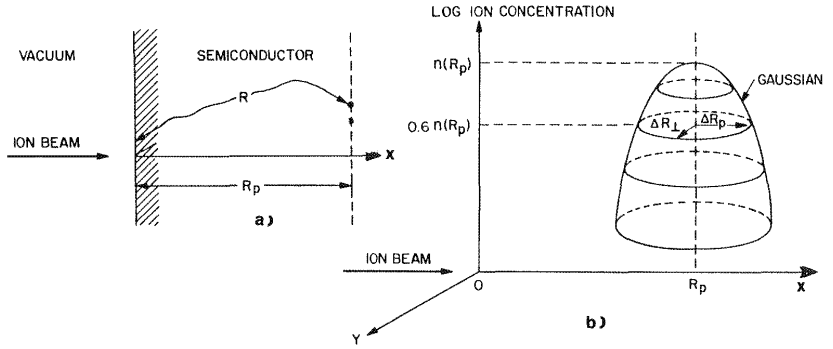


Fig. 10-2 a) Schematic of the ion range, R, and projected range, R_p. b) Two-dimensional distribution of the implanted atoms. From S.M. Sze, *Semiconductor Devices, Physics and Technology*, Copyright © 1985 John Wiley & Sons. Reprinted with permission of John Wiley & Sons.

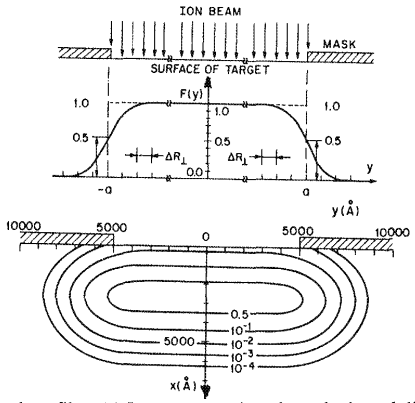


Fig. 10-3 Illustration of lateral profiles. (a) Ion concentration along the lateral direction (y) for a gate mask with a >> $\Delta R_\perp$ and infinite extension in the x-direction. (b) Contours of equal-ion concentrations for 70 keV B^+ (R_p = 2710 Å, ΔR_p = 824 Å, and $\Delta R_\perp$ = 1006 Å) incident into silicon through a 1 μm slit.[6] Reprinted with permission of the Japanese Journal of Applied Physics.

If the total distance that the ion travels before coming to a complete stop is given by R, then:

$$R = \int_0^{E_o} \frac{dR}{(dE/dR)} \qquad \textbf{(10.3)}$$

where E_o is the initial incident ion energy.

The *nuclear stopping process* can be visualized with the aid of the extreme simplification that treats the event as a collision between two hard spheres (Fig. 10-4). A more correct approximation assumes that the scattering is described by a Coulombic force-at-a-distance interaction. In the latter description, an appropriate atomic scattering potential V(r) must be determined. The most successful model for predicting implantation profiles based on the ion stopping approach is the so called *LSS model*, which is discussed in the following section. The LSS model utilizes a modified Thomas-Fermi screened potential for V(r). Calculations based on this model show that nuclear stopping increases linearly in effectiveness at low energies, reaches a maximum at some intermediate energy, and decreases at higher energies (because at high velocity, ions move past target nuclei too quickly to efficiently transfer energy to them). Values of $S_n(E)$ for boron, phosphorus, and arsenic are shown in Fig. 10-5. It is important to note that $S_n(E)$ increases with the mass of the implanted ion, and thus heavy ions such as arsenic will transfer much more of their energy through nuclear collisions than will B atoms.

The *electronic stopping process* can be considered as similar to the stopping of a projectile in a viscous medium, and the stopping magnitude can be approximated to be proportional to the square root of the ion energy:

$$S_e(E) = k_e (E)^{1/2} \qquad \textbf{(10.4)}$$

where k_e is a constant that depends weakly on the ion and target atomic masses and numbers. Figure 10-5 also shows $S_e(E)$ for B, P, and As. As can be seen in Fig. 10-5 the crossover energy at which electronic stopping becomes more effective than nuclear stopping is higher for heavier

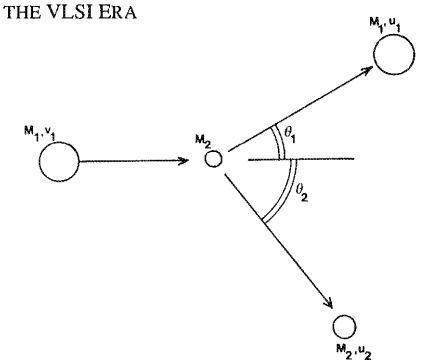


Fig. 10-4 Collision between two particles treated as if they are hard spheres.

ions. For example, for boron $S_e(E)$ is the predominant energy loss mechanism down to ~10 keV, while for P and As, the energy loss due to nuclear stopping predominates for energies up to 130 keV and 700 keV, respectively.

10.2.3 Models for Implantation Profiles in Amorphous Solids

The range-energy relation given by Eq. 10-2 was reformulated by Lindhard, Scharff, and Schiott (LSS) for implantation into amorphous material in terms of the reduced parameters ε and ρ, as:

$$\frac{d\varepsilon}{d\rho} = \left(\frac{d\varepsilon}{d\rho}\right)_n + k_\varepsilon (\varepsilon)^{1/2} \qquad (10.5)$$

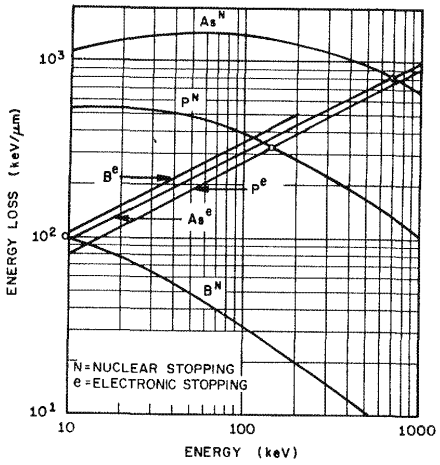


Fig. 10-5 Calculated values of *dE/dx* for As, P, and B at various energies. The nuclear (N) and electronic (e) components are shown. Note the points (o) at which nuclear and electronic stopping are equal. (After Smith, Ref. 7, redrawn by Seidel in Ref. 83. Copyright 1983, Lucent Technologies, Incorporated, reprinted by permission.)

where ρ and ε are dimensionless variables related to the range R, and the incident energy E_o by:

$$\rho = \frac{(RNM_1M_2 4\pi a^2)}{(M_1 + M_2)} \tag{10.6a}$$

and

$$\varepsilon = \frac{E_o aM_2}{[Z_1 Z_2 q(M_1 + M_2)]} \tag{10.6b}$$

where: M_1 and M_2 are the mass of the incident ions and target atoms, respectively; N is the number of atoms per unit volume; a is the screening length [equal to $0.88a_o/(Z_1^{1/3} + Z_2^{2/3})^{1/2}$, in which a_o is the Bohr radius and Z_1 and Z_2 are the atomic numbers of the ion and target].

LSS used a modified Thomas-Fermi screened potential to calculate the energy loss due to nuclear stopping, together with the assumption that the energy loss due to electronic stopping is given by Eq. 10-4 (Fig. 10-6). Using this approach, they calculated values of ρ for different values of ε. (Note that these calculations are quite complex, but can be found in the classic paper by LSS.[3]) The value of ρ was then converted to R (using Eq. 10-6a), and finally a value for R_p was obtained from the approximate expression:[8]

$$R_p \cong \frac{R}{1 + [\frac{M_2}{3M_1}]} \tag{10.7}$$

LSS assumed that the distribution of the implanted ions in amorphous materials could be described by a symmetrical Gaussian curve. If this assumption is valid then the implanted ion concentration, n, as a function of depth, x, can be described by:

$$n(x) = \frac{\phi}{\sqrt{2\pi}\Delta R_p} \exp\left[\frac{-(x - R_p)^2}{2\Delta R_p^2}\right] \tag{10.8}$$

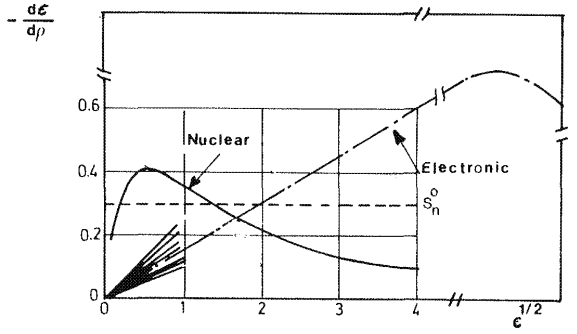


Fig. 10-6 (a) Nuclear stopping power for Thomas-Fermi potential (solid line) and electronic stopping power (dash-dot line) for k = 0.15 in terms of the reduced variables ε and ρ, based on LSS theory.[3]

where: ϕ is the dose (in number of implanted ions/cm^2), and ΔR_p is the standard deviation of the Gaussian distribution (or *projected straggle* of the distribution in the direction of incidence of the beam). The value of ΔR_p is calculated in terms of R_p and the mass of the implanted ions M_1 and the target atoms M_2, by the approximate expression:[8]

$$\Delta R_p \cong \frac{2R_p}{3}\left[\frac{\sqrt{M_1 M_2}}{M_1 + M_2}\right] \tag{10.9}$$

The concentration is maximum at R_p, and Eq. 10-8 at $x = R_p$ reduces to:

$$n(x = R_p) = \frac{\phi}{\sqrt{2\pi}\Delta R_p} \cong \frac{0.4R_p}{\Delta R_p} \tag{10.10}$$

The assumption by LSS that the distribution of implanted atoms in amorphous materials is well approximated by a Gaussian curve is not completely correct, but is nevertheless very useful as a first order description. Indeed, the fit to experiment is almost always good for all implantations near the peak. For example, the peak value predicted by Eq. 10-10 is generally within 1% of the measured value (except for very shallow [low energy] implants). On the other hand, significant asymmetries begin to appear in experimental implant profiles in amorphous targets once the concentration levels drop by a factor of 10 below the peak value (and which are not taken into

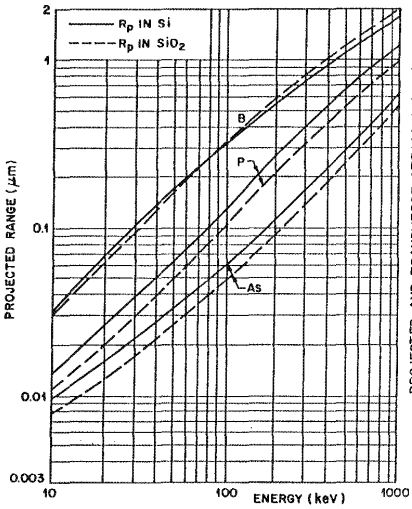


Fig. 10-7 Projected range for B, P, and As in Si and SiO_2 at various energies. The results pertain to amorphous silicon targets and thermal SiO_2 (2.27 g/cm^3). After Smith, Ref. 7, redrawn by Seidel in Ref. 83, Copyright, 1983, Lucent Technologies, Incorporated. Reprinted by permission.

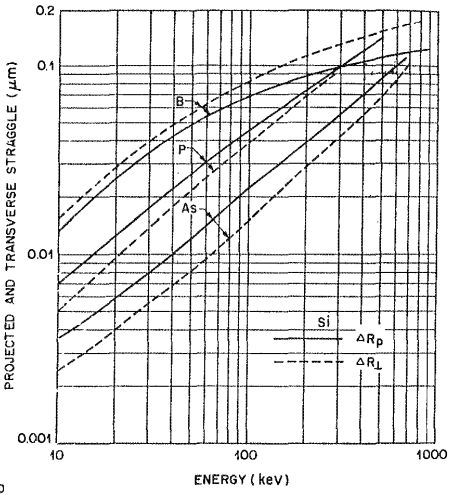


Fig. 10-8 Calculated ion projected straggle ΔR_p, and ion lateral straggle $\Delta R_\perp$ for As, P, and B ions in silicon. After Smith, Ref. 7, redrawn by Seidel in Ref. 83, Copyright, 1983, Lucent Technologies, Incorporated. Reprinted by permission.

account by a symmetrical Gaussian approximation). Therefore, various distribution curves with a higher number of moments than a Gaussian (which can be described with only two moments, R_p and ΔR_p), have been examined for their ability to fit the experimental implant profiles (Fig. 10-1). In addition, LSS theory and the Gaussian curve fail to account for several effects that occur when implants are made into *single-crystal* material. Thus, modifications (or even other analytical models) must be used to obtain a good fit to data obtained in such situations.

Nevertheless, in practice the Gaussian distribution is still commonly used to provide quick estimates of doping distributions into amorphous and single-crystal targets. The higher moment distributions (or alternate models) are subsequently utilized to fine-tune the dose or energy to obtain better results. Several workers have calculated R_p and ΔR_p values for many of the elements that are commonly implanted into Si and SiO_2 (using LSS theory to perform the calculations). Figure 10-7 is an example of such data for B, P, and As into silicon. Values for projected lateral straggle, $\Delta R_\perp$, have also been compiled, and are given in Fig. 10-8 (together with ΔR_p).

Example 10-1: A 150 mm wafer is to be implanted with 100 keV boron atoms to a dose of 5×10^{14} ions /cm^2: 1) Determine the projected range, projected straggle, and peak concentration using Figs. 10-7 and 10-8. 2) If the implantation time is 1 min, calculate the required ion beam current.

Solution: 1) From Figs. 10-7 and 10-8 it is found that the *projected range* and *projected straggle* are 0.32 μm and 0.07 μm, respectively. The peak concentration, $N(x = R_p)$, can be calculated from Eq. 10-10, or:

$$N(x = R_p) = \frac{0.4\,\phi}{\Delta R_p} = \frac{0.4\ (5 \times 10^{14}\ \text{cm}^{-2})}{(0.07 \times 10^{-4}\ \text{cm})} = 2.8 \times 10^{19}\ \text{ions/cm}^3$$

2) To find the beam current, first calculate Q, the total number of implanted ions (where Q = dose * wafer area):

$$Q = (5 \times 10^{14}\ \text{ions/cm}^2) * [\,\pi\,(15/2)^2\,] = 8.8 \times 10^{16}\ \text{ions}$$

Then, the required scanned beam current is determined by dividing the total charge, qQ, by the time of implantation:

$$I = (q\text{Q/t}) = [(1.6 \times 10^{-19}\ \text{C})\ (8.8 \times 10^{16})]/60\ \text{sec} = 0.23\ \text{mA}.$$

10.2.3.1 Higher Moment Distributions for Implant Profiles in Amorphous Material: As noted earlier, even when implanting into amorphous material, the experimental profiles exhibit some asymmetry, or *skewness*. This is not surprising if one considers the forward momentum of the ions. That is, when relatively light atoms make collisions with target atoms (e.g., B in Si), they experience a significant degree of backscattering. Hence more will come to rest at a distance closer to the surface than R_p (causing the concentration near the surface to be higher). On the other hand, heavier atoms will undergo little backscattering, and the concentration on the deep side will be higher. Thus, even if a Gaussian distribution is used to approximate the implantatation profile, such non-Gaussian effects on the behavior of devices can be anticipated. That is, when boron is utilized to implant deep p-wells (e.g., in CMOS technology), higher doping close to the surface is observed than is predicted by a Gaussian distribution. On the other hand, the skewness in arsenic implants will produce deeper junctions than predicted when implanting n^+ source/drain (or n^+ emitter regions) with arsenic.

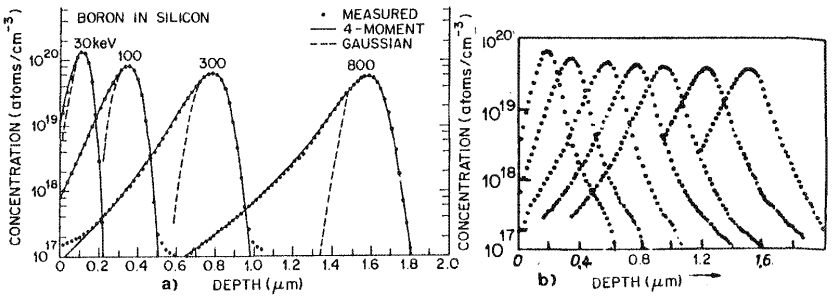


Fig. 10-9 (a) Boron implanted atom distributions, showing 1) measured data points, 2) four-moment (Pearson-IV) curves, and 3) symmetric Gaussian curves. The boron was implanted into *amorphous silicon* without annealing. (b) Depth distributions of boron implanted in *crystalline silicon* in a <763> direction - a dense crystallographic direction at various energies. *Tails* occur because of channeling effects.[5] Reprinted with permission of Philips Research Reports.

To theoretically account for the skewness found in measured profiles, probability distributions with higher order moments must be used to approximate impurity distributions. Gibbons and Mylroie[9] have demonstrated that use of a distribution with a third central moment can closely approximate depth profiles if the asymmetry of the profile is not excessive (i.e., the value of the third central moment is less than ΔR_p). When this approach is taken, the distribution is actually represented by two half-Gaussian profiles, each with their own projected straggle ΔR_{p1} and ΔR_{p2} (Fig. 10-6b), joined together at the depth of their modal range, $R_m = R_p - 0.8(\Delta R_{p1} - \Delta R_{p2})$. The joined half-Gaussian approximation produces a good fit to the profiles of phosphorus and arsenic atoms implanted into Si. The concentration values of such distributions as a function of depth can be calculated from:

$$n(x) = \frac{2\phi}{\sqrt{2\pi}(\Delta R_{p1} + \Delta R_{p2})} \exp\left[\frac{-(x - R_M)^2}{2\Delta R_{p1}^2}\right] \qquad x \geq R_M \qquad \textbf{(10.11a)}$$

$$n(x) = \frac{2\phi}{\sqrt{2\pi}(\Delta R_{p1} + \Delta R_{p2})} \exp\left[\frac{-(x - R_M)^2}{2\Delta R_{p2}^2}\right] \qquad x \leq R_M \qquad \textbf{(10.11b)}$$

and the values for ΔR_{p1} and ΔR_{p2} are found in the tabulated data of Gibbons *et al.*[2]

An approach that represents the implanted profile with a distribution described by *four moments* is more exact than the three-moment approach (and is applicable even when the third central moment is large). In fact, Hofker[5] demonstrated that excellent agreement with measured B profiles in amorphous Si is obtained (Fig. 10-9), by assuming that the implantation distribution can be described by a *Pearson-IV distribution function,* a type of distribution function which can be specified by four moments. The four moments describe various characteristics of the implant profile curve: 1) μ_1 (mean range); 2) μ_2 (straggle); 3) γ_1

(skewness); and 4) β (*kurtosis* - which characterizes the *tail* aspect of the distribution). It was later shown by Ryssel *et al.*,[10] that equally good agreement is found using a Pearson-IV curve for other implanted species into amorphous targets. The mathematical expressions for the Pearson- IV curve are quite lengthy and fall beyond the scope of this text. However, they can be found in Reference 1.

10.2.4 Implanting Into Single Crystal Materials: Channeling

LSS theory matched with an appropriate distribution curve can accurately predict concentration profiles for implantation into amorphous materials. In practice when fabricating ULSI, however, most implants are made into single-crystal Si. In single-crystal lattices there are some crystal directions (known as *channels*) along which the ions will not encounter any target nuclei, and will instead be *channeled*, or *steered* along such open channels of the lattice. Figure 10-10 illustrates the most likely channeling directions. They are (in order), <110>, <111>, and <100>. These are the directions that exhibit channels of decreasing "openness." As the implanted atoms travel along channels, the slowing down is accomplished mainly by electronic stopping, and the ions can penetrate the lattice several times more deeply than in amorphous targets (Fig. 10-11a). At first, this might appear to offer the advantages of being able to produce deeper implanted junctions and less lattice damage. It turns out, however, that the large sensitivity to incident beam direction, and the unpredictable effects of *dechanneling* (leading to anomalous profile tails), make channeling effects difficult to control. These issues have kept the apparent benefits of channeling from being exploited to produce deep implanted profiles.

Thus, instead of seeking to harness the effect of channeling, techniques have been sought to avoid its occurrence. The most widely adopted procedure to minimize channeling has been to tilt the wafer surface relative to the incident beam direction (most commonly by ~7°), so that the lattice presents a dense orientation to the incident beam (i.e., the approximate <763> direction). There are two reasons why this technique *reduces* (but does not eliminate) channeling:

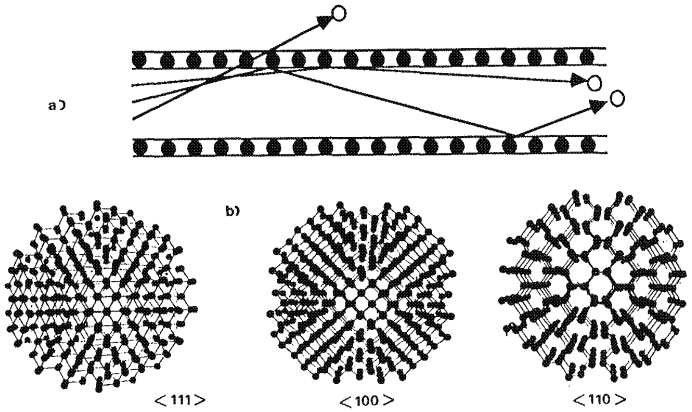


Fig. 10-10 (a) Schematic representation of ion trajectories in an axial channel for various entrance angles. (b) Ball model showing relative degree of "openness" of the diamond (Si) lattice when traversing in <111>, <100>, and <110> directions.

1. Even when atoms enter the lattice in an apparently random, "dense-appearing" direction, they can be diverted into one of the open channels after entering the lattice (Fig. 10-11b). This effect leads to the *tails* that are observed in implanted profiles into single-crystal lattices. Figure 10-9b shows the depth profiles of boron implanted into crystalline Si in a <763> direction. Compared to Fig. 10-9a, it can be seen that tails in the profile exist. The combination of "random" and "channeled" atoms can be used to con-struct a highly accurate description of the net implant profile, as illustrated in Fig. 10-12.

2. If the 7° tilt is performed such that the Si atoms are still inadvertently aligned in a highly symmetric array of planes, the phenomenon of *planar channeling* can still produce channeling effects (Fig. 10-13). Turner *et al*[13] reported that in order to insure that both planar and axial channeling effects are avoided, wafers must also be oriented with an appropriate azimuthal (or *twist*) direction, in addition to a proper tilt angle.

Recommended optimum twist angles for various implantation species, energy, and orientation are shown in Fig. 10-13. It should be noted that control of the twist angle was not possible with many of the gravity-fed wafer handling systems used on older medium-current implanters. However, most implanter suppliers now provide end-stations that make provision for controlling twist angle.

In addition to tilting and twisting wafers, other methods investigated to further reduce

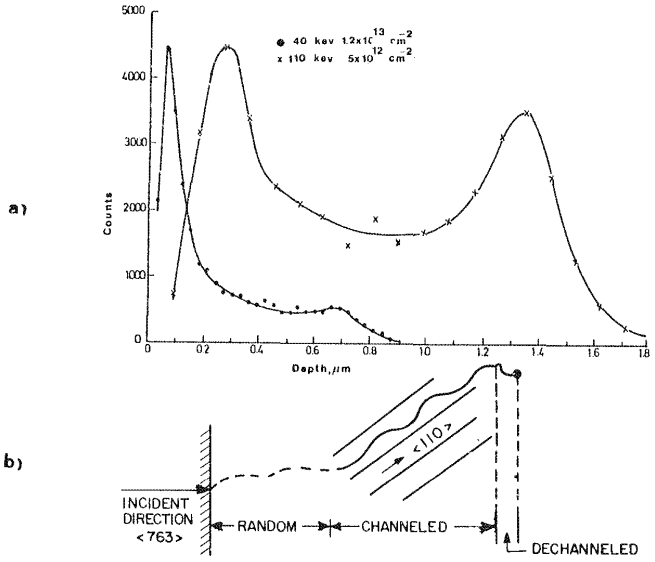


Fig. 10-11 (a) Variation of P^{32}-concentration profiles with ion energy for implantation along <110> axis, plotted on a linear scale.[11] Reprinted with permission of Canadian Journal of Physics. (b) Schematic diagram of an ion path in a single crystal for an ion incident in a "dense" <763> direction. The path shown has non-channeled and channeled behavior.[63] Copyright, 1983, Lucent Technologies, Incorporated, reprinted by permission.

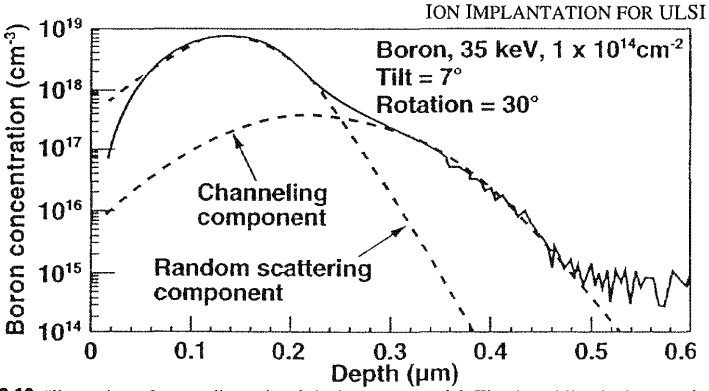


Fig. 10-12 Illustration of a one-dimensional dual Pearson model. The dotted line is the model prediction, and the solid line is the experimental profile measured by secondary ion mass spectroscopy.[12]

channeling effects include: a) pre-amorphizing the lattice by a prior implantation (e.g., with Si or Ge);[14,15] b) implanting through a surface oxide (to randomize the directions of the ions as they enter the lattice);[14,15,16] and c) using heavy ions for the implantation (e.g., BF_2 rather than B). These and other issues associated with channeling are discussed in detail by Simonton.[17]

10.2.5 Calculating Implantation Profiles: Boltzmann Transport Equation and Monte Carlo Approaches

When fabricating integrated circuits, ion implantation is frequently performed through a thin surface layer (e.g., SiO_2, Si_3N_4, or even a composite layer). As discussed in a later section, if the stopping powers of the various layers are not too different, simple approximations can be used to fit actual profiles fairly well. More accurate predictions of profiles into multilayered materials, however, require either a numerical solution of the Boltzmann transport equation (BTE)[18,19] or a Monte Carlo simulation.[20,21] The assumption used by the LSS theory that the target material is homogeneous and isotropic makes the LSS model invalid for predicting implantation ranges in multilayer structures. In both the BTE and Monte Carlo methods, the predicted distribution is calculated directly (rather than inferred from a range value and an assumed distribution curve).

Accurate and "computationally efficient" as-implanted profiles are needed as the inputs to the process simulators that calculate dopant diffusion (such as SUPREM and FLOOPS). Such as-implanted doping profiles are available from a large body of experimental data on dopant range profiles that has been organized into multi-parameter fits for "random" (or "amorphous material") and "channeled" (or crystalline material) profiles. The two types of profiles are each fit to a Pearson-IV distribution with the ratio of the random to channeled portions related to the damage and accumulated dose. In recent implementations of this approach, the Pearson-IV distributions have been replaced by the more general Legendre polynomials.[12,22,82]

10.2.5.1 Monte Carlo Calculations: In the Monte Carlo approach, ion implantation is simulated by following the history of an energetic ion through successive collisions with target atoms using the binary collision assumption. The calculation of each trajectory begins with a given energy, position, and direction. A large number of ion trajectories are calculated (>10^3), and the

depth at which each ion stops is determined. The predicted profile is generated by plotting histograms of the number of ions stopped within each depth interval. Monte Carlo simulations can be performed for either amorphous or crystalline targets. In the amorphous material simulation, the position of the target atoms follows a Poisson distribution, while in crystalline target simulation the atom positions are specified to correspond to the positions that they would assume on a lattice.

A widely used, PC-based Monte Carlo range and damage simulation program is named TRIM (for Transport *of Ions in Matter*).[23] It assumes that the target material is amorphous (i.e., no channeling effects occur), and consequently provides a very useful first-approximation to the atom range and damage distributions in multi-element and multi-layer materials.[24] A more advanced Monte Carlo code has been developed for Si-like crystalline targets (UT-MARLOWE). It runs on Unix-based workstations and provides highly accurate atom range and damage distributions (including the effects of damage accumulation on channeled profiles).[84,85]

10.3 ION IMPLANTATION DAMAGE ACCUMULATION AND ANNEALING IN SILICON

As discussed earlier, ion implantation has many advantages for ULSI processing, the most important being the ability to control (to a high degree of precision) the number and depth of impurity atoms introduced into a substrate. The price for such benefits is high, however, as implantation cannot be achieved without damage to the substrate material. That is, high-energy ions collide with substrate atoms and displace them from their lattice sites in large numbers. Furthermore, only a small percentage of the as-implanted atoms end up on electrically active lattice sites. In order to successfully fabricate devices, the damaged substrate regions must be

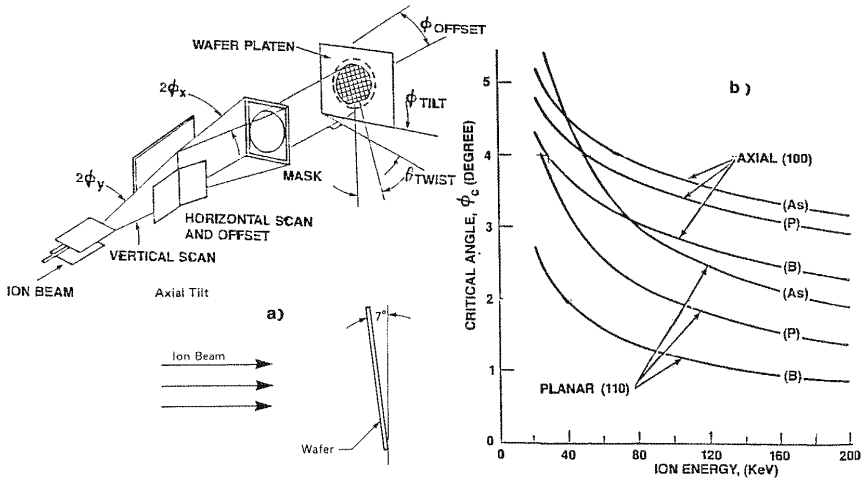


Fig. 10-13 (a) Channeling avoidance by using axial tilt and azimuthal twist. (b) Calculated critical angles vs. ion energy for channeling along the axial Si <100> direction and planar <110> direction for B^+, P^+, and As^+ ions.[13] Reprinted with permission of Solid State Technology, published by PennWell.

must be restored to their pre-implanted structure, and the implanted species must be electrically activated. Described in this section are: a) the aspects of implantation damage; b) the mechanisms of damage annealing; and c) the electrical activation of implanted dopants. Figure 10-14 illustrates the subjects that will be covered. At the conclusion of the section it will be noted that as devices are designed with very shallow junctions (e.g., ≤250 nm deep), the remaining residual implantation damage (even after the annealing process) represents a mechanism that degrades device operation. An example is the excess reverse-bias leakage current observed when such shallow junctions are fabricated.

10.3.1 Implantation Damage in Silicon

When energetic ions strike a silicon substrate they lose their energy in a series of nuclear and electronic collisions, and thus rapidly come to rest some hundreds of atom layers below the surface. Only the nuclear collisions result in *displaced silicon atoms* (also referred to as *damage* or *disorder*). An individual nuclear collision can result in different types of displacement events, depending upon the magnitude of the energy transferred.

If the energy transferred to a silicon atom (ΔE_n) is less than the energy required to displace it from its lattice site, E_{di}, no displacement event results. If $2E_{di} \geq \Delta E_n \geq E_{di}$, a single displacement and simple isolated point defects are created. If $\Delta E_n \geq 2E_{di}$, point defects and *secondary displacements* (i.e., recoiled lattice atoms with enough energy to generate additional lattice disorder) are produced. Finally, if $\Delta E_n >> 2E_{di}$, multiple secondary displacements and defect clusters are created. (Note, E_{di} [Si] ~15 eV.)

Because of their extremely small size, the exact nature of isolated defects and defect complexes from ion implantation are hard to characterize. However, each displacement of a lattice atom (whether by a primary beam ion or an energetic recoil lattice atom), produces a *Frenkel defect* (see Chap. 2). In addition, it is widely agreed that the defects include: a) vacancies [V]; b) di-vacancies [V^2] (i.e., two vacancies bound together); c) higher order vacancies and (vacancy-impurity complexes); and d) interstitials [I]. The ions create zones of gross disorder populated by such defects in regions where they deposit their kinetic energy (Fig. 10-15). These zones are vacancy-rich at the center, and are surrounded by [I], since each displaced Si atom moves into the lattice with velocity components perpendicular to the ion track.

The lattice in these disordered regions exhibits several different damage configurations:

1. *Isolated point defects* or *point defect clusters in essentially crystalline silicon* (i.e., the type of damage that results from implanting light ions, or when $\Delta E_n \cong 2E_{di}$).
2. *Local zones of completely amorphous material* in an otherwise crystalline layer (i.e., an *amorphous region* is defined as a region in which the displaced atoms per unit volume approach the atomic density of the semiconductor). Local zones of amorphous damage are associated with low-dose implants of heavy ions (i.e., $\Delta E_n >> E_{di}$).
3. *Continuous amorphous layers* which form as the damage from the ions accumulates. That is, as the dose of ions (typically heavy ions) increases, the locally amorphous regions eventually overlap, and a continuous amorphous layer is formed.

In our discussion Type-1 and Type-2 damage will be grouped into the category of *primary crystalline-defect damage.* Type-3 damage will be referred to as *amorphous layer damage.* The basis for this grouping (as shall be seen in the sections that discuss damage annealing), is that the annealing strategy for Type-1 and Type-2 damage are the same, but a different annealing procedure is employed for Type-3 damage.

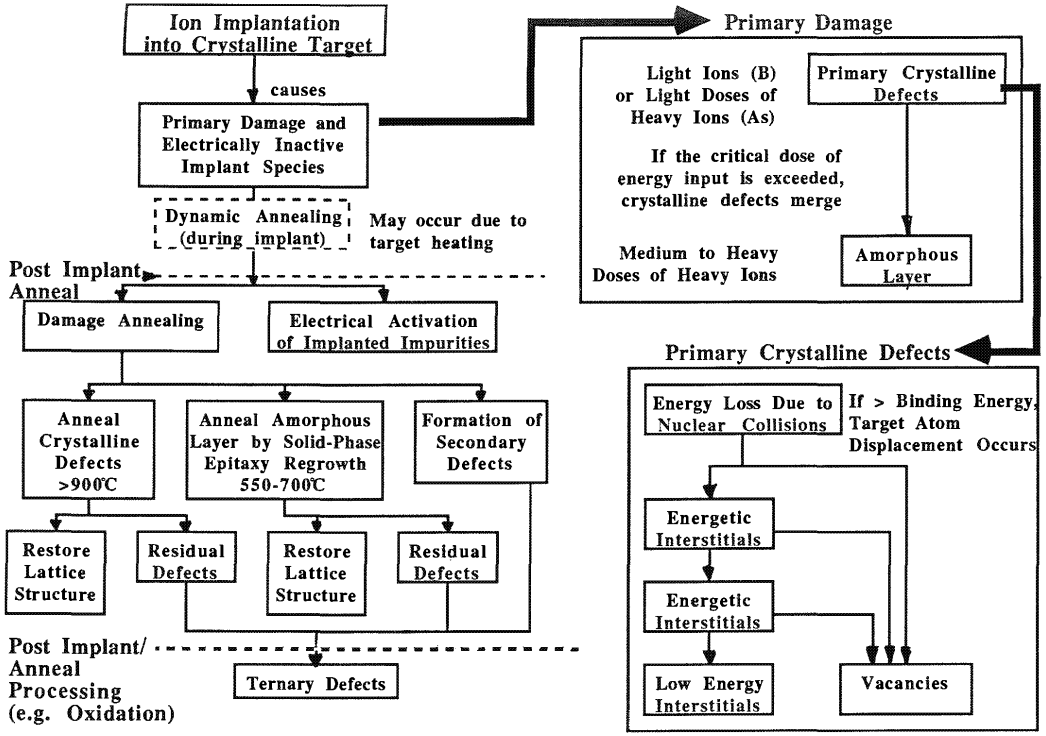


Fig. 10-14 Ion implantation damage and annealing.

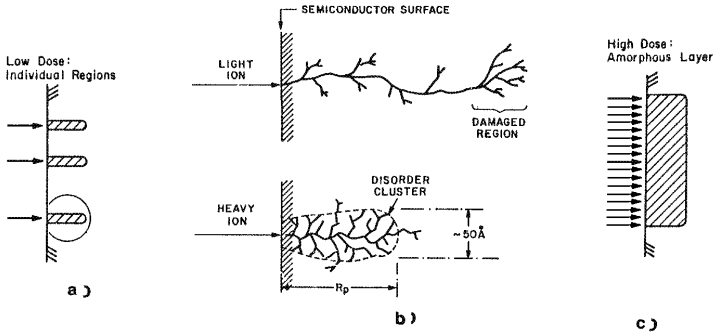


Fig. 10-15 Schematic representation of the disorder produced by ion implantation. (a) Low dose light ions - individual regions with degree of disorder increasing as ions penetrate deeper into substrate. (b) Low dose heavy ions - individual regions of more uniform disorder along entire ion trajectory. (c) Heavy doses - formation of an amorphous layer.

Regardless of the form of the damage configuration, the number of displaced atoms after implant is almost always larger than the number of implanted atoms. These displaced atoms reduce the mobility in the damaged regions and produce defect levels in the band gap of the material (i.e., deep-level traps, for both electrons and holes) which have a strong tendency to capture free carriers from the conduction and valence bands. Due to the relatively large number of displaced atoms (and the fact that few as-implanted impurities occupy substitutional sites), non-annealed implanted regions typically exhibit high resistivity.

10.3.2 Primary Crystalline Defect Damage

Primary crystalline defect damage is observed after implanting a Si crystal (28 atomic mass units [28 amu]) with relatively light ions (e.g., B [11 amu]), or with light doses of heavier ions (e.g., P [31 amu], As [75 amu], and Ar [40 amu]). As noted in the previous section, the damage configurations from light ions are in fact quite different than those from heavy ion implantation. The reason for the difference is discussed below.

As shown in Fig. 10-5, at the initial impact energies most of the energy loss for *light ions* (in this example, boron) is due to electronic collisions (which do not produce displacement damage). However, as B ions penetrate deeper into the lattice, they lose energy. Eventually, the cross-over energy is reached, below which nuclear stopping predominates (~10 keV for B). As shown in Fig. 10-15, *most lattice damage occurs in the part of the light ion trajectory beyond that point.*

An estimate of the damage can be calculated by considering the case of an 80-keV boron ion. Its projected range is ~250 nm (Fig. 10-7), and its initial *nuclear energy loss is* ~35 eV/nm (Fig. 10-5). The boron ion will thus lose ~8.75 eV from nuclear collisions for each lattice plane that it passes (as the lattice spacing in Si is ~0.25 nm). Since 8.75 eV is less than E_{di} [Si], the B atom (upon initially entering the Si), does not transfer enough energy per nuclear collision to displace silicon atoms from their lattice sites. When the ion energy has decreased, however, to 40 keV (at a depth of ~130 nm in the lattice), the energy loss due to nuclear collisions rises to ~15 eV per

lattice plane (i.e., 60 eV/nm), which is sufficient to displace Si lattice atoms. (Note that although electronic stopping is still predominant, nuclear stopping between 10–40 keV can still displace Si atoms.) Assuming that one Si atom is displaced per lattice plane for the remainder of the ion trajectory, 480 lattice atoms are displaced (i.e., 120 nm/0.25 nm) by the time the boron atom comes to complete rest. If each displaced atom is moved roughly 2.5 nm by such collisions, the damage volume is found from $V_{dam} \cong \pi\ (2.5nm)^2(120\ nm) = 2.4x10^{-18}\ cm^3$. The damage density is $480/V_{dam} \cong 2x10^{20}\ cm^{-3}$, which amounts to only ~0.2% of the atoms. This calculation implies that very large doses of light ions are required to produce an amorphous layer, and for the most part each ion produces a trail of well-separated primary recoiled Si atoms in the wake of the implanted ion. Furthermore, displaced atoms will be separated by short distances from the vacancies they leave, because the energies of their recoils are low. This suggests that only a relatively small input of energy to the lattice could cause such separated pairs to rejoin. In fact, as will be explained in the following section on annealing, a large fraction of the disorder produced during boron implantations is *dynamically annealed during the implantation*, and therefore at room temperatures even high-dose boron implantations may not produce an amorphous layer. The damage from boron implantations is thus characterized by *primary crystalline defects*. Damage density is distributed versus depth as shown in Fig. 10-16a, which shows a sharp buried peak concentration and qualitatively fits the description of the damage-creation process.

When *heavy ions* are implanted, the energy loss is predominantly due to nuclear collisions over the entire range of energies experienced by the decelerating heavy ions (Figs. 10-5 and 10-15). Thus, substantial damage is expected. Examine the case of 80-keV arsenic atoms, which will have a projected range of ~50 nm. The average energy loss due to nuclear collisions will be ~1200 eV/nm over the entire range. As a result, the As atoms lose ~300 eV for each Si atomic plane that they pass. Most of this energy is transferred to a single lattice atom. The recipient Si atom, however, will subsequently produce ~20 displaced lattice atoms. The total number of displaced atoms is thus 4000. Again, assuming an average distance moved for each displaced atom of ~2.5 nm, the damage volume is $V_{dam} \cong \pi\ (2.5\ nm)^2(50\ nm) = 0.8x10^{-18}\ cm^3$. The damage density is then $4000/V_{dam} \cong 5x10^{21}\ cm^{-3}$, or ~10% of the number of atoms in the lattice within the damage volume. Since a single ion is capable of producing such heavy damage, it is reasonable to expect that some local regions of a silicon substrate (which when subjected to even light doses of heavy-ion bombardment) will suffer enough damage to become amorphous. Some damage density distributions due to heavy-ion (e.g., As) implants are shown in Fig. 10-16b. They exhibit a *broad* buried peak that is a replica of recoiled range distribution.

10.3.3 Amorphous Layer Damage

Simple qualitative concepts illustrate how continued bombardment by heavy ions will lead to the formation of continuous amorphous layers. That is, heavy-ion damage accumulates with ion dose through an increase in the density of localized amorphous regions. Eventually these regions overlap, and a continuous amorphous layer is the result. The evolution of a continuous amorphous layer from the accumulation and overlap of damage formed by individual atoms has been observed (Fig. 10-17) using Rutherford backscattering spectroscopy in a channeling mode. In Fig. 10-17a it can be seen that damage produced by 1.7 MeV Ar^+ ions in Si builds up to an initial damage distribution with a peak at a depth of ~1.3 μm. At that depth, individual amorphous zones are likely to be created by each ion (Fig. 10-17b). Closer to the surface the damage consists predominantly of isolated defect clusters (akin to the damage caused by light

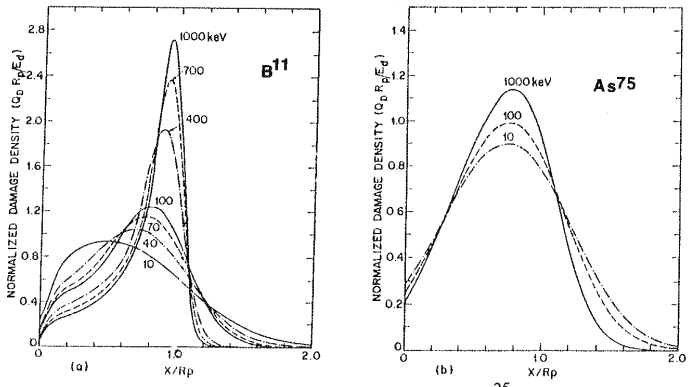


Fig. 10-16 Calculated damage density profiles of (a) boron, and (b) arsenic.[25] Reprinted with permission of Plenum Publishing Company.

ions.) As the dose is increased, both types of damage increase, and finally a 1.5 μm-thick, (almost continuous, as in the example of Fig. 10-17a) amorphous layer extends to the surface. This also illustrates another aspect of the formation of amorphous layers. That is, amorphization begins at the depth of the maximum nuclear collision energy deposition, at slightly less than the projected range, and spreads towards the surface (as well as towards deeper positions in the target). In addition, the interface with the single-crystal region below is not a well-defined plane, due to the statistical nature of the penetration. Beyond the interface, primary crystalline damage as well as a considerable concentration of Si interstitials (which had diffused out of the damage clusters during implantation) is expected.

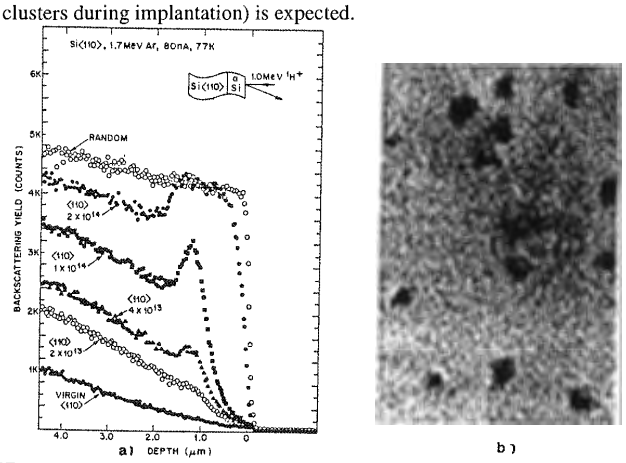


Fig. 10-17 (a) RBS channeling spectra of damage accumulation in Si at liquid nitrogen temperature following bombardment with 1.7 MeV Ar^+ ions.[26] By permission of the Harvard University Archives. (b) Bright-field electron micrographs of amorphous regions in Si produced by bombarding to doses of $3x10^{11}$ cm^{-2} with 10 keV bismuth ions.[27] Courtesy of the Institute of Physics Conference Series.

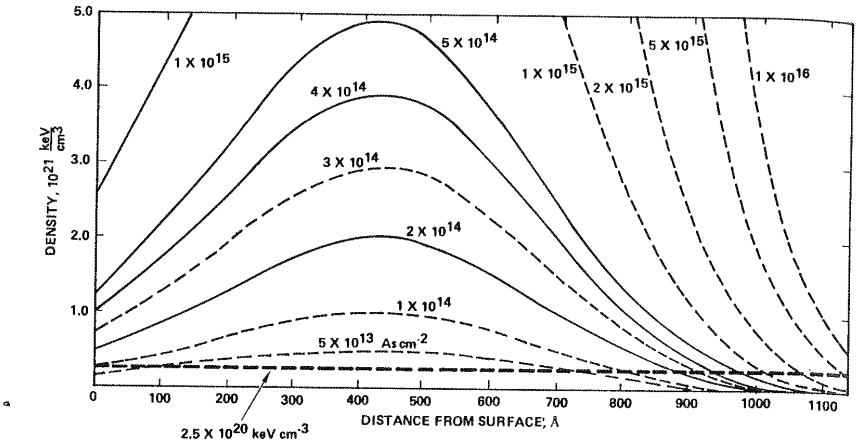


Fig. 10-18 Damage density distribution for 100-keV As implantations for doses of $5x10^{13}$–$1x10^{16}$ As cm^{-3} calculated from Brices[25] curves.[28] Reprinted by permission of American Physical Society.

The minimum (or *threshold*) dose required to convert a crystalline material to an amorphous layer can be arrived at from a number of different viewpoints. First, it is certain that when the number of stably displaced Si atoms reaches the number of Si atoms/unit volume (e.g., $5x10^{22}$ cm^{-3}), the material has become "amorphous." When the damage density reaches this value, an amorphous condition is reached. Another view holds that there is a critical energy density that must be placed into the crystal to make it amorphous.[29] This critical energy E_c is given by:

$$E_c \cong f\,(10^{21}\ \text{keV})/\text{cm}^3 \qquad \textbf{(10.12)}$$

and the pre-factor f is 0.1–0.5 for Si. Based on this model, Brice published tables that predict the extent of the amorphous layer formation for different doses (and constant implant energy). Prussin *et al.*,[28] plotted the data from these tables for As implanted into Si (Fig. 10-18). A simple estimate can also be obtained by assuming that if enough energy is applied to the crystal to cause melting (for Si, 10^{21} keV/cm^3); this would also produce an amorphous layer. For 100-keV As ions, the dose D_{crit} calculated from this approach (where E_o is the beam energy in keV, and R_p in cm), is:

$$D_{crit} = [(10^{21}\ \text{keV/cm}^3)\,R_p]/E_o = 6x10^{13}\ \text{ions/cm}^2 \qquad \textbf{(10.13)}$$

From the discussion on minimum implant dose for creating an amorphous layer, three more important aspects of amorphous layer structure can be inferred:

1. From Brice's model (Fig. 10-18), it can be seen that some implant doses will cause an amorphous layer to form below the surface of the Si, but that near the surface not enough energy has been transferred to the lattice to cause the material to become amorphous (e.g., in the case of As at 100 keV, for doses less than $1x10^{14}$ cm^{-2}, the energy density at the surface drops below the threshold required to create an amorphous layer). Thus the amorphous region does not extend all the way to the surface, but is a *buried amorphous layer.* Such buried amorphous layers have been found to exhibit profoundly different annealing characteristics than those that extend all the way to the surface.

2. In the crystalline layer just below the amorphous layer, the silicon was the recipient of impurity atoms, albeit in lower quantities than necessary to cause an amorphous region. Yet, heavy primary-crystalline-defect damage from the implanted ions exists in this zone, as well as implanted impurity atoms that must be electrically activated in order to achieve the full electrical activation of the dose. These cannot be ignored when considering an annealing strategy for amorphous layers.

3. Prussin *et al.*,[28] have shown that the critical energy, E_c, required to create an amorphous layer depends on the implanted species, and is 2.5×10^{20} keV/cm^3, 1.0×10^{21} keV/cm^3, and 5.0×10^{21} keV/ cm^3, for As, P, and B, respectively.

The temperature of the substrate during implantation also impacts amorphous layer formation. Figure 10-19 shows a plot of the critical dose versus reciprocal temperature for various ions.[28] It can be seen that for light ions such as B^{11}, a 50°C rise above room temperature prevents the formation of an amorphous layer at any dose. This is due to the fact that B generates only a few stably displaced atoms during implantation, and a small increase in T above room temperature allows such closely spaced vacancy-interstitial pairs (Frenkel defects) to recombine.

10.3.4 Electrical Activation and Implantation Damage Annealing

In this section the methods that can be used in attempts to restore the silicon to its pre-implanted structure, and to electrically activate the implanted impurity atoms are considered. The subjects will be covered in the following order: a) electrical activation of implanted impurities; b) annealing primary crystalline defect damage; c) annealing of amorphous layers; d) dynamic annealing effects; and e) diffusion of implanted impurities.

10.3.4.1 Electrical Activation of Implanted Impurities: Since most as-implanted impurities do not occupy substitutional sites, a subsequent thermal step is employed to bring about electrical activation. The degree to which the thermal procedure is effective in electrically activating impurities is commonly determined by Hall effect measurements, but can also be checked more simply by measuring the sheet resistance, R_S.[30] To a first approximation, R_S is inversely proportional to the mobility μ and the implanted dose ϕ (cm^{-2}):

$$R_S \propto 1/(q\,\mu\,\phi) \qquad \textbf{(10.14)}$$

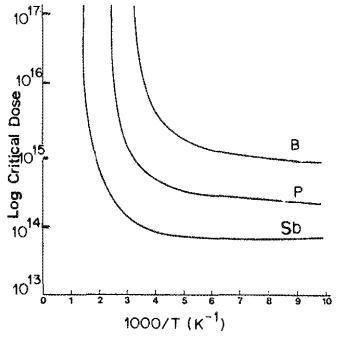


Fig. 10-19 Theoretical relationship between amorphous threshold and temperature for B^+, P^+, and Sb^+ into Si.[29] Reprinted with permission of Gordon and Breach Publishing Company.

where q is the electronic charge. Although μ depends strongly on the concentration of doping atoms and implantation damage, values of R_S have been tabulated utilizing known mobility data.[31] For a known dose, full electrical activity is reached when the predicted R_S is reached.

Electrical activation of implanted impurities in amorphous layers proceeds differently than in layers with primary crystalline damage. As will be discussed, electrical activation in amorphous layers occurs as the impurities in the layer are incorporated onto lattice sites during recrystallization. Electrical activation in crystalline damaged regions exhibits more complex behavior.

For example, Fig. 10-20 shows the *isochronal* electrical activation behavior of implanted boron (i.e., anneals performed at varying temperatures, but for identical times). In this curve, the measured surface carrier concentration (normalized for different junction depths to the dose, in cm^{-2}) is used to indicate the degree of activation. That is, when $p_{Hall}/\phi = 1$, full activation is reached. Note that other impurities exhibit similar behavior to that shown in Fig. 10-20, provided the implantation does not cause a continuous amorphous layer to be formed.

The temperature range up to 500°C (Region 1 of Fig. 10-20) shows a monotonic increase in electrical activity. This is due to the removal of trapping defects and the concomitant large increase in free carrier concentration as the traps release the carriers to the valence or conduction bands. In Region 2 (500–600°C), substitutional B concentration actually decreases. This is postulated to occur as a result of the formation of dislocations at these temperatures. Some boron atoms that were already on substitutional sites are believed to precipitate on or near these dislocations. In Region 3 (>600°C), the electrical activity increases until full activation is achieved at temperatures ~800–1000°C. The higher the dose, the more disorder, and the higher the final temperature required for full activation. At such elevated temperatures, Si self-vacancies are generated. They migrate to the B precipitates, allowing boron to dissociate and fill the vacancy (i.e., a substitutional site).

Activation of implanted impurities by rapid thermal processing (RTP, see Chap. 9) has also been studied. The time-temperature cycle to reach minimum sheet resistance for As, P, and B is ~5–10 sec at 1000–1200°C, the exact condition being dependent on implanted species, energy, and dose.[33,34]

10.3.4.2 Annealing of Primary Crystalline Damage: Isolated point defects and point defect clusters (that predominantly occur during light ion implantation), and locally amorphous zones (that are typically observed from light doses of heavy ions), are both regions of primary crystalline damage that exhibit comparable annealing behavior. At low temperatures (up to ~500°C), vacancies and self-interstitials that are in close proximity undergo recombination, thereby removing trapping defects. At higher temperatures (500–600°C), as described above, dislocations start to form, and these can capture impurity atoms. Temperatures of 900–1000°C are required to dissolve these dislocations. Note that the activation energy of impurity diffusion in Si is always smaller than that of Si self-diffusion (see Chap. 9). Therefore, the ratio of defect annihilation to the rate of impurity diffusion becomes greater as the temperature is raised. This implies that *the higher the anneal temperature the better*, with the upper limit being constrained by the maximum allowable junction depth dictated by the device design.[30]

It is also important that the steps used to anneal implantation damage be conducted in a neutral ambient, such as Ar or N_2. That is, dislocations which form during annealing[35] can serve as nucleation sites for oxidation induced stacking faults (OISF, see Chap. 2) if oxidation is carried out simultaneously with the anneal (i.e., the annealing is performed in an oxygen ambient).

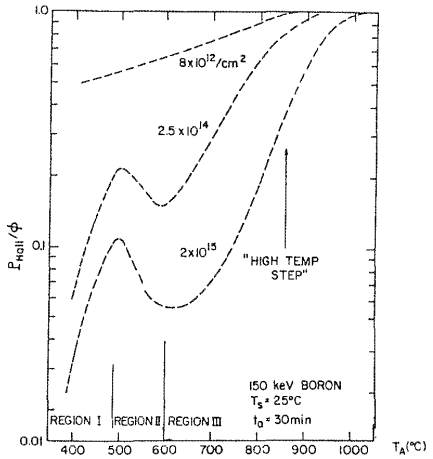


Fig. 10-20 Isochronal annealing behavior of boron. The ratio of free-carrier content, p_{Hall}, to dose ϕ is plotted against anneal temperature T_A for three doses of boron.[32] Reprinted with permission of Gordon and Breach Publishing Company.

10.3.4.3 Annealing of Amorphous Layers: The annealing of continuous amorphous layers that extend to the Si surface has been found to occur by solid-phase epitaxy (SPE),[36] at temperatures between 500–600°C (Fig. 10-21). That is, a recrystallization process occurs on the underlying crystalline substrate, and regrowth proceeds toward the surface. The amorphous layer regrows at varying rates, depending on the annealing temperature, crystal orientation, and implanted species, although at 600°C regrowth is usually completed in a matter of minutes. Regrowth is faster on (100) than (111) Si, and impurities such as B, P, and As enhance regrowth, while O, C, N, or Ar retard regrowth. In practice, 550°C is favored, since (100) regrowth at this temperature occurs at a controllable and reproducible rate.

The impurity dopant atoms are swept into substitutional lattice sites during the SPE regrowth, and thus full electrical activation within the amorphous layer can be obtained at relatively low temperatures. The impurities that are implanted into the region beyond the amorphous layer, however, must be subjected to the higher temperatures needed to electrically activate impurities in regions of primary crystalline damage (Fig. 10-22). Therefore, to fully activate an amorphous layer, and the region of heavy primary crystalline damage behind it, higher temperature anneals than 600°C must be used (normally 800–1000°C). Note that specific *minimum* times and temperatures depend on the particular implanted species and dose.

If the amorphous layer does not extend all the way to the surface (i.e., a buried amorphous layer), the annealing proceeds differently.[38] That is, SPE occurs at both amorphous-single crystal interfaces, and the regrowing interfaces meet below the surface. This meeting point has been found to be a heavily damaged layer, with properties likely to cause device degradation. Thus, it is prudent to select implantation conditions that avoid the formation of buried amorphous layers. Note that if a wafer exhibits color bands after implantation and anneal, they are probably symptomatic of a subsurface damage layer left behind by a buried amorphous

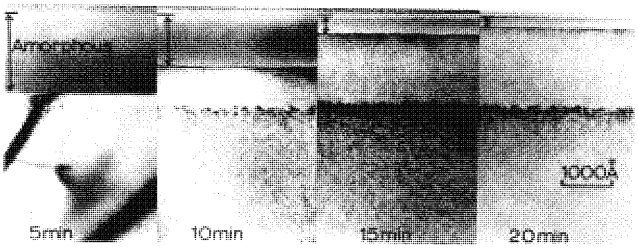


Fig. 10-21 Solid phase regrowth of a 200 keV, $6x10^{15}/cm^2$ antimony implantation at 525°C. TEM cross section micrograph. Courtesy of Institute of Physics, Conference Series.[36]

layer. The gives rise to optical interference effects from the light reflected off the subsurface damage layers and the surface.

Some of the crystalline defects in the region beyond the amorphous layer are annealed out during subsequent thermal cycles, but others give rise to extended defects (such as dislocation loops and stacking faults), which then grow and interact. Under some implantation and annealing conditions, these defects move to the surface and eventually disappear, while in others they grow into larger structures which intersect the surface or remain in the bulk.

In a large number of cases, the total number of "excess" Si atoms found in dislocation loops after thermal annealing fits a remarkably simple model. Even though hundreds to thousands of atoms are displaced from their positions in the Si lattice by each ion impact, the number of residual Si atoms left out of the lattice after the completion of thermal annealing increases with ion dose and is closely proportional to the number of ions. This is known as the "+1" model, where the number of excess Si interstitial atoms is equal to the number of implanted dopant atoms that occupy lattice sites after thermal annealing.[38,39] These Si interstitials arise from the dopant atom replacing the Si-atom in the lattice.

10.3.4.4 Dynamic Annealing Effects: The heating of the wafer during implantation can impact the implantation damage and the effects of subsequent annealing. A rise in temperature increases the mobility of the point defects caused by the damage, and this gives rise to healing of damage even as the implant process is occurring, hence the name *dynamic annealing.*[40] In the case of light ions, sufficient damage healing may occur to prevent the formation of amorphous layers, even at very high implantation doses. In the case of heavy ion implantations, dynamic annealing can cause amorphous layer regrowth during the implantation step.

A study by Prussin *et al.,*[41] showed that the wafer cooling capability of an ion implanter can impact the structure of the damage following implantation because of dynamic annealing effects. That is, if a wafer is prevented from being significantly heated above room temperature by adequate heat sinking during implantation, dynamic annealing is minimized. On the other hand, if no heat sinking is provided and wafers are allowed to rise to temperatures ~150–300°C, dynamic annealing effects can produce changes in implantation damage structures. This typically occurs in non-reproducible and unwanted ways, such as the formation of buried amorphous layers, or crystalline layers containing high densities of dislocation loops.

10.3.4.5 Diffusion of Implanted Impurities: As described in Chap. 9, the diffusion of impurities in single-crystal Si is a complex phenomenon. The diffusion of impurities in implanted Si is

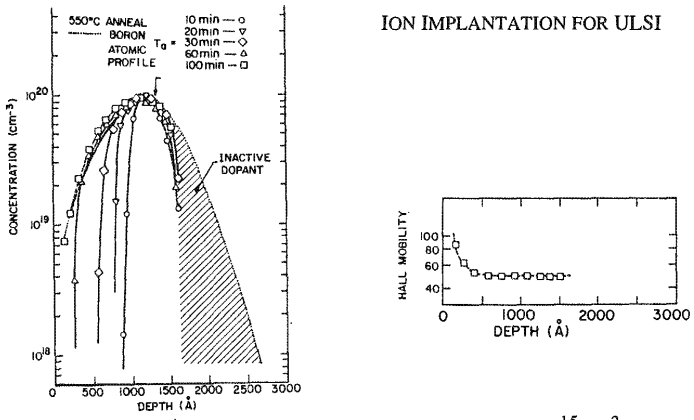


Fig. 10-22 a) Profiles for BF_2^+ implanted into <100> Si at 150 keV and $10^{15}/cm^2$ after different isothermal anneals. The dotted curve is the as-implanted atomic profile from SIMS analysis. The original amorphous-crystalline interface is denoted by the arrow, SPE is complete after ~100 minutes. The hatched region is electrically inactive.[64] b) Free-carrier concentration and mobility for implanted layers which illustrate dopant incorporation by solid phase epitaxy (SPE). Reprinted with permission of the American Physical Society.

even more complicated as a result of the presence of implantation damage. As an example, empirical studies of the *diffusion of B* in implanted single-crystal Si indicate that at high temperatures (≥ 1000°C), the data appears to obey ordinary diffusion theory (Fig. 10-23). At lower temperatures, however, ordinary diffusion theory does not accurately predict the diffusion behavior. That is, at 900°C the boron profile can be closely fit only if a diffusion constant is used that is about three times the value observed under a chemical 900°C diffusion. In addition, at 700–800°C, the depth of profile peak remains fixed, but the concentration in the tail is much deeper than is predicted from normal boron diffusion constants. This increase in the B diffusion (by as much as 4x) in the implanted profile tail occurs during the initial stage of the thermal anneal, but slows to the equilibrium diffusion rates for the rest of the anneal. Consequently, this

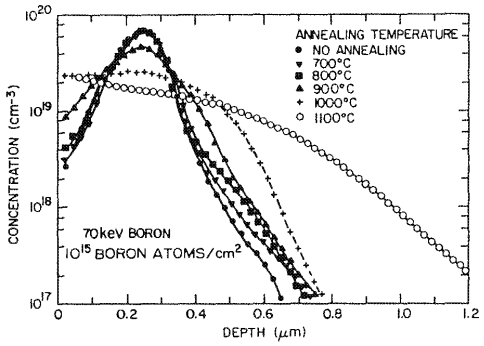


Fig. 10-23 Boron concentration as a function of annealing at various temperatures. The anneal time is 35 minutes.[5] Reprinted with permission of Philips Research Reports.

phenomenon is called *transient enhanced diffusion* (or *TED* – which is discussed in detail in Chap. 9). It is driven by the extra defects and Si interstitial atoms that are present after the implant, but which are rapidly annihilated during the first stages of thermal anneals.

In early semiconductor device fabrication processes, diffusion of impurities during annealing was often used to drive them to depths beyond the range of implantation damage. This resulted in the production of junctions which were not degraded by lattice damage. Present demands for shallow junctions in CMOS (as well as for narrow-base and shallow-emitter regions in bipolar devices), no longer allow extensive dopant redistribution during anneal. Therefore rapid thermal processing (see Chap. 8) is used to anneal implantation with minimal impurity redistribution. RTP cycles of ~1000°C for 10 sec can activate implanted layers as effectively as 30 minute furnace anneals at 1000°C (Fig. 10-24), but with impurity redistribution distances of only a few hundred angstroms (compared to several thousand angstroms for furnace anneals). RTP cycles for shallow junction annealing often use "spike" anneal temperature profiles, where a fast temperature ramp up is followed immediately by a cooling down after the anneal temperature is reached. Shallow junctions have also been fabricated by "pre-amorphizing," in which a combination implantation is performed as a way to reduce the depth of the tail from channeling. That is, an implant of Si or Ge is first carried out to amorphize the Si surface, and then the dopants (such as B) are implanted. Annealing with RTP is conducted after implanting the desired impurity (see also Sect. 11.6.4 which deals with *Shallow Junctions*).

The redistribution of impurities in polysilicon should also be considered. It is observed that dopants redistribute themselves much more rapidly in polysilicon than in single-crystal Si (as a result of grain boundary diffusion). Thus, even short RTP cycles that anneal implantations in single-crystal Si without appreciable redistribution are likely to uniformly distribute impurities throughout a thin film of polysilicon. This is useful for producing polysilicon gates with reduced boron depletion effects. For some impurities (e.g., As), a capping oxide must be present to prevent significant As outdiffusion from the polysilicon by such cycles.

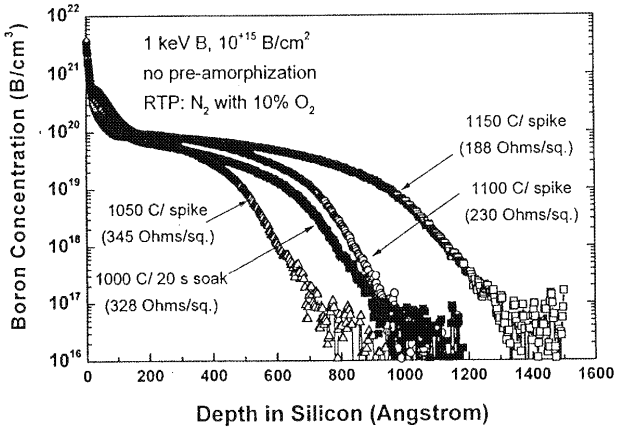
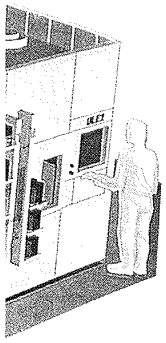


Fig. 10-24 Atomic profiles for 1 keV boron implants after a soak anneal at 1000°C for 20 sec or spike anneals at 1050°C, 1100°C, and 1150°C.[103] Reprinted with permission of Elsevier.

8. A *computer* and *control system* provides recipe-driven operation of the implanter.[45] Linkage between the implanter control system and a fab-host computer network is used to track wafer lots, to specify implant process conditions, and to forward run statistics to process control and tool performance functions.



terials, Inc.

eam.
y measures
lant time),
on systems
[> 5 mA]
ers in front
end station
um-current
ing wafers

-real time),
he entrance
fering with
ons created
is could es-
ild result.
; area, the
he vacuum
^{5} torr) are
cceleration
th residual
in regions
ifer charge
end-station
ion source

10.4.2 Ion Implanter Types

Ion implantation systems have evolved into a number of distinct tool types because of the wide range of ion implantation process applications that are encountered in ULSI fabrication. When ion implanters began to be widely used in IC fabrication in the early 1980's, two distinct types of implanters were commercially offered:

1. *Medium-current implanters*, which produced beam currents ranging from a few μA up to about 1 mA, and which operated over a useful energy range of 20–200 keV. Such medium-current machines were used for the lower-dose implants that comprised the majority of the doping steps.

2. *High-current implanters*, which could produce beam currents up to 30 mA and could operate at maximum energies from 80 to 200 keV. The high-current implanters were used for doping steps which required doses above 10^{15} ions/cm.[1] Such high doses are primarily needed in processes used to form the source/drain regions in MOSFETs and the collector/emitter regions of bipolar transistors.

As integrated circuit fabrication migrated toward the ULSI era (and to deep-sub-micron feature sizes), the designs for implanters were driven to divide functionality in a different way than before. Implanter categories were instead split into two groups according to their useful energy range. These two new implanter classifications are;

3. *Low-energy implanters*, which are high-current tools capable of operating with maximum beam currents of 1 to 20 mA over an energy range of 0.2–80 keV.

4. *High-energy implanters*, which can implant ions at energies from 0.5 to 2–3 MeV.

The low-energy machines are used to form the source/drain and channel regions of the transistor (with junctions depths of 30 to 100 nm), and the high-energy machines are used to produce the deeper (1-2 μm) CMOS wells and isolation junctions. We will now describe each of these four implanter types will now be described.

10.4.2.1 Medium-Current Implanters: *Medium-current implanters* are distinguished not only by their maximum beam-current output, but also by the fact that they are *serial processing machines* (i.e., machines that implant one wafer at a time). The maximum beam-current is ~2 mA and several sets of magnetic or electrostatic deflection and focus elements scan the ion beam in a square pattern in the x and y directions (Fig. 10-14a). The scanning frequencies range from 10 Hz to 1 kHz. Modern medium-current machines use beamline and mechanical wafer scanning designs that allow for a parallel beam path (i.e., constant beam incidence angle on the wafer) as the beam passes over the wafer. These systems are typically used in CMOS applications to perform the threshold-voltage-adjust, well, and isolation implantations. In bipolar technology they are employed to perform the resistor, base, and isolation implantations. The maximum practical dose in these machines is in the range of 10^{14}–10^{15} ions/cm^2.

How kinetic energy possessed by the bombarding ions is transferred to the lattice, and the fact that this causes the wafer temperature to rise was described earlier. Even for relatively low

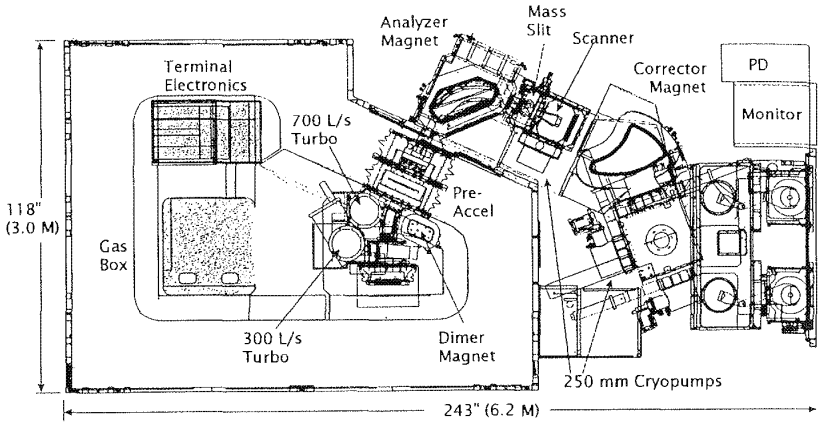


Fig. 10-30 Beamline for a VIISta-810,® ribbon-beam implanter with a mechanical single-wafer scan. Courtesy of Varian Semiconductor Equipment Associates.

10.4.2.4 High-Energy Implanters: *High-energy implanters* are machines which operate with ion beam energies from ≈400 keV up to several MeV, but with relatively low beam currents (0.005-0. 5 mA). The high beam energies are achieved by using either a LINAC (LINear ACcelerator) or Tandem beamline between the source and magnet analysis, together with endstation sections borrowed from high-current system designs (Fig. 10-31). High-energy implantation was first used for the formation of buried grids as protection against "soft errors" in memories devices[47] and for the direct formation of buried collectors in bipolar transistors.[48] High-energy implanters are now widely used for the formation of retrograde wells and buried layers for control of latchup in CMOS devices,[49] as well as for special applications for ROM-programming and CCD photo-sensor devices. A key advantage of MeV implantation is that most of the ion damage is confined within the ion range distribution, which is buried deep to a region near the bottom of the CMOS wells, so that near-surface regions (such as the transistor channel), are relatively undisturbed.

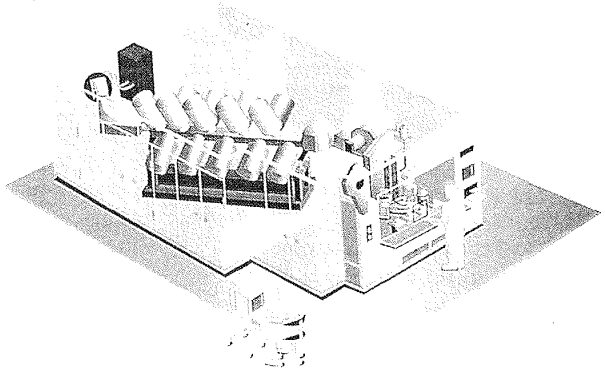


Fig. 10-31 Beamline for an Eaton GSD-VHE,® a high-energy implanter using a LINAC (*LINear ACcelerator*) beamline. Courtesy of Eaton Corp.

10.4.2.5 High-Angle Implanters: *High-angle implanters* are a variation on the standard machine types described above, in which the beam incidence angle on the wafer can be varied from the usual 7° or 0°. For several batch machines the implant angle can be varied from 0° up to 7–10°. In the Eaton GSD (Gyroscopic Super Disk) machines, the beam angle can be varied by tilting the wheel up to ±10° along both the horizontal and vertical directions achieving a beam strike from four separate directions (known as a "quad" implant). Many modern single-wafer machines have wafer tilt and rotation fixtures that allow for implantation angles up to 60 degrees from the wafer normal and full rotation of the wafer in the ion beam. These high-angle implanters are used to form regions under the edges of the gate mask (e.g., source/drain extensions called *LATID* [*Large-Angle-TIlt Drain*], and punchthrough stoppers called *LATIPS* [*Large-Angle-TIlt Punchthrough Stopper*]).

10.4.3 Ion Implantation Equipment System Limitations

The level of precision and variety of process conditions used in ion implantation leads to a very high level of expectation that the process will be performed exactly as designed. This is not always the case and a number of process non-uniformities and defects that can limit ID device yield.[50,51] In the previous sections the impact of maximum beam current and vacuum pumping issues on wafer throughput, and planar channeling and wafer heating as process limitations associated with implantation equipment was discussed. Some other limitations of the equipment will be described here, including: a) elemental and particulate contamination; b) dose monitoring inaccuracies due to beam charge state change effects; c) implantation mask issues; d) wafer charging during implantation; e) poor dose matching from machine-to-machine in production implanters; f) tilt-angle and "scan lock-up" non-uniformities in electrostatically scanned machines; and g) dose variations on the scale of die sizes or "micro-uniformity."

Table 10.2 ION IMPLANTATION SYSTEM TYPES

Years	Era	Machine Type	Beam Energy (keV)	Beam Current (mA)	Key Application	Tool Examples
1970-1985	LSI	Medium Current	20–200	< 1	Threshold Voltage Junction Isolation	Varian DF-4
		"Pre-Dep" High Current	20–80	1–10	S/D Contact	Eaton NV-10-80/160 Varian 120-10
1985-1995	VLSI	Medium Current/ High Tilt (40-60°)	20–200	< 1	Gate-Overlap LDD (GOLDD)	Varian E-220
		High Current	20–200	0.01–25	Source/Drain; Channel; Poly	Applied P!900/9200 Eaton NV-20A Varian E-1000
1995-2005	ULSI	High-Energy	400–3000	0.02–1	CMOS Wells	Eaton GSD-HE
		Low-Energy	0.2–80	0.01–25	S/D Extensions/ Contacts	Applied xR LEAP Eaton ULE
		High-Current/ High Tilt (40-60°)	0.5–80	0.01–10	FLASH S/D	Varian VIISta-810

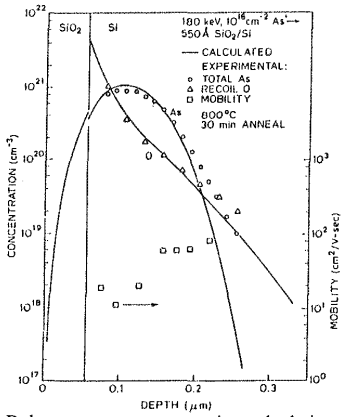


Fig. 10-38 Comparison of a Boltzmann transport equation calculation and experimental SIMS results for the recoil oxygen distribution resulting from an implantation of As into 55 nm of SiO_2 from a dose of 10^{16} cm^{-2}. Reprinted with permission of the American Physical Society.[19]

and

$$\frac{R^2}{\sigma_p^2} = \left[\sum_i \gamma_i \frac{C_i}{R_i} \frac{R_i^2}{\sigma_{pi}^2}\right]\left[\sum_i \gamma_i \frac{C_i}{R_i}\right]^{-1} \qquad \textbf{(10.20)}$$

where

$$\gamma_i = 4\, M_1\, M_i/(M_1 + M_i)^2 \qquad \textbf{(10-21)}$$

and C_i represents the fraction of the total mass of the i^{th} component, and R_i and σ_i, the range and projected range of the i^{th} component in mass units per cm^2. Monte Carlo methods should be used when higher accuracy is required.

10.6.3 Threshold-Voltage Control in MOS Devices

One of the most important applications of ion implantation is the *control of the threshold voltage of MOS devices*. When boron atoms are implanted into the channel region below the gate oxide of MOS devices, the threshold voltage V_T changes by $\Delta V_T = -\Delta Q_B/C_{ox}$, where ΔQ_B is the change of the sheet ionized dopant charge in the channel (the ion dose), and C_{ox} is the gate oxide capacitance per unit area. To achieve adequate control over the transistor threshold voltage, the ion dose and energy need to be controlled to within a few percent at a dose of $\approx 10^{12}$ ions/cm^2 (about 0.1% of a monolayer of atoms). The ability of ion implantation to allow this level of control was a critical factor in its initial introduction into IC fabrication process.

As channel regions are scaled to lengths of 0.1 μm (100 nm), a new problem occurs because the total number of dopant atoms in the channel becomes a countable and statistically fluctuating quantity. For a dose of 10^{12} ions/cm^2, there are approximately 100 atoms implanted into a channel region that is 100x100 nm^2 (0.1x 0.1 μm) in area. This means although the average dose uniformity can be controlled to less than 1% over a wafer (and over a large number of lots), the variation in the number of dopant atoms in individual, adjacent channel regions will be $\approx \pm 10\%$ ($[100]^{1/2}/100$). This variation leads to a large spread in the threshold voltage over a device and to large sub-threshold leakage currents.

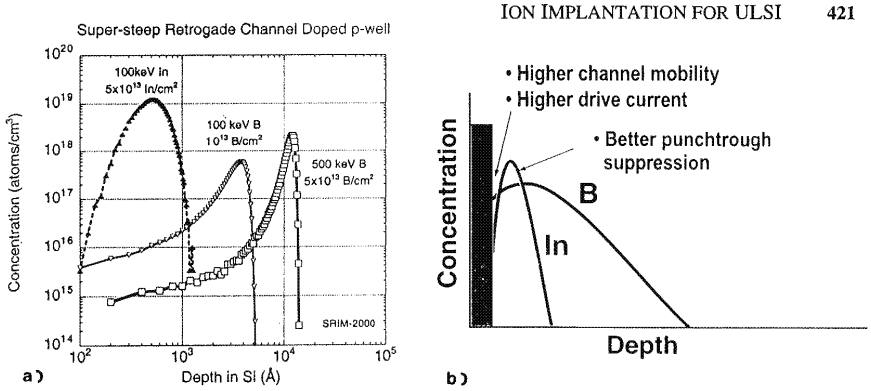


Fig. 10-39 (a) Calculated profiles for a "super-steep retrograde doped" channel implant with 100 keV at a dose of $5x10^{13}$ indium ions/cm^2 in a multi-implanted p-well background.[91] © IEEE (1995). (b) Schematic comparison of channel profiles using In and B. The super-steep In retrograde profile improves punchthrough suppression, and electron mobility is increased by lowering interface dopant levels.[104]

An alternative approach to threshold control is to leave the channel region just below the gate oxide relatively undoped and to confine the source to drain current flow by the use of a rapidly rising well doping profile, a "super-steep, retrograde channel doping" implant.[91] Since the channel implants are done relatively early in the fabrication process (i.e., before the gate oxidation, gate formation and source/drain doping steps), it is important to use dopants which do not diffuse far from their as-implanted profiles during the later thermal processes. For this reason, indium (for p-wells) and antimony (for n-wells) are the leading candidates for this process. An example of the implant profiles for an In-implanted channel is shown in Fig. 10-39.

10.6.4. Shallow Junction Formation by Ion Implantation

As transistor lateral dimensions are reduced to achieve higher speed performance and higher density of functional components on a die (most notably by shrinking the size of the MOSFET gate), the junction depths and doping profiles are also restricted to shallower locations (i.e., closer to the contact plane at the base of the interconnect network). There are two main regions where this "junction scaling" is important: 1) the "channel region" which encompasses the contacts and doped areas around the CMOS source and drain (including the channel under the gate); and 2) the "well" regions (which determine the background doping for the transistor, see Fig. 10-40). The channel region is extremely shallow, with channel depths for MOSFETs with 100 nm gate sizes in the range of ≈30 nm. The well doping extends to much greater depths (typically to ≈1 μm), and well formation is discussed in the next section. As the junctions are scaled to shallower depths, the ion energies used to implant dopants into both the channel and well regions decrease in rough proportion to the gate size (see Fig. 10-41).[92]

To obtain the shallow junctions required for source/drain extensions and channels, low energy ions must be used (e.g., below 1 keV in the case of boron implantation). Furthermore, all other physical mechanisms which allow dopants to penetrate deeper into Si than the as-implanted profile must be strictly controlled or eliminated. The principal factors to be controlled are ion channeling and dopant diffusion. To accomplish these ends, shallow-junction doping pro-

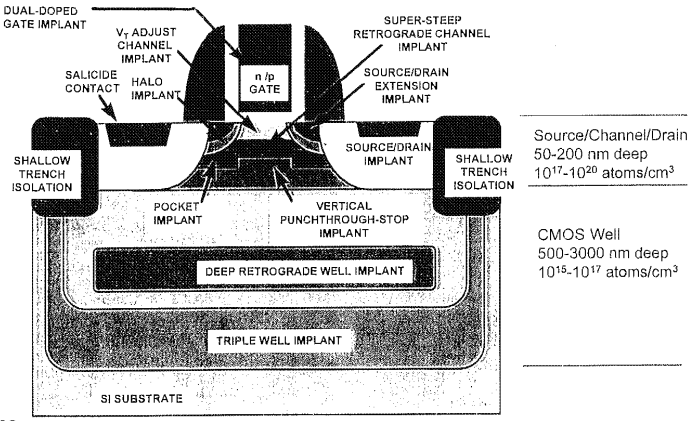


Fig. 10-40 CMOS transistor doping regions.

cesses often use combinations of: 1) high-dose, "pre-amorphization" implantation steps before the actual dopant implantation; and 2) rapid thermal annealing cycles. For example, a 5-keV pre-amorphization Ge implant (at a dose of 10^{15} Ge/cm^2) may be used to control channeling. This is combined with a short RTA step (≈1 sec) at a temperature of ≈1050°C (to anneal the lattice damage created by the implant and to electrically activate the dopant atoms, without allowing appreciable diffusion of the implanted species). In addition to the peak anneal temperature, such factors as the temperature ramp rates and gas ambient (trace levels of oxygen, for example) are important process-control factors.

The source extensions are particularly critical junctions. In order to suppress the roll-off of the threshold voltage as the gate size is decreased, the junction depth of the extensions must also decrease proportionately (i.e., 30–50% as much as the gate length is scaled). To minimize the series resistance from the source contact to the drain, the sheet resistance of the extension region must also be minimized. This is accomplished by achieving a high percentage of dopant activation while still limiting the dopant diffusion (which must be controlled to maintain a shallow junction depth). In addition, the abruptness (change in doping concentration) of the

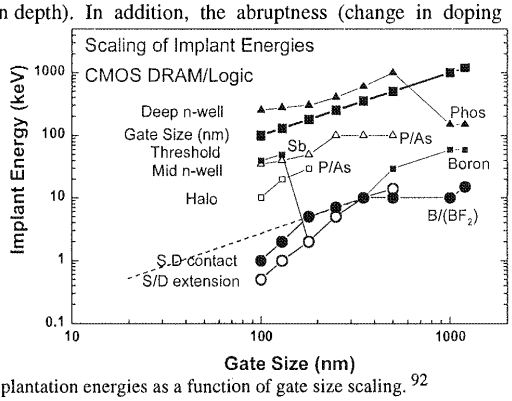


Fig. 10-41 Ion implantation energies as a function of gate size scaling. [92]

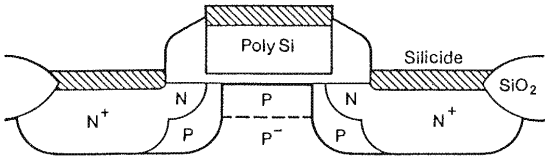


Figure 1. DI-LDD device cross section, vertical doping profile through the LDD region, and lateral doping profile at the surface.

Fig. 10-42 Cross-section of an NMOSFET fabricated with a lightly-doped drain (LDD) and a *p*-halo implant beneath the *n*-drain extensions.[86]

extension boundary into the channel region must also be kept as sharp as possible to minimize resistance effects.

One method for simultaneously increasing the abruptness of the extension junction doping profile, decreasing the extension junction depth, and increasing the punchthrough voltage is to implement a *halo implantation.* In a halo implantation, dopant of the same type as the major well dopant is implanted beneath the drain extension junction (Fig. 10-42).[86] The "halo" is used because the implantation is done after the formation of the gate, and so the implanted dopant is self-aligned to the gate edge and surrounds the gate with a ring of dopant. For *n*-well transistors, arsenic is often used for the halo implants, while P is used to dope the deeper portions of the well. Arsenic is used for the halo doping since it has a lower diffusion rate and the halo profile can be better controlled during the source/drain activation cycle.

The source/drain contact doping must be deep enough to allow for the formation of low contact resistance interfaces between the junction and the interconnect metal network. The contact junctions must also have a low enough resistance that the current flow from the source contact to the channel region (and the mirror current path at the drain) spreads out evenly though the junction and does not concentrate at the inside edges of the contact pad. If the sheet resistance of the contact junctions is too high, heating of the transistor through "current crowding" effects can not only limit the transistor drive current, but also in the worst case cause circuit failure by contact electromigration effects (see Vol. 2, Sect. 4.7.3.2 of this series).

The contact junctions must also have a rapid change in doping concentration at the location of the bottom of the junction to obtain the lowest sheet resistance and form the shallowest junctions that still have the low-leakage currents of *pn* diodes (see Fig. 10-43).[92]

10.6.5 High-Energy Implantation

Even as junctions of the source/drain and channel region call for ion energies of a few keV or less, the doping of CMOS well junctions requires ion energies ranging from a few hundred keV to several MeV[93] (see Fig. 10-41). The particular ion energy used depends on the application and device type. For example, high-speed logic (microprocessors) and DRAM devices use relatively shallow well structures while SRAM and FLASH memories use deep wells, sometimes in combinations with even deeper "triple wells" (which allow for the *n*- and *p*-well junctions to be biased separately from the substrate).

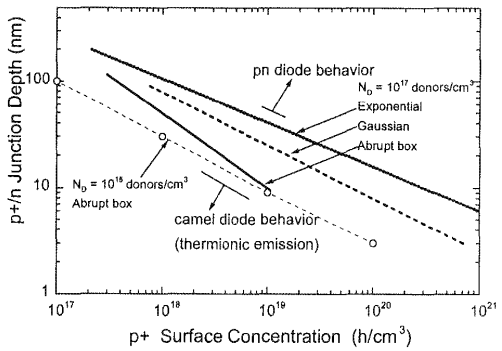


Fig. 10-43 The boundary between *pn* diode behavior and high-leakage currents from thermionic emission of metal contacts through various contact dopant profiles. Note that much shallower junction with *pn* diode behavior can be formed with abrupt (or "box profile") junctions than the more gradual Gaussian or exponential dopant profiles.[92]

Modern CMOS processes use directly implanted profiles to set doping levels in the wells. The shape of a high-energy ion-implantation profile (which rises from a low concentration at the wafer surface to a peak deep within the Si), is the basis of the name "retrograde" well (see Fig. 16-23). This doping profile contrasts with the doping profile of earlier conventional CMOS processes, in which the well dopants were diffused more deeply into the substrate from their near-surface implanted locations. Such a drive-in diffusion following a near-surface well-implant step results in a maximum concentration at the wafer surface and a monotonic decrease in doping concentration as the well junction is approached. The directly implanted, retrograde doping profile provides much higher levels of protection from accidental well-to-well "latch-up" events in CMOS than is possible with diffused wells (see Vol. 2, Sect. 6.4 of this series for a detailed discussion of latchup in CMOS). In addition, in DRAM devices the risk of bit state loss (also called "soft errors" - see Vol. 2, Sect. 8.3.5 of this series for a description of soft errors in DRAMS) due to transient charges arising from ionizing radiation, is significantly reduced. The directly implanted processes also allow for separate doping of the channel region (for threshold voltage control) and the region beneath the channel (anti-punchthrough implant - for increasing the punch-through voltage - see Fig. 10-39). Details of punchthrough effects in MOSFETs are given in Vol. 3, Chap. 5 of this series. In addition to these device advantages, the use of high-energy implantation to replace the long-cycle well diffusion provides significant process time and wafer stress reduction.

High-energy implantation is also used to program ROM cells by changing the threshold voltage through direct implantation into the channel region. The implantation is done through the gate stack, after the source/drain and gate formation process have been completed. Programming ROMS late in the process cycle not only reduces product cycle time to completion but also provides an added degree of security, since the ROM code cannot be detected in the final devices by standard etching and imaging techniques. This method of programming ROM cells is used to produce high-security, chip specific, "scramblers" for data links between distributed microprocessors and data bases.

Chapter 16

CMOS PROCESS INTEGRATION

In the first fourteen chapters of this book the focus was on the individual process modules employed in IC fabrication. This chapter will consider how these modules can be combined to define a set of manufacturing steps that becomes a recipe for fabricating monolithic silicon integrated circuits. The task of specifying such a sequence of individual process steps is called *process integration. Complimentary MOS* (or *CMOS*) technology is used to demonstrate some examples of such *process sequences* (or *process flows*). CMOS was chosen for our examples because it is the dominant silicon IC technology of the 1990's. Figure 16-1 shows that about 80 percent of all semiconductor devices and circuits manufactured in 1995 were CMOS ICs. By the end of the 1990s this percentage grew even larger. Since this text is intended to provide an introduction to *mainstream* silicon IC process technology, it is appropriate to use CMOS as the vehicle for illustrating process flows. Since CMOS technology was preceded by other MOS technologies, their evolution will also be briefly reviewed. How and why CMOS became the dominant semiconductor technology will be presented. Note that CMOS is so-named because it uses both *p*- and *n*-type MOS transistors in its circuits. Figure 16-2 depicts a CMOS inverter.

In the examples of CMOS process integration two CMOS process flows will be described. The first is a generic process flow for the generations of CMOS logic ICs extending from 1.2–0.5 μm. The second pertains to CMOS ICs having minimum feature sizes of 0.25 μm and smaller. For such *deep-submicron CMOS technologies* (i.e., $\leq$ 0.25 μm), several major changes occurred in the process flow and device architecture. Hence, the second process flow differs significantly from the first. Some of new processes used are: high-energy implanted retrograde wells, shallow-trench isolation, dual-doped polySi, $TiSi_2$ and $CoSi_2$ salicides, chemical-mechanical polishing, and formation of interconnects with damascene processing.

The purpose of this chapter is to present examples of how Si processes are integrated, resulting in the creation of fabrication sequences for CMOS ICs. A discussion of process integration involves the relationships between process technology and device physics. However, the operation of MOS devices is beyond the scope of this text, and is not presented here. Readers interested in more information on device physics should consult such texts as those by Chen,[1] Muller and Kamins,[2] and Tsividis[3] (as well as Vol. 3 of this series).

16.1 INTRODUCTION TO CMOS TECHNOLOGY

Integrated circuits built with MOS transistors were introduced in the early 1960s, and these were fabricated using PMOS technology. However, it was known that since electrons have higher mobility than holes, ICs fabricated with NMOS would have better performance than those made with PMOS. By the 1970s, the technological barriers that blocked NMOS fabrication were overcome (see Vol. 2, Chap. 5 for details), and NMOS became the dominant technology for MOS IC fabrication. NMOS remained the main MOS IC technology until the integration level of devices on a chip became too large (i.e., the late 1980's), when CMOS took over.

NMOS is inexpensive to fabricate, very functionally dense and faster than PMOS. The earliest NMOS technologies required only 5 masking steps, including the pad mask.* On the other hand, NMOS logic-gates (e.g., inverters) consume dc power while in one of their two logic states. Therefore, at least some of the logic gates in an NMOS IC are always ON. Such circuits draw a steady current, even when being operated in the standby mode (when no signals are propagating through the circuit). Hence, as the number of logic gates on NMOS chips increased, the current being drawn (and hence the power dissipated) also rose. Although this was always a limitation of NMOS for such applications as space-borne or portable electronic systems, it did not represent a major drawback for most others. Such was the situation at the level of device integration that existed up to the mid-1970s. By the mid-1980's, with the dawning of the VLSI era (i.e., device counts on chips began to exceed 10^5), power consumption in NMOS circuits began to surpass tolerable limits. A lower-power technology was needed to exploit the VLSI fabrication techniques. CMOS represented just such a technology.[1]

From a quantitative perspective, the ascendancy of CMOS was the inevitable result of the two-hundred-fold increase in functional density and the twenty-fold increase in speed of ICs between 1968 and 1987. For example, in 1969, 256-bit SRAM circuits (e.g., the Intel 1101) used 12-μm design rules to create six-transistor SRAM cells; each cell occupied 20,600 μm^2. By 1998, SRAM cells using 0.25-μm design rules occupied less than 10 μm^2. (The memory access time in these respective memory circuits decreased from 1 μs down to 10 ns.) By 1998 the decreased device dimensions and the attendant increase in chip size allowed 16-Mbit SRAMs to be built. To take another example: The Intel 4004 4-bit microprocessor (introduced in 1971), had 2300 devices and was built with PMOS. The 8086 model, a 16-bit microprocessor introduced in 1978, had 29,000 devices and was built in NMOS, and the 80386, introduced in 1985, had 275,000 devices and was built in CMOS.

Chips dissipating in excess of 5 W of power can be housed in ceramic packages. But such packages are expensive. To be able to use the much cheaper plastic packages, the maximum power dissipation is limited to about 1 W. The Intel 8086 dissipated around 1.5 W of power

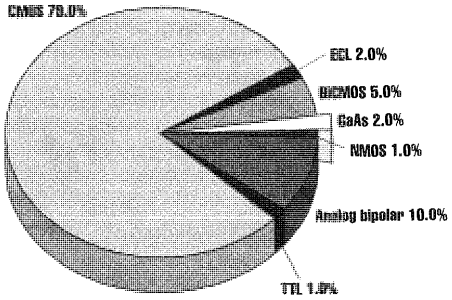


Fig. 16-1 Estimated market share of IC technologies in 1995. Reprinted with permission of Oxford Press.

* Early NMOS logic gates used butted contacts as well as enhancement-mode (E) transistors for both the driver and load. Depletion-mode loads (D) and buried contacts came later. Hence, of the seven masks of the E-D NMOS process, only five were required in early NMOS technology. This resulted not in lower manufacturing costs per wafer and higher yields, and thus ultimately cheaper NMOS circuits.

Fig. 16-2 CMOS inverter. (a) Circuit schematic. (b) Device cross section.

when operated at 8 MHz. By the late 1970s it was clear that chips would very soon be manufactured with power dissipation values that were unacceptably high. (Note that when the 8086 was later reissued in CMOS technology under the model number 80C86, its power consumption dropped to about 0.25 W.)

Unlike in an NMOS inverter, in a CMOS inverter (Fig. 16-2) only one of the two transistors is driven at any one time (if the inverter is not in a transient switching state). This means that a high-impedance path exists between V_{DD} and ground, regardless of the state the inverter is in. Hence, virtually no current flows between the power and ground lines, and almost no dc power is dissipated by the CMOS circuit. Thus, logic gates consuming only microwatts of standby power can be implemented with CMOS (Fig. 16-3).

The problem of power dissipation can also be considered from both a *chip* perspective and a *system* perspective. From the chip perspective, microprocessors of the 64-bit generation are built using CMOS and yet some still dissipate more than 15 W of power. If only NMOS technology was available, the power dissipation would probably be an order of magnitude larger! Now consider memory chips from the system perspective. Although a 1-Mbit DRAM may consume 120 mW of power in NMOS, it consumes even less (~50 mW) in CMOS. Since there may be hundreds of memory chips in a system (versus only a few microprocessors), the ramifications of lower power-dissipation at the system level are significant. Smaller power-supplies and smaller cooling fans are but two of these.

16.1.1 Historical Evolution of CMOS

Although CMOS is now the dominant integrated-circuit technology, during its first 20-years it was considered to be a runner-up for the design of MOS ICs. The pairing of complementary *n*- and *p*-channel transistors to form low-power IC logic gates was originally proposed by

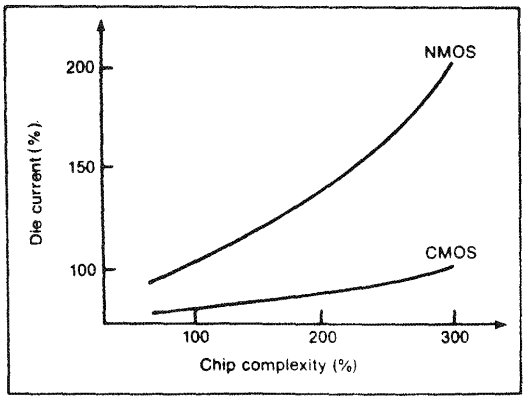


Fig. 16-3 As circuit complexity increases, NMOS power consumption rises to levels that eventually prevent further growth. In contrast, CMOS power consumption increases only slightly as the device count on a chip rises.

by Sah and Wanlass in 1963.[4] The first CMOS ICs were fabricated in 1966. Subsequent development of the technology was spearheaded by the RCA Corporation. The earliest volume commercial application of CMOS was its use as a logic-inverter in the frequency divider circuits of digital watches.

CMOS technology at that time had many disadvantages compared even to PMOS and, later, to NMOS. Drawbacks included significantly higher fabrication cost, slower speed, susceptibility to latchup, and much lower packing density. As a result, until the 1980s CMOS was limited to applications that mandated either the lower power dissipation of CMOS (e.g., watches and calculators), or it's very-high noise margin (e.g., radiation-hard circuits). Furthermore, the advances made in NMOS fabrication were not rapidly transferred to CMOS. For many years CMOS languished behind advanced Si-gate-NMOS and bipolar technologies. Except for the special applications mentioned above, it lay dormant for nearly two decades.

There were several reasons why this occurred. Recall that both *n*- and *p*-channel transistors must be fabricated on the same wafer in CMOS technologies. Obviously, only one type of device can be fabricated on a given starting substrate. To accommodate the device type that cannot be built on this substrate, regions of a doping type opposite to that present in the starting material must be formed. These regions of opposite doping, called *wells* (or sometimes *tubs*) are the first features to be defined on a starting wafer. In the earliest CMOS technologies, a single-well approach was used. A lightly doped *n*-type wafer was the starting material, and only *p*-wells were formed in the wafer. This was called *p*-well CMOS. At the time these circuits were being fabricated (late 1960s), the processes of ion implantation and local oxide isolation (LOCOS) had not yet been developed. Metal gates were still being used in MOS devices and control had not been gained over the large and quite variable positive oxide charges in the gate oxide. These issues made *p*-well technology the only viable technique for fabricating CMOS.

While PMOS enhancement-mode transistors could be successfully fabricated in a lightly doped (e.g., 10^{15} cm^{-3}) *n*-substrate with a V_T of about -2 V, NMOS enhancement-mode transistors could not be fabricated on lightly doped *p*-substrates because the V_T of NMOS metal-gate transistors on such substrates is negative. In addition, it was likely that parasitic channels would be established in the field regions between NMOS devices built on lightly doped substrates. These parasitic channels could not be reliably suppressed because of three other factors: 1) the oxide charge at the silicon/SiO_2 interface was large, and procedures for lowering it had not yet been discovered; 2) the segregation of boron at the field-oxide/silicon-substrate interface reduced the V_T of parasitic MOSFETs; and 3) the need to use relatively thin field oxides (to permit adequate step coverage of metal layers running over these oxide steps) also reduced the parasitic MOSFET V_T (see Vol. 2). Therefore, the only reliable way to manufacture enhancement-mode NMOS transistors for CMOS inverters without excess leakage was on regions with boron surface concentrations high enough to overcome these problems.

The *p*-well approach to building CMOS provided such regions, since the *p*-well had to be doped about ten times as heavily as the *n*-substrate for adequate control of doping in the well to be achieved. As a result, *p*-well technology became established in the companies that pioneered CMOS technology. Long after the problems of fabricating NMOS on lightly doped substrates had been solved (by controlling V_T with ion implantation, and the interface charge by annealing), most of these companies continued to use *p*-well technology to design new ICs.

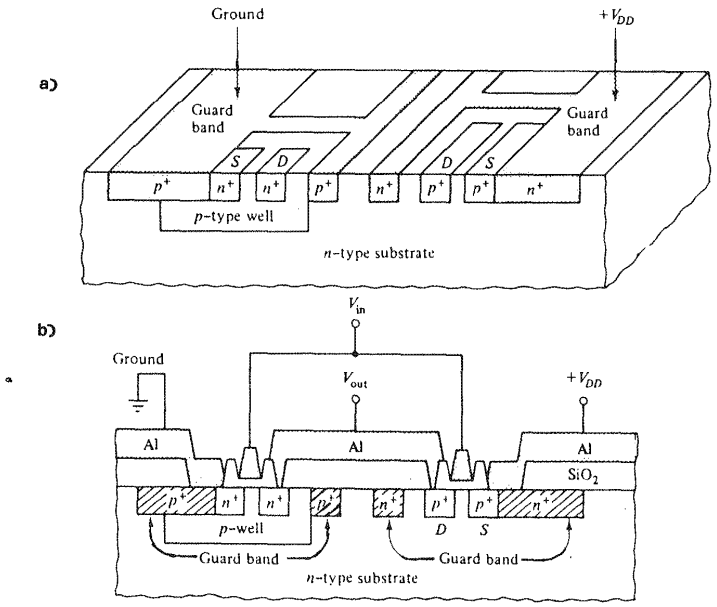


Fig. 16-4 CMOS structure with guard bands. A CMOS inverter circuit is depicted.[5] From W. C. Till and J. T. Luxton, *Integrated Circuits: Materials, Devices, and Fabrication.* Copyright 1982 Prentice-Hall. Reprinted with permission.

The packing-density limits of the early p-well CMOS technology, however, were responsible for its poor performance. The packing density was much worse than that of NMOS, primarily because the n- and p-devices had to be surrounded by guard rings (n^+ or p^+ diffusions that surround the device, as shown in Fig. 16-3)[5] to prevent the inversion of the field regions and latchup (see Vol. 2).

Before LOCOS isolation and ion implantation became available, guard rings were needed to provide adequately large values of V_T in the field regions. However, their use results in very large-area devices. Until oxide isolation and channel-stop implants were developed, interconnect capacitance and resistance severely degraded the speed performance.* Once guard rings were no longer needed, however, the devices could be brought closer together, and the speed of CMOS circuits improved dramatically.

When it finally became apparent in the late 1970s that the increases in power density and dissipation make it impossible to design future generations of circuits with NMOS,

* It should be noted that the fabrication of NMOS devices in heavily doped p-wells also increased the device junction-capacitances and reduced the magnitude of the drive current. This further decreased the performance of the circuits, but to a lesser degree than did the interconnect capacitance.

companies that had been stubbornly continuing to use NMOS for design and fabrication finally began to consider CMOS. It was natural for them to seek a technology that was compatible with the modern high-production-volume Si-gate-NMOS processes that they had successfully perfected. Since *n*-well CMOS (the other single-well CMOS type) offers near compatibility with such processes, and since it allows NMOS performance to be optimized (through their fabrication in the lightly doped *p*-substrate regions), it became the technology of choice for many companies that had formerly been manufacturing NMOS integrated circuits.[6]

Some time later, however, it became evident that neither *p*-well nor *n*-well would be the optimum choice for submicron CMOS. It was realized that *twin-well CMOS* would be more effective. As a result, many twin-well processes were subsequently developed. Examples of such twin-well CMOS process flows in Sects. 16.2 and 16.3 will be considered.

16.1.2 The Operation of CMOS Inverters

The CMOS inverter (for which the circuit schematic and a sample layout are shown in Fig. 16-2a and 16-5, respectively) uses enhancement-mode transistors for both the NMOS driver and the PMOS load transistors. (In enhancement-mode transistors the devices are *OFF* if the gate [or input] voltage is zero.) It can be seen that the gates of the two transistors in a CMOS inverter are connected, and serve as the input to the inverter. The common drains of each device are also connected to the inverter output. Assume the inverter is driving some load capacitance, C_L,

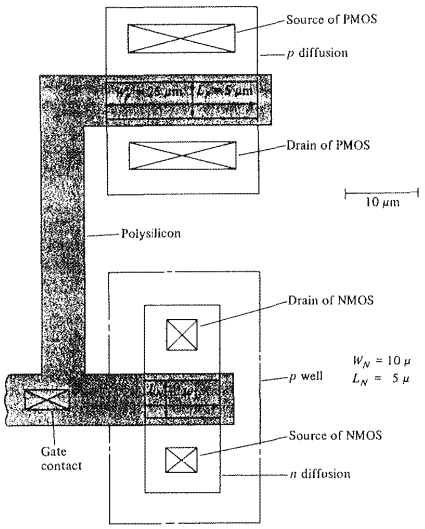


Fig. 16-5 CMOS inverter layout. From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright 1988, 2nd Ed., McGraw-Hill Book Co. Reprinted with permission.

(e.g., the input to another CMOS logic gate). Both the source and body of the NMOS transistor are connected to ground, while those of the PMOS transistor are connected to V_{DD} (e.g., 5 V).

The threshold voltage of the NMOS driver transistor V_{Tn} is positive (e.g., V_{Tn} = 0.6 V), while that of the PMOS load transistor V_{Tp} is negative (e.g., V_{Tp} = -0.6 V). Figure 16-6a shows the inverter's output voltage V_o as a function of the input voltage V_{in}. This curve is known as the *static voltage input/output characteristic*, or the *transfer characteristic*.

When V_{in} = 0, V_{GSn} = 0, the NMOS transistor is *OFF* (since V_{GSn} is <0.6 V). The PMOS transistor, however, is *ON*, since V_{GSp} = -5 V, which is much more negative than -0.6 V. Thus, when V_{in} = 0, C_L is charged to V_{DD} through the turned-on PMOS load transistor, and V_o = 5 V.

When V_{in} = 5 V, the NMOS transistor is turned *ON*, since V_{GSn} = 5 V (which is > 0.6 V). The PMOS transistor is turned *OFF*, since V_{GSp} = 0 V (which is more positive than -0.6 V). Consequently, when V_{in} = 5 V (V_{DD}), the output is connected to ground through the turned-on NMOS driver, allowing C_L to be discharged. Since the PMOS device is off, C_L will be completely discharged, and V_o will be 0 V.

The most important property of the CMOS inverter is that when the gate is sitting quiescently in either logic state (V_o = V_{DD} or 0), one of the transistors is *OFF* and the current conducted between V_{DD} and ground is negligible (i.e., it is equal to the leakage current of the *OFF* device). This feature can be seen in Fig. 16-6b, which plots the current through the inverter, I_{DD}, as a function of V_{in} (solid curve). The power dissipated in the static (or standby) mode is then determined by the product of the leakage current and the supply voltage. Since the leakage current of an MOS transistor being operated in cutoff is so small, very little power is consumed in the static mode. Another important feature is that the output voltage V_o swings all the way from V_{DD} to 0 as the inverter changes state. This switching characteristic is referred to as *swinging from rail to rail.*

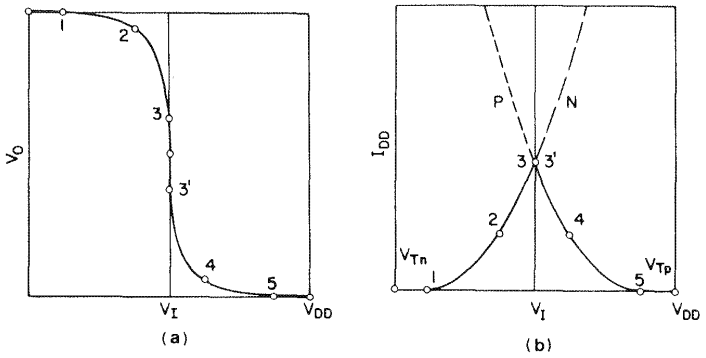


Fig. 16-6 (a) Output (V_o) versus input (V_I) voltage of CMOS inverter. (b) Current through inverter as a function of input voltage (solid curve); I-V characteristics of *n*– and *p*–channel transistors (dashed curves). The numbers correspond to different points on the inverter transfer characteristic.[8]

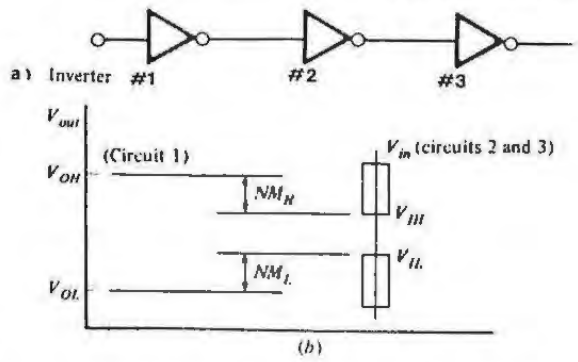


Fig. 16-7 (a) String of inverters connected in series. (b) Definition of noise margins.

Two operational features of the CMOS inverter must also be mentioned. First, as can be seen in Fig. 16-6b, significant current is conducted through the inverter only when both transistors are *ON* at the same time (i.e., during switching). Therefore, most of the power dissipation is due to the charging and discharging of C_L. In fact, it can be shown that the power dissipation is essentially given by $f C_L V_{DD}^2$, where f is the switching frequency. Because only a small fraction of the logic gates in an IC are likely to switch during any given clock cycle, CMOS power dissipation remains much smaller than in NMOS ICs.

Second, because the output voltage swings from rail to rail, the CMOS inverter inherently provides excellent noise margins. Noise margins are usually defined in terms of the *logic-gate threshold voltages* V_{OH}, V_{OL}, V_{IH}, and V_{IL} (which are *not* the same parameters as *device threshold-voltages*, see Fig. 16-7a). The noise margins NM_L and NM_H (Fig. 16-7b) are defined according to the following equations:

$$NM_L = V_{IL} - V_{OL} \quad \textbf{(16.1a)}$$

$$NM_H = V_{OH} - V_{IH} \quad \textbf{(16.1b)}$$

Briefly, the argument for Eq. 16-1a is that the logic gate (e.g., *Inverter 1* in Fig. 16-7a) should provide a maximum "low output" that is less than the maximum "low input" that the subsequent logic gate (e.g., *Inverter 2*) can accept without causing the output voltage of *Inverter 2* to be driven into the ambiguous portion of the transfer characteristic. The difference between V_{IL} and V_{OL} (the noise margin) then specifies how much noise voltage can be tolerated at the input node of *Inverter 2* before its output will be driven to a voltage value that is logically undefined.

In CMOS, V_{OL} approaches zero, and V_{OH} approaches V_{DD}. Because of the steep transition region in the transfer characteristic, V_{IL} and V_{IH} can be designed so that noise margins are on the order of $V_{DD}/4$ for expected process variations and operating temperatures. Thus, for a 5-V CMOS technology, V_{NML} can be 1.25 V, whereas in NMOS, V_{NML} is typically only 0.3 V.

For the performance of the logic gates to be maximized, the threshold voltages of the *p*- and *n*-channel transistors should be comparable (and, ideally, of equal magnitude). In addition, the threshold voltages should be as small as possible. The minimum values selected for V_{Tn} and V_{Tp} will be determined by the subthreshold leakage current, I_{Dst}.

16.2 PROCESS SEQUENCES FOR TWIN-WELL CMOS FOR THE GENERATIONS FROM 1.2 μm TO 0.5 μm

This section describes an 11 mask, twin-well CMOS process flow that is representative of process sequences used to fabricate CMOS ICs from 1.2 μm down to about 0.5 μm. The process flow is nevertheless useful as a vehicle for illustrating the general sequence of steps employed in CMOS fabrication. Twin-well CMOS technology has been generally adopted for CMOS generations below about 1 μm because this approach allows two separate wells to be formed in a lightly doped substrate region. The doping profiles in each well can thereby be tailored independently so that neither device will suffer from excessive doping effects.

16.2.1 Starting Material: All MOS technologies use silicon wafers with a <100>-orientation. Typically, in twin-well CMOS the starting material is also either a lightly p-doped wafer (p-bulk wafer, Fig. 16-7a), or a heavily p-doped wafer on which a thin, lightly p–doped epi layer is grown (Fig. 16-7b). The concentration range of the p doping in both the bulk wafer and the p–epi layer is $8x10^{14}$–$1.2x10^{15}$ cm^{-3}, while the p^+ substrate beneath the epi layer is doped to ~10^{20} cm^{-3}. The p-epi-on-p^+ wafers (see Chap. 7) provide several advantages including: 1) improved latchup protection (see Vol. 2); 2) gate-oxides with better reliability are grown on epi layers than on bulk-Si surfaces; and 3) improved gettering capability (see Chap. 2). Their chief disadvantage is higher cost. (In 1998, the price of a 200-mm bulk starting wafer was ~$100 US, while that of a 200-mm epi wafer was ~$180 US.) For this example process flow, assume p-epi-on-p^+ wafers are used (Fig. 16-8b). The exact doping concentration of the epi layer is chosen to provide the best overall set of the following device characteristics: low source/drain-to-substrate capacitance; high source/drain-to-substrate break-down voltage; high carrier mobility; and low sensitivity to source-substrate bias effects. As feature sizes shrink, the epi-layer doping concentration value is expected to increase.

16.2.2 Formation of Wells and Channel Stops: The wells are the first features to be formed in these generations of CMOS technology. A number of different procedures have been developed to form twin-wells. The most obvious method is to use two masking steps. Each blocks one of the two implant steps used to introduce the well dopants. A single masking-step procedure, however, was also developed. It is probably the one most commonly used, and it will be described here (see Figs. 16-8 and 16-9).

In this method a single mask is used to pattern a nitride/oxide film that has been formed on the bare silicon surface (*Mask #1*). A *pad oxide* (which serves as stress-relief layer between the silicon nitride film and the silicon substrate) is first thermally grown to a thickness of 40–50 nm

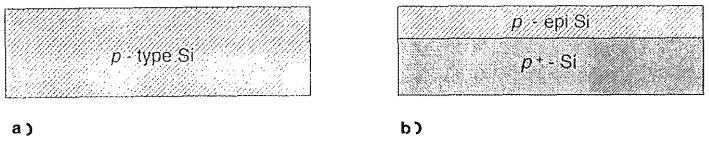


Fig. 16-7 Silicon starting substrates for twin-well CMOS: (a) p-bulk wafer; (b) p-epi-on-p^+ wafer.

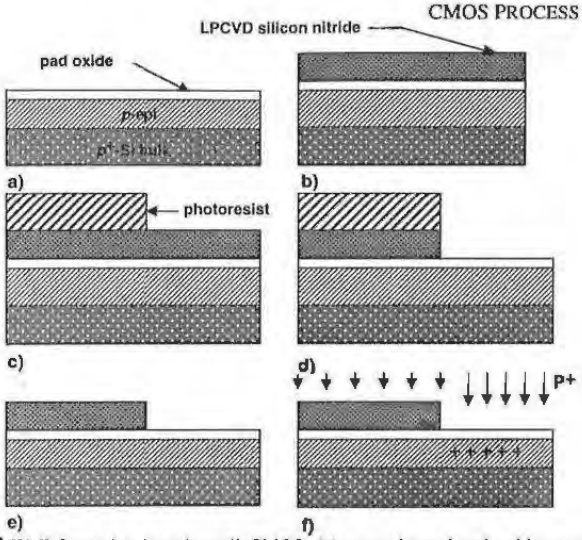


Fig. 16-8 Well formation in twin-well CMOS. (a) grow thermal pad oxide on p-epi-on-p^+-bulk starting wafer; (b) deposit LPCVD silicon-nitride film; (c) use *Mask #1* to define n-well regions; (d) etch nitride; (e) strip resist (f) implant n-well dopants (phosphorus).

(Fig. 16-8a). This is followed by an LPCVD of silicon nitride (140–200 nm thick- Fig. 16-8b). After the photoresist is applied, exposed and developed (Fig. 16-8c), the nitride is dry etched (Fig. 16-8d). The openings in the nitride film define the n-well regions. The required n-dopants are next introduced into these regions by ion implanting phosphorus (Fig. 16-8f). This implant is performed at 80 keV with a dose of ~1×10^{12} cm^{-2}, using a medium-current implanter. The resist and unetched nitride regions (which define the p-well regions) act as an implant mask, preventing the phosphorus from entering the substrate in those regions. Next, a thermal oxidation step is carried out, using a wet oxide process (Fig. 16-9a). This grows an oxide to a thickness of 350–400 nm over the n-well regions. (Since nitride acts as a diffusion barrier, oxide grows only over the regions not covered by nitride.) The thickness of the thermal oxide is selected so it will block the subsequent boron implant used to form the p-wells. The nitride is next stripped using hot phosphoric acid (an etchant that does not attack the thermal oxide grown over the n–well regions). This leaves the silicon in the p-well regions covered only with the pad oxide, while the n-well regions remain covered with the thicker oxide. Thus, during the p-well implant (e.g., B at 50 keV, 1×10^{12} cm^{-2}, using a medium-current implanter), boron ions enter the silicon substrate only in the areas where the p–wells are to be formed (Fig. 16-9b). Next the well dopants are diffused into the substrate using a long, high-temperature drive-in step (e.g., 1100–1200°C for 3–20 hours). This is performed in an oxidizing ambient so that an oxide is also grown on the entire wafer surface during this step. At the conclusion of the drive-in step, the dopant concentration at the Si surface in the wells of 0.8-μm-CMOS is ~1×10^{16}/cm^3 for the p-well, and ~3×10^{16}/cm^3 for the n-well. The n-well might have a higher doping concentration to improve the punchthrough performance of the PMOS devices and to eliminate the need for a

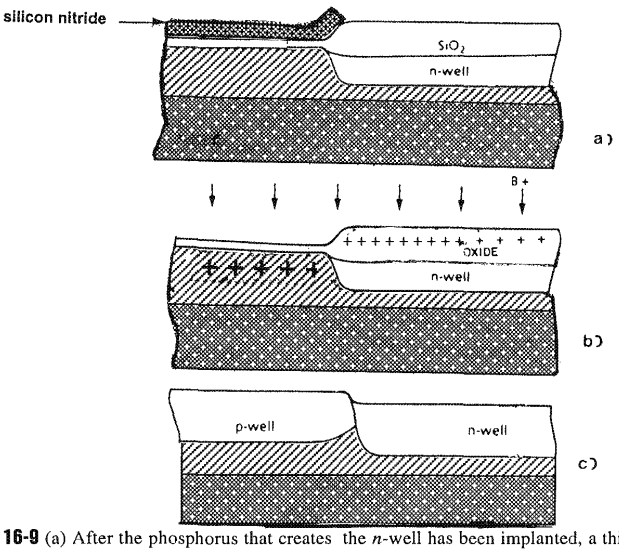


Fig. 16-9 (a) After the phosphorus that creates the *n*-well has been implanted, a thick oxide is grown, with the nitride preventing oxide form growing beneath it; (b) after stripping the nitride, a boron implant is performed that will form the *p*-well regions. The thick oxide above the *n*-wells prevents the boron from penetrating there; (c) wafer-surface topography after wells are driven-in and the oxide is stripped.

separate channel-stop implant step for the *n*-well. The oxide thickness is different over the *n*-well than the *p*-well, creating a step in the oxide film at the well boundaries. This oxide is nevertheless next etched away, but the step in the silicon remains behind (Fig. 16-9c). This step serves as an alignment mark to allow subsequent layers to be aligned to the well boundaries. (An alternative procedure that uses an additional mask can be used to create alignment marks. That is, deep pits can be etched into the wafer prior to the well formation step. Such marks provide improved registration accuracy for lithographic steps, and have become standard in deep-submicron CMOS technologies.)

It is important to observe that when wells are formed in this manner the well doping profile is at a maximum at the surface and then monotonically decreases with depth (see Fig. 16-10). The well doping concentrations are selected to be at least an order of magnitude higher than the epi-layer doping, so the expected variations in epi doping will not significantly affect the well concentration. The junction depth of the wells is typically ~2 μm deep.

16.2.3 Definition of Active and Field Regions: In conventional CMOS processes a standard LOCOS process (described in more detail in Vols. 2 and 3) is next used to define the active and field regions, and to create isolation structures between the active devices (Fig. 16-11). This is done by selectively growing a thick thermal oxide over the field regions, while keeping the active regions devoid of oxide. The steps that accomplish this begin with the growth of a pad oxide by thermal oxidation over the entire wafer to a thickness of 40–50 nm. Next, a blanket layer of silicon nitride (140–200 nm thick) is deposited by LPCVD (Fig. 16-11a). Then photo-

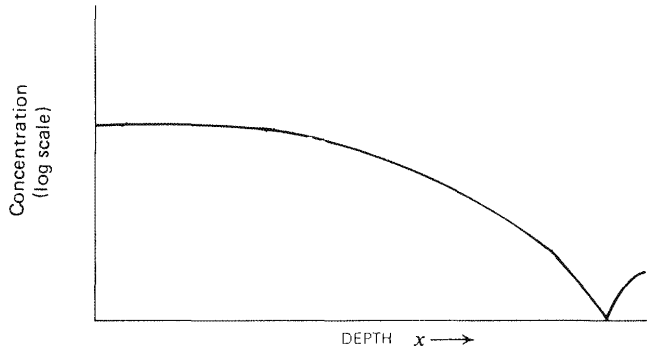


Fig. 16-10 Doping profile in a conventional well formed by a shallow implant and a drive-in step.

photoresist is spun onto the wafer and *Mask #2* is used to create the patterns that define the field and active regions of the circuit (Fig. 16-11b). After exposing and developing this pattern, the nitride is etched and the resist is stripped (Fig. 16-11c).

The next step creates the channel-stop dopant layer in the field regions of the *p*-wells. This is done by implanting a layer of boron atoms near the Si surface in the *p*-wells. This layer of implanted B atoms increases the threshold voltage of the parasitic PMOSFETs in the field regions of the *p*-wells. To prevent the boron from being implanted into the field regions of the *n*-wells, another mask is used. A layer of resist is applied prior to the boron channel-stop implant. *Mask #3* is used to form a pattern that covers the *n*-wells (Fig. 16-11d). Boron is then ion implanted into the surface (B with a dose of 10^{12}–10^{13} atoms/cm^2 at 40-80 keV, using a medium-current implanter). Boron only enters the silicon substrate in regions not covered by *Mask #3* or nitride (Fig. 16-11e).

A thick oxide (called a *field oxide*) is next grown selectively to a thickness of 400-700 nm over the field regions using a wet oxidation process (Fig. 16-11f). The combination of a thick field-oxide and channel-stop dopants beneath this oxide forms *parasitic MOSFETs* in the field regions whose threshold voltages that are too high to be turned on by normal voltages used to run the circuits. This keeps *active MOSFETs* in the IC electrically isolated from one another.

No oxide grows over regions covered with nitride (the active regions). Such a process of selectively growing field oxide has been named *local oxidation of silicon* (*LOCOS*). Observe, however, that some oxide does grow under the edges of the nitride film as a result of the oxidizing species diffusing laterally as well vertically. This forms a tapered oxide region at the edge of the active region called a *birds beak* (Fig. 16-12 and Chap. 8). As the active regions get smaller with device scaling, bird's beak begins to limit the use of LOCOS.

16.2.4 Threshold-Adjust Implantation Step: After the field oxide is grown, the nitride is stripped and the circuit is almost ready for carrying out the threshold adjust (or V_T-adjust) implant process and for growing the gate oxide. Prior to this, however, a problem caused by the previous field oxidation step of the LOCOS process must be rectified. During the wet oxidation process

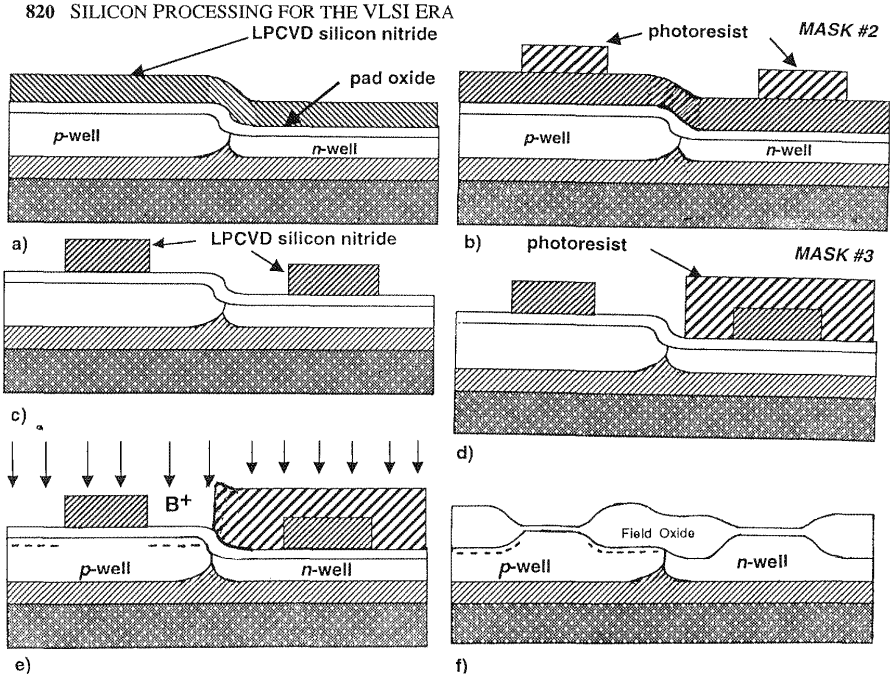


Fig. 16-11 Formation of the active and field regions using LOCOS isolation: (a) grow pad oxide and deposit silicon nitride film by LPCVD; (b) Use *Mask #2* to define active regions; (c) etch the nitride film; (d) Use resist and *Mask #3* to cover *n*-well regions; (e) implant B into the field-regions of the *p*-wells (channel-stop implant); (f) selectively grow field oxide with a wet-oxidation process, strip nitride.

a thin layer of nitride forms on the silicon surface of the active regions (near the edges of the nitride film). Such layers would interfere with the gate oxidation process, insofar as the gate oxide will not grow as rapidly over such regions (and hence it would be thinner there). This problem is called the *Kooi effect.* To overcome it a thin *sacrificial* (or *dummy*) *oxide* is next grown to a thickness of 50–100 nm. This also oxidizes the thin nitride layer so it can be removed with an HF etch. The V_T-adjust implant is usually performed with this dummy oxide layer in place, as the oxide serves another useful function. That is, it also acts as a screen oxide, preventing contaminants from the implantation process from entering the Si substrate.

The V_T-adjust implant step (Fig. 16-13a) sets the threshold voltage of both the NMOS and PMOS FETs, typically to values of 0.6 V and –0.6 V, respectively. The V_T-adjust implant process in CMOS technologies using only n^+ poly can be accomplished with either a single boron implant step (if the background doping is chosen correctly, see Fig. 16-13b), or with two separate boron implant steps (in which case an additional mask is required). A separate punchthrough-prevention implant for the NMOS devices is also usually needed, especially when the minimum device dimension decreases below about 1 μm (see Vol. 3). The punchthrough

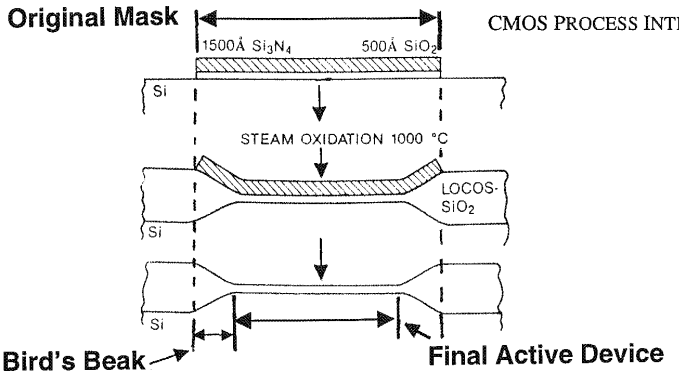


Fig. 16-12 "Bird's beak" in LOCOS isolation structure intrudes into the active regions of the wafer surface. This limits the use of LOCOS in deep-submicron IC process technologies.

implant can be performed sequentially using the same implanter, along with the V_T-adjust implant. Once the V_T-adjust implant step is completed, the dummy oxide is stripped with HF, leaving a residue-free Si surface upon which high-integrity gate oxides can be reliably grown.

16.2.5 Gate Oxide Growth: In the next step the gate oxide is formed. The growth of the gate oxide is a critical step. A defect-free, very thin (6-20-nm), high-quality oxide without contamination is essential for proper device operation. The gate oxide is grown only in the exposed active regions (actually, the field-oxide thickness is also very slightly increased as a result of the gate-oxide growth step). Since the drain current in an MOS transistor is inversely proportional to the gate-oxide thickness (for a given set of terminal voltages), the gate oxide is normally made as thin as possible – commensurate with oxide breakdown and reliability considerations. See Chap. 8 for details on thin gate-oxide growth.

16.2.6 Polysilicon Deposition and Patterning: Once the gate oxidation step is completed, a heavily *n*–doped polysilicon gate structure is fabricated. Polysilicon is the preferred gate material for several reasons. First, it can withstand the high-temperature steps required to form the source/drain junctions. Second, the polysilicon-SiO_2 interface is well understood and electrically stable. The gate formation procedure begins with the deposition of a 0.4–0.5 μm thick, undoped polysilicon film by LPCVD (Fig. 16-14a, and see Chap. 6 for details). This layer is then doped with phosphorus by ion implantation or chemical doping (Fig. 16-14b), producing a film with a sheet resistance of 20–30 Ω/sq While this sheet-resistance value is small enough for MOS circuits with gate lengths ≥ 2 μm, lower values are required for submicron ICs. In such applications, composite layers of refractory metal silicides and polysilicon (called *polycides*) are used to reduce the sheet resistance to ~1–3 Ω/sq. This allows the benefits of the polysilicon-SiO_2 interface to be preserved and the gate sheet resistance to be lowered. For the CMOS processes being discussed here, CVD WSi_x became the most popular material used to form such polycide structures (see Chap. 6 for details on CVD WSi_x deposition).

The gate structure and polysilicon interconnect structures are then patterned using *Mask #4* (Fig. 16-14c). Following exposure and development of the resist, the polysilicon (or polycide) film is dry-etched (Fig. 16-14d). This is a critical etch step for several reasons. First, due to the

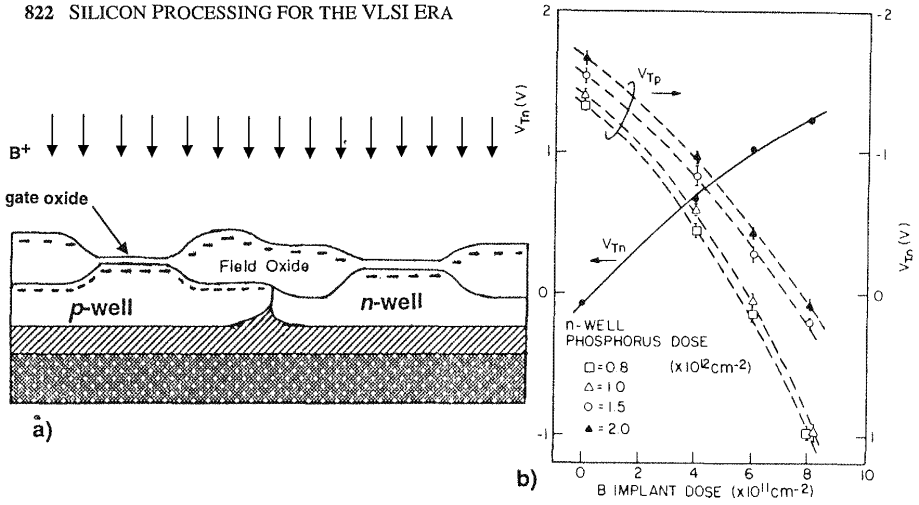


Fig. 16-13 (a) Threshold-voltage adjustment carried out using a single boron implant; (b) threshold voltages of NMOS (V_{Tn}) and PMOS (V_{Tp}) transistors as a function of boron dose. The CMOS structure uses an *n*-well implanted into a *p*-substrate (with doping level of $6x10^{14}$ atoms/cm^3). V_{Tp} results are shown for various implant doses of the *n*-well.[10] (© 1980 IEEE).

self-aligned nature of silicon-gate technology, the channel length of the device depends on the width of the polysilicon line. Hence, the gate-length dimension must be precisely maintained across the entire wafer, and from wafer to wafer. If the gate length is too long, the drain current of the MOSFETs will decrease, slowing down the performance of the IC. If the gate length is too short, the source and drain may punch through. Second, the profile of the etched poly gate structure should be vertical. This will prevent variation of channel lengths by the penetration of the ions of the thinner regions of the gate sidewalls during formation of the source/drain regions by ion implantation (Fig. 16-14e). Third, to achieve the above goals, an anisotropic polysilicon etch process must be employed. This process, however, requires overetching to remove the locally thicker regions of polysilicon that exist wherever it crosses steps on the wafer surface. During the overetch time, areas of the thin gate oxide are exposed to the etchants. Thus, it is necessary to use a polysilicon etch process that is highly selective with respect to SiO_2.

After the polysilicon is etched, a step called *poly reox* is usually performed (16-14f). This grows a thin thermal SiO_2 layer on the polySi (on the order of 10 nm thick). The reasons for performing this step are discussed in Vol. 3, Chap. 9.

16.2.7 Formation of Source/Drain Regions: The next step in the process flow is the formation of source and drain (S/D) regions of the MOSFETs. Such regions are paths of current flow in the silicon between the metal interconnect lines and the channel of the transistor. As such, it is important that they have the lowest possible resistance. In addition, submicron MOSFETs require S/D junctions to be as shallow as possible to suppress such short-channel effects as *punch-through,* see Vol. 3. To obtain low resistivity, the S/D regions are doped as heavily as possible (typically using ion implantation, with a dose on the order of ~10^{15} cm^{-3}). The NMOS S/D

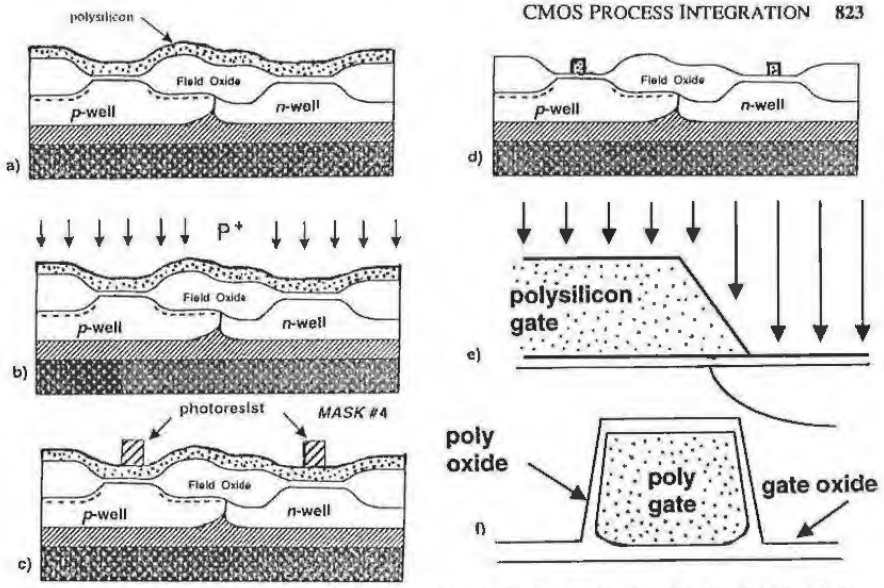


Fig. 16-14 n^+-polysilicon gate formation: (a) deposition of undoped polysilicon film by LPCVD; (b) *ex-situ* doping of the polysilicon with phosphorus by chemical diffusion of ion implantation; (c) Use *Mask #4* to define the polysilicon gates; (d) etch polysilicon and strip resist; (e) the etch process used to pattern the gate must yield poly features with vertical sidewalls (to maintain control of the effective gate length during the implant process that forms the source and drain regions); (f) *polysilicon re-ox* step grows a thin oxide on the poly by a short thermal oxidation process.

regions are doped with As, because it has a high solubility, low diffusivity, and a shallow projected range at the low ion implant energies that are used. Boron or BF_2^+ are used to dope the PMOSFET S/D regions. However, boron has a higher diffusivity than As. Hence, shallow S/D junctions in PMOSFETs are harder to achieve than in NMOSFETs. Since it is necessary to selectively implant source/drains to form n^+ regions for the *n*–channel FETs and p^+ regions for the *p*-channel FETs, this usually requires two masks (i.e., *Mask #5* and *Mask #6*, Fig. 16-15).

In submicron-CMOS processes, gate lengths become so small that *lightly doped drain* (*LDD*) structures must be used to minimize hot-electron effects, especially in NMOS devices. Thus, procedures are integrated into the CMOS process flow to fabricate such LDD structures. If these structures are used only for NMOS transistors, *Masks #5* and *#6* will still allow the sources and and drains of the two transistor types to be selectively implanted.* However, if LDD structures are needed for both PMOS and NMOS devices, two more masking layers may be needed.

* A procedure that requires only one masking step for creating both types of source and drain regions has also been reported. In this approach, the heavy boron implant is carried out non-selectively. The mask is then used to cover the PMOS device active regions. An *n*-type implant heavy enough to overcompensate for the boron implant in the active regions of the *n*-channel devices is used to form the sources and drains of the NMOS transistors.

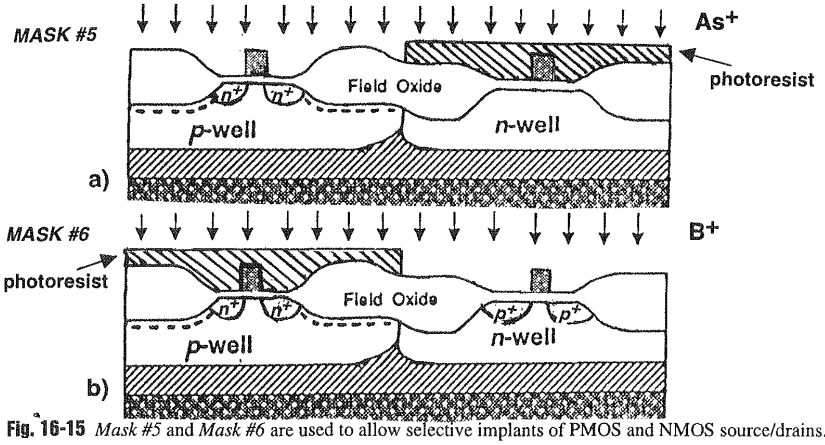


Fig. 16-15 *Mask #5* and *Mask #6* are used to allow selective implants of PMOS and NMOS source/drains.

The LDD structure is formed in the following way. The lightly doped regions of the source and drain are created with an ion implant step (Fig. 16-16b). A dose of approximately $3x10^{13}$–$3x10^{14}$ cm^{-2} dopant ions is implanted at low energy (30–50 keV). The implantation process causes the edge of these implanted ions to be automatically aligned to the edge of the gate (i.e., it is a *self-aligned process*). A conformal layer of dielectric material (usually SiO_2 or silicon nitride) is then deposited over the entire wafer. An anisotropic etch process is used to clear the oxide in the flat areas while leaving sidewall spacers on the sidewalls of the poly gates (Fig. 16-16c). These spacers cover and protect the regions beneath them from the subsequent high-dose implant that forms the rest of the S/D regions. A dose of $1x10^{15}$–$5x10^{15}$ atoms/cm^2 at 40–80 keV is used for this step (as shown in Fig. 16-16d).

16.2.8 Premetal Oxide Deposition and Contact Formation: Following formation of the source and drain regions, a doped *premetal dielectric* (*PMD*) is deposited by CVD. (This layer is also known as an *interlevel dielectric* or *ILD*.) Contact windows are opened in this dielectric layer to allow electrical connections to be made between Metal 1 and the following structures: 1) source/drain contact regions: 2) gate contacts; 3) substrate-contact regions; and 4) well-contact regions. A CVD process is used to deposit doped SiO_2, about 1 μm thick (Fig. 16-17a), onto the wafers (see Chap. 6 for details of this process). The dopant in the SiO_2 is either phosphorus (in which case the material is referred to as *phosphosilicate glass* or *PSG),* or both phosphorus and boron (making it *borophosphosilicate glass* or *BPSG*).

The doped CVD glass layer plays several roles in the fabrication and operating aspects of the circuit. First, it acts as an insulating layer between polysilicon and Metal 1. Second, it reduces the parasitic capacitance between Metal 1 and the substrate. Third, the addition of phosphorus to the glass makes the layer an excellent getter of Na ions (contamination by Na can destabilize the V_T of an MOS device). The PSG (or BPSG) binds otherwise mobile Na atoms within the doped-glass layer, preventing them from reaching the gate oxide and altering V_T. Finally, the dopants in the glass make it viscous at elevated temperatures (1000–1100°C for PSG, 850–950°C for BPSG, see Ch. 6). This allows the layer of glass to be flowed after it is deposited (Fig. 16-17b).

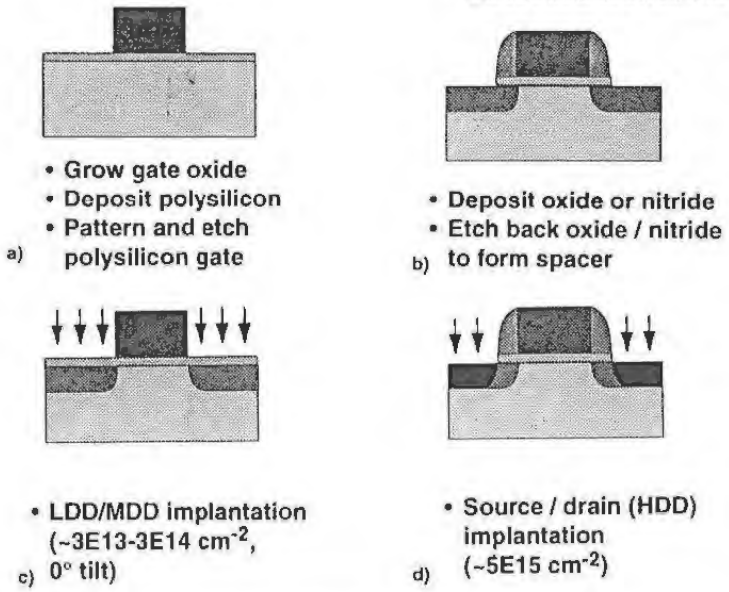


Fig. 16-16 Basic process to form an LDD NMOSFET: (a) after poly etch; (b) LDD implant; (c) after spacer material deposition and etch; (d) n^+ implant.

Through this procedure, a rounding of the glass contours and smoothing of any sharp steps is achieved. This produces better step coverage of the metal (which is deposited next) over the otherwise severe wafer topography. The high-temperature glass-flowing step also serves as an annealing process that activates the dopants implanted to create the source/drain junctions.

Contact openings are next created by a lithography-and-etch step (Fig. 16-17c). *Mask #7* is used to define contact patterns in a photoresist film. A dry-etch process is then used to open the contact windows through the interlevel dielectric to the underlying polysilicon and the source/drain regions in the silicon.

This contact-opening step can be critical, as the contact size and its alignment to underlying patterns limit the minimum size of the device. The source/drain regions must also be large enough for the contact to fit, with an allowance for alignment tolerance. If the contact opening exposes a part of the substrate, the drain or source will be shorted to it. Likewise, if the contact opening overlaps the both the source/drain and the gate, a short will be created between them.

After the contact etch is completed and the resist is stripped, the doped CVD glass is subjected to another flow step. This procedure rounds the corners of the top of the oxide windows so metal step coverage into the contact windows is improved (called *reflow* of the contact windows, see Chap. 6). It may still be too difficult to deposit metal into these contact holes with adequate step coverage. A number of techniques have been studied to improve the coverage of Metal-2 into the vias (Chap. 15). In Fig. 16-17d, the step coverage of Metal-1 into

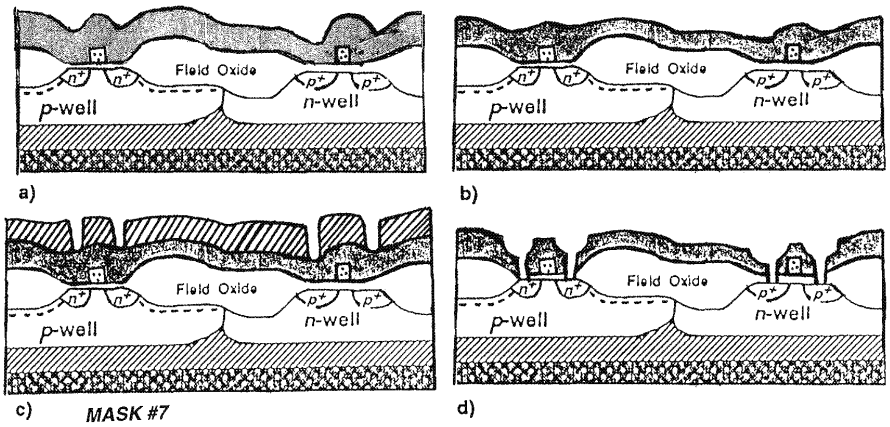


Fig. 16-17 (a) Deposit interlevel dielectric BPSG layer (ILD) by CVD; (b) flow BPSG. (c) use *Mask #7* to define contact holes; (d) etch contact holes with "wine glass" etch and strip resist.

the contact hole is facilitated by the use of a *wine-glass* etch process. The contact hole is formed with a two-step via-etch process (i.e., using an isotropic etch process to etch half of the ILD thickness, and an anisotropic etch step to etch the remaining half). Reflow may still be carried out after such a two-step etch.

16.2.9 Metal Deposition and Patterning: After the contacts have been opened, the metallization layer is deposited. Because the metal layer is highly conductive, it is used whenever possible to interconnect circuit elements and to carry large amounts of supply current. The metal interconnect lines that are fabricated must have sufficient thickness, width, and step coverage to keep the current density in each line below the value that could produce electromigration failure (Vol. 2, Chap. 4). In addition, the space between adjacent metal lines must be kept large enough that the lines never touch, even under worst-case process variations.

Although evaporation was the method employed to deposit Al in the early days of MOS, it has generally been replaced by sputtering. The change was made because Al alloys with tightly controlled compositions became the materials of choice for the metal layer. Sputtering allows alloys to be deposited with much better compositional fidelity (see Chap. 11).

The metal alloy eventually chosen for NMOS was Al:1wt% Si. Silicon is added to the aluminum film to prevent spiking of the contacts during subsequent annealing steps (see Chap. 11). In CMOS, however, such Al:Si alloys were replaced with a metallization system that includes a diffusion barrier (typically Ti/TiN, see Chap. 15), and Al:Cu alloys for the main interconnect layer (Fig. 16-18a). Dry etching is used to pattern the interconnect films using *Mask #8* (Fig. 16-18b).

If a single level of metal is used in the CMOS process, a sintering step occurs after the metal has been patterned. This step brings the metal and the n^+ and p^+ regions in the silicon into intimate contact. Such intimate contact between the metal and the heavily doped Si regions

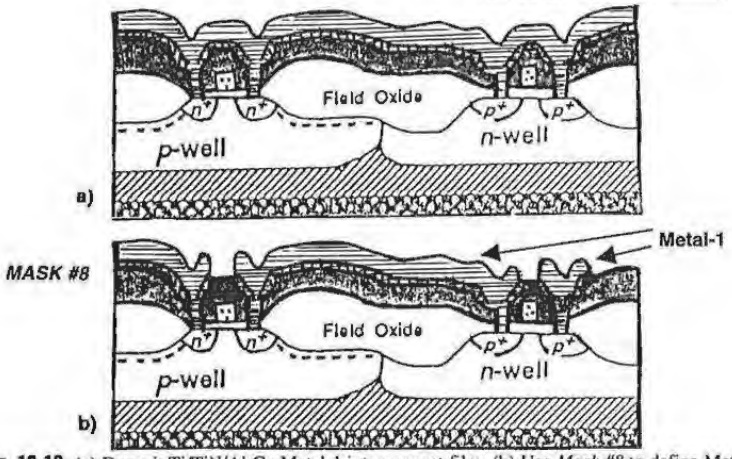


Fig. 16-18 (a) Deposit Ti/TiN/Al:Cu Metal-1 interconnect film. (b) Use *Mask #8* to define Metal-1 pattern and then etch the metal with a dry-etch process.

establishes a low-resistance ohmic contact. The annealing process exposes the wafer to a 375–500°C temperature in an H_2, or $N_2 + H_2$ (5%) ambient for about 30 minutes. This step is also used as the annealing process for reducing the interface trap density in the gate oxide that was introduced by earlier processing steps (see Chap. 8).

16.2.10 Intermetal Dielectric Deposition and Via Patterning: If a two-level-metal interconnect structure is used, an *intermetal dielectric* (*IMD*) must be deposited to electrically isolate the Metal-1 layer from the Metal-2 layer (Fig. 16-19a). A variety of CVD processes for depositing such dielectric films have been developed (see Chap. 6).

Deposition of this layer may make the wafer topography once again too severe to allow Metal-2 films to be deposited with adequate step coverage. One of the planarization processes discussed in Chap. 15 may need to be used to overcome this difficulty.

If such techniques are successfully implemented, the wafer topography will be relatively planar, and any steps on its surface will be gently sloped rather than severely vertical. It is also necessary to open vias in the intermetal dielectric layer so an electrical connection can be established between Metal-2 and Metal-1 at desired locations (*Mask #9*). Metal-2 must be deposited with adequate step coverage into these vias, but reflow is not possible because Al is present on the wafer (Metal-1). Thus, in this example, a champagne glass etch process is again used to etch the vias (Fig. 16-19b).

16.2.11 Metal 2 Deposition and Patterning.: The processing issues of depositing and patterning (*Mask #10*) the Metal-2 layer (Fig. 16-19c) are discussed in Chaps. 11, 14, and 15.

16.2.12 Passivation Layer and Pad Mask: Finally, a *passivation* (or *overcoat*) layer, such as CVD PSG or plasma-enhanced CVD silicon nitride (or both), is put down onto the wafer surface (Fig. 16-19d). This layer seals the device structures on the wafer, protecting them from contaminants and moisture. It also serves as a scratch protection layer.

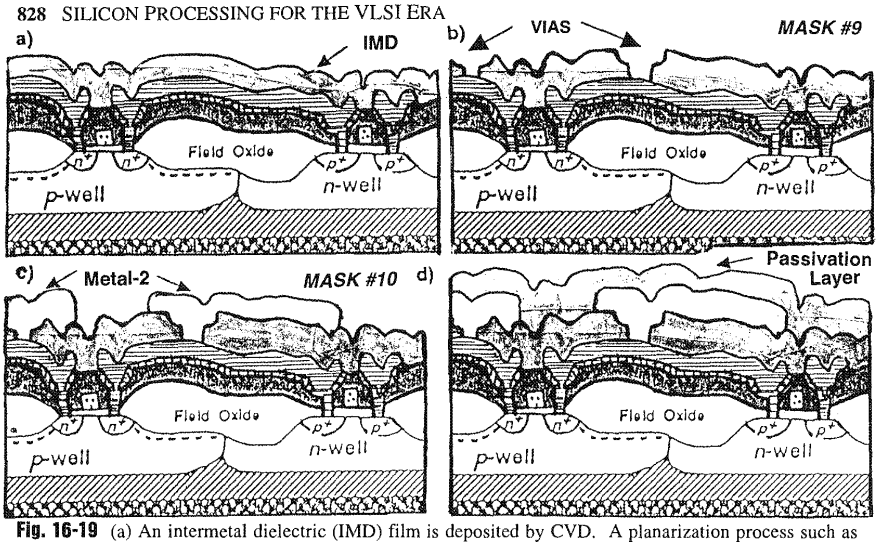


Fig. 16-19 (a) An intermetal dielectric (IMD) film is deposited by CVD. A planarization process such as resist etchback, or spin-on-glass, may be used to smooth this surface. (b) Vias are defined in the IMD film with *Mask #9* and an etch step. (c) Metal-2 is deposited and defined with *Mask #10* and a metal etch process. (d) A final passivation layer is deposited on the wafer by CVD.

Openings are etched into this layer so that a set of special metallization patterns under the passivation layer is exposed. These metal patterns are normally located in the periphery of the circuit and are called *bonding pads* (Fig. 16-20). Bonding pads are typically about 100 x 100 μm in size and are separated by a space of 50 to 100 μm. Wires are connected (bonded) to the metal of the bonding pads and are then bonded to the chip package. In this way connections are estab-

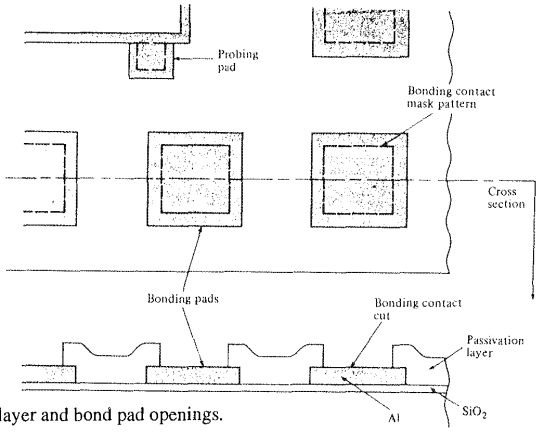


Fig. 16-20 Passivation layer and bond pad openings.

lished from the chip to the package leads (see Chap. 17 for details on chip bonding).

The bond pad openings are created by patterning the passivation layer with *Mask #11*. If a PSG layer is used, the phosphorus (2-6 wt%) in the glass not only causes it to act as a getter for Na but also prevents the film from cracking. Care must be taken to ensure that not more than 6% phosphorus is incorporated into the PSG, as this can cause corrosion of the underlying metal if moisture enters the circuit package (see Chap. 11). When silicon nitride is used, care must be taken to ensure that the deposited nitride film exhibits low stress (either tensile or compressive), so it will not crack. Cracking would compromise the sealing capability of the film.

16.3 PROCESS FLOW FOR 0.25 μm CMOS

An example process flow for 0.25 μm CMOS (and perhaps for the 0.18 μm generation as well) will be presented in this section. (At this book's writing, the process flows for 0.25 μm CMOS, and below, were not yet "standardized," and many possible variations of the flow postulated here existed.) This flow is "compiled" from the literature, where a number of 0.25 μm CMOS technologies are reported, including those in Refs. 11, 12, 13, 14. Since the literature does not provide all the details of such flows, some of the information here is necessarily conjectural. Despite these issues, the 0.25 μm flow presented here should be a reasonable representation of those found in actual use.

A number of general characteristics of 0.25 μm CMOS process technology should be noted. First, the lithography technology used for the critical masking layers is DUV optical lithography, which is based on 248-nm KrF laser light sources (and *i*-line lithography perhaps being used for the non-critical layers). Second, the power supply voltages of the systems using 0.25 μm CMOS is 2.5 V, and for 0.18 μm CMOS is estimated to be 1.8 V. This is a deviation from the 5-V power supply used from the mid-1970s's and down to when CMOS reached 0.65 μm. However, when CMOS reached 0.5 μm, the gate oxide thickness had been scaled so much that the electric field across the gate oxide exceeded the value that would permit reliable operation for the designed life of the systems (e.g., 10 years). To ensure adequate device lifetime, the power supply voltage had to be changed for 0.5-μm CMOS to 3.3 V, and again at 0.25-μm CMOS to 2.5 V. (This is an example of how the limits of process technology impact the electronic systems specifications.) Third, the number of masking layers is likely to exceed 20, especially if 5 or more interconnect levels are needed. Finally, 0.25-μm CMOS was initially being manufactured on 200-mm wafers. It is likely that 0.18-μm CMOS will also use such wafer sizes. However, it appears that 0.15-μm CMOS, or 0.12-μm CMOS will enter large scale production on 300-mm wafers.

As mentioned in the introduction, there are a number of process modules that had to be fundamentally changed as CMOS technology was scaled below about 0.35 μm. Figure 16-21 demonstrates this quite clearly.[13] The order of carrying out some of the steps has also changed. Why these changes in the process flow were necessary will be indicated.

16.3.1 Starting Material: The starting material for deep-submicron CMOS is predicted to shift more to p-epi-on-p^+ wafers (Fig. 16-7b), because the gate-oxide-integrity is better on epi than on bulk wafers. However, the higher cost of epi compared to bulk wafers remains a concern, especially for such low-price commodity chips as DRAMs. Less expensive alternatives to epi are being studied. One process that shows great promise is the implantation of a high dose of

Generation	0.5um	0.35um	0.25um	0.18um	0.15um	0.13um
Isolation	LOCOS		STI			
Well	Diffused		Retrograde			
Gate	WiSix n+ Polycide		Dual Poly (n+/p+) Salicide			
Gate Oxide	Single Oxide		Dual/Triple Oxide => Oxynitride			
Gate Litho	I-line		248nm, Hi NA		PSM	193nm
Silicide	None	TiSix		TiSix => CoSix		
Contact/Via	Etchback		CMP			
BEOL Dielectric	SOG	Oxide		Low-k		
Metal	Al			Al =>Cu Damascene		

Fig. 16-21 The process modules used for successive submicron-CMOS generations.[13] © IEEE (1998).

boron deep below the surface of a bulk wafer, using high-energy ion implantation. This step creates a heavily doped B buried layer that acts in the same way as the heavily doped substrate of the epi wafer.[16] Good latchup protection and GOI are reportedly achieved by this approach. The cost for carrying out this step is 10–20% that of depositing an epi layer on a bulk wafer.

16.3.2 Formation of the Shallow-Trench-Isolation Structure: The isolation structures are formed next. Note that not only is a different type of isolation structure formed (*shallow trench isolation*, or *STI*) instead of LOCOS, but it is formed prior to the well structures. STI replaces LOCOS because STI has no bird's beak and provides a planar surface for further processing.[15]

The sequence of steps for forming STI structures begins with the growth of a pad oxide, followed by an LPCVD nitride layer. Then resist is applied and *Mask #1* is used to pattern the STI trench openings. The nitride and pad oxide are etched first. Then the trench is anisotropically etched (Fig. 16-22a) to a depth of 400–500 nm (e.g., using an etch gas mixture of $HBr/Cl_2/O_2$). The trench etching process should yield smooth trench sidewalls with angles of 70–85°, rounded bottom-corners (to minimize stress), and a residue-free silicon surface after etch (see Fig. 16-22f). After the trenches are etched the resist is stripped and a thin thermal oxide is grown on the trench walls (Fig. 16-22b). Next, a CVD dielectric film is used to fill the trench. This layer also covers the areas of the wafer where the nitride remains (Fig. 16-22c). A CMP step is used to polish back the dielectric layer until the nitride is reached (the nitride acts as a CMP-stop layer, Fig. 16-22d). The dielectric material is densified at about 900°C. Then the nitride is stripped, leaving the STI structure in place (Fig. 16-22e).

16.3.3 Formation of Retrograde Wells and Carrying Out of V_T-Adjust Implants: The well formation sequence occurs next. In addition, the V_T-adjust implants and the punchthrough implants are often carried out on the same ion implanter and during the same pumpdown as the well-formation implants. This process of performing multiple implants during the same pumpdown is termed a *chained implant.* The well formation step employs high-energy ion implantation to locate the well dopant peak ~1–2 μm below the Si surface. The implanted dopants will

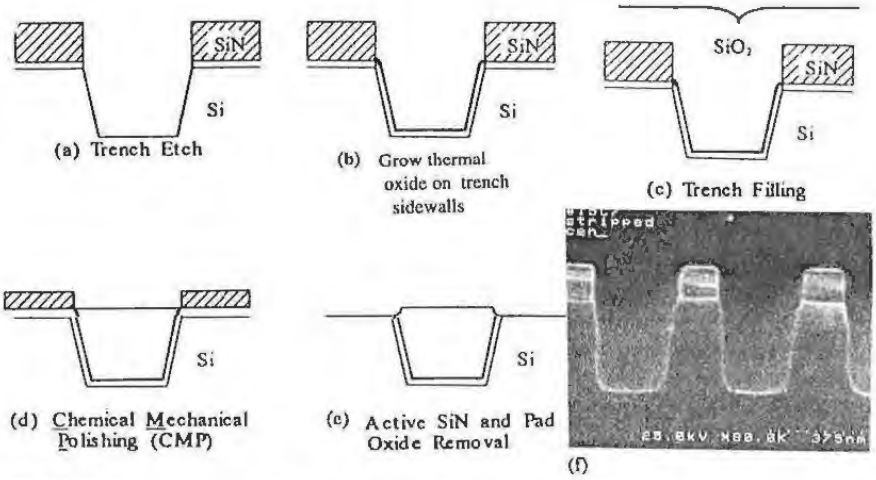


Fig. 16-22 Process sequence used to form shallow-trench isolation structures (STI): (a) etch trench in silicon substrate; (b) grow thermal oxide trench liner to improve trench Si/SiO_2 interface (and for trench corner rounding); (c) fill trench with CVD oxide; (d) use CMP to etch back the oxide to the nitride layer; (e) Strip nitride to leave the STI structure; (f) SEM of trench profile with 87° trench sidewall slope, and top and bottom trench corner rounding. Courtesy of Lam Research Corp.

have a *retrograde* concentration profile toward the surface from this peak (Fig. 16-23). Such a doping profile has several device performance advantages (discussed in Vol. 3). Since the dopants are placed deeply enough by the implant itself, there is no need for the long, high-temperature drive-in step required previously. Instead, only a lower-temperature anneal (carried out in a furnace at 900°C for 30 min), or a very-short, high-temperature RTP anneal (1100°C for 30 sec), is required to activate the implanted dopants. However, the high-energy ion beam used in the implantation process (500 keV–1 MeV) requires the use of a high-energy ion implanter and very thick resist films (~5 μm). Such thick resists are needed to prevent high-energy ions from reaching the Si surface in regions covered by the resist.

The well-formation process is begun by growing a layer of sacrificial oxide on the wafer. After applying the thick resist layer, *Mask #2* is applied to open the patterns for where the *n*-wells will exist. Then a high-energy P implant is carried out (700–850 keV, $1x10^{13}$ cm^{-2}). The peak of the P implant is ~1 μm below the Si surface (Fig. 16-24b). Both the V_T-adjust implant for the PMOSFETs (using P, since the dual-doped poly will allow the use of surface-channel PMOS devices – see Vol. 3), and the punchthrough implant are carried out in this implanter in the same pumpdown. Then the resist is stripped, and a new coat of thick resist is applied. *Mask #3* is then used to pattern the *p*-well openings. A high-energy B implant is then carried out (450–500 keV, $1x10^{13}$ cm^{-2}, Fig. 16-24c). A retrograde profile is again created, with the B peak concentration residing at 1-2 μm below the surface (Fig. 16-25). The V_T-adjust implant and the

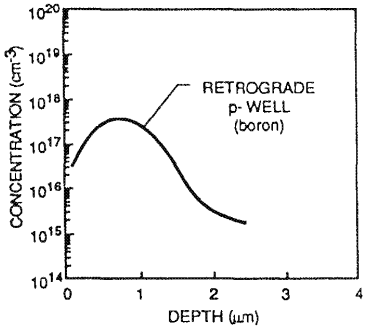


Fig. 16-23 Example of the doping profile of a retrograde *p*-well (boron) implant (400 keV).

punchthrough implant for the NMOSFETs (now using B) are also carried out during this "chained-implantation" procedure.

As discussed in Chap. 10, V_T-adjust implants with retrograde doping profiles have recently been introduced. A high-mass dopant of the same conductivity type is substituted for the ordinary dopant (e.g., indium for boron in NMOSFETs, antimony for phosphorus in PMOSFETs). The higher mass of the dopant reduces its diffusivity, which significantly decreases dopant redistribution during subsequent processes. Thus, the desired as-implanted profile is kept more intact. Such high-mass dopant V_T-adjust implants end up having a reduced

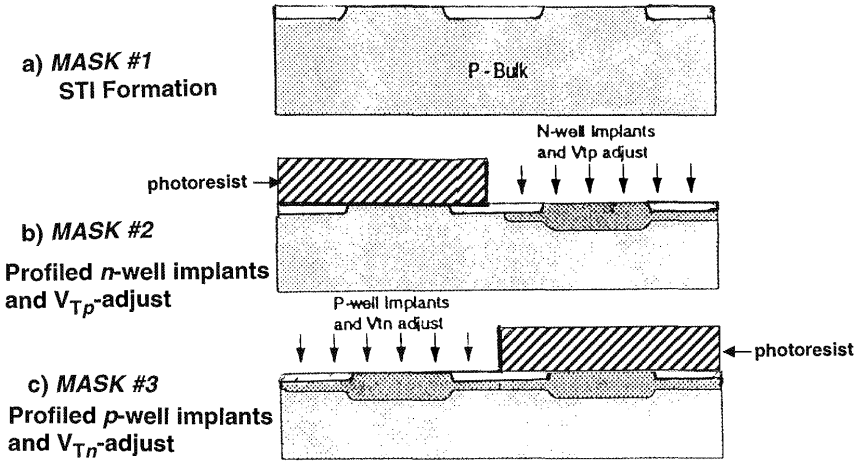


Fig. 16-24 (a) *Mask #1* patterns the shallow-trench isolation features; (b) *Mask #2* defines the *n*-well regions. A high-energy phosphorus (P) implant creates the retrograde *n*-wells, and a lower energy P (or Sb) implant is used to adjust V_{Tp}; (c) *Mask #3* defines the *p*-well regions. A high-energy boron (B) implant creates the retrograde *p*-wells, and a lower energy B (or In) implant is used to adjust V_{Tn}.

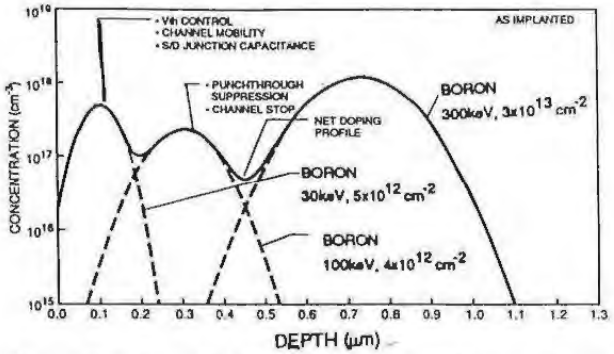


Fig. 16-25 Doping profile in the *p*-well showing the as-implanted: (a) retrograde *p*-well implant (B, 300keV); (b) the anti-punchthrough implant (B, 100 keV); and (c) the V_{Tn} implant (B, 30 keV).

doping concentration (~$5x10^{16}$ cm^{-2}) at the interface while at the doping peak (about 80 nm below the oxide interface) the concentration is ~$1x10^{18}$ cm^{-2}. These profiles have been named *super-steep retrograde* (SSR) channel profiles (Chap. 10, Fig. 10-39). Such SSR profiles allow deep-submicron MOSFETs to be fabricated in which short-channel V_T-rolloff and drain-induced barrier lowering (DIBL) effects are significantly suppressed (see Vol. 3).

16.3.4 Formation of the Gate Oxide: The gate oxides of deep-submicron CMOS will continue to scale, with gate oxides for logic CMOS predicted to get thinner than 30 nm (see Chap. 8). The gate oxide, however, will have to prevent boron (from the p^+-poly) from penetrating into the Si substrate. The addition of nitrogen to the gate oxide has been reported to combat this problem, and so gate dielectric layers may be more like oxynitrides than SiO_2.[13] Different techniques have been investigated for incorporating the nitrogen into such films (see Chap. 8), including annealing the as-grown SiO_2 film in N_2O[18] or NO,[13] implanting N into Si and then growing SiO_2,[19] or exposing the SiO_2 to a high-density nitrogen discharge and then performing a post-nitridation anneal.[20]

16.3.5 Formation of the Gate-Stack Structures: Dual-doped polysilicon (i.e., n^+-doping of NMOS poly gates and p^+-doping of PMOS poly gates) is needed for 0.25-μm CMOS technologies (and smaller). This permits both the NMOS and PMOS devices to be fabricated as surface channel FETs (see Vol. 3), a necessity for being able to control short-channel effects.

The polysilicon layer is deposited undoped by LPCVD, to a thickness of 300-400 nm. But the selective doping of the appropriate poly structures is not carried out immediately. (Instead, it is performed later, simultaneously with the formation of the heavily doped regions of the source and drain junctions by ion implantation.) *Mask #4* is used to pattern the poly layer, and the gate structure is etched with an anisotropic dry-etch step (Fig. 16-26a).

16.3.6 Formation of the Source/Drain Junctions: After the poly gate structure is etched, the formation of the source/drain regions is undertaken. These regions in the substrate have to provide a low-resistance path from the contact to the channel edge. They are formed in two steps. In the

first, so-called *source/drain extensions* (SDEs) are formed. In the previous flow these parts of the S/Ds were termed the *lightly doped areas* of the source and drain. Since they were formed by an ion implantation step, the doping in these regions is characterized in terms of the implant dose (typically $3x10^{13}$–$3x10^{14}$ cm^{-2}). In deep-submicron FETs, the doping concentration is increased in the SDE areas above the values used in the first process flow presented earlier. If the SDEs are also formed by ion implantation, the doping in them can again be characterized by the implant dose, which in the present process flow is raised to about $1x10^{15}$ cm^{-2}.

However, the primary function of the SDE is no longer to act as a region that reduces the local electric field for the purpose of reducing hot-carrier effects (because this problem is reduced by the smaller power supply voltages used). Instead, the SDE has two other principle roles that must simultaneously be met: 1) to provide a low resistance path from the contact to the channel edge; and 2) to have extremely shallow junction depths (e.g., 50–100 nm for 0.25 μm CMOS, and 36–72 nm for 0.18 μm CMOS), so that short-channel effects can be effectively suppressed. It is imperative that a very high doping concentration be established within such a shallow structure so that the parasitic series resistance of the SDE becomes as small as possible. Unfortunately, forming a shallow junction at the wafer surface with a high doping concentration is a challenging task, especially for B-doped regions. Several techniques are being developed to try to achieve the above goals. Ion implantation at very low energies followed by an RTP anneal has been the most widely used process for forming these drain extension regions (see Sect. 10.6). Other techniques being investigated for this process include *gas immersion laser doping* ([*GILD*] - see Chap. 8), *plasma immersion doping* (see Chap. 10), and out-diffusion from a layer of deposited doped oxides (e.g., B from a layer of BSG), as described in Refs. 21 and 22.

Regardless of the technique being used, two masks are needed (i.e., *Mask #5* and *Mask #6,* shown in Fig. 16-26b and 16-26c). This allows the NMOS and PMOS source/drain extension (SDE) regions to be selectively doped. If ion implantation is being used, a "halo implant" (to control punch-through) can also be performed within the same pumpdown as the SDE implant step. This "halo implant" is also self-aligned with the channel edge: and it puts dopants at a depth beneath those dopants implanted into the SDE regions. "Halo" implants can be performed either with a 0° tilt angle or a high (30°) tilt angle (see Chap. 10).

The spacer formation is carried out next. Typically LPCVD silicon nitride is used as the spacer material. The deposited nitride layer is then anisotropically etched-back, to form the spacer. At that point the heavily doped, deeper part of the source/drain regions are formed. This is done with ion implantation, using two masks, so the NMOS S/D regions are formed with one implant, and the PMOS S/D regions with the other (*Mask #7* and *Mask #8,* as shown in Figs. 16-27a and 16-27b). Doses of $1–5x10^{15}$ cm^{-2} are used for this step. Note these regions have deeper junction depths than the SDEs (Fig. 16-28). The poly gates are also doped simultaneously by these implants. A single rapid-thermal annealing step is used to activate the implanted dopants in both the SDE and the heavily doped S/D regions.

16.3.7 Salicide Formation: A *salicide* (self-aligned silicide) structure is fabricated next. It consists of a metal silicide formed atop the lines of polysilicon that make up the gates and the local interconnects (akin to the polycide structure described in the first flow). However, in this case it is also formed on the source-drain regions. The silicide on the source/drain regions

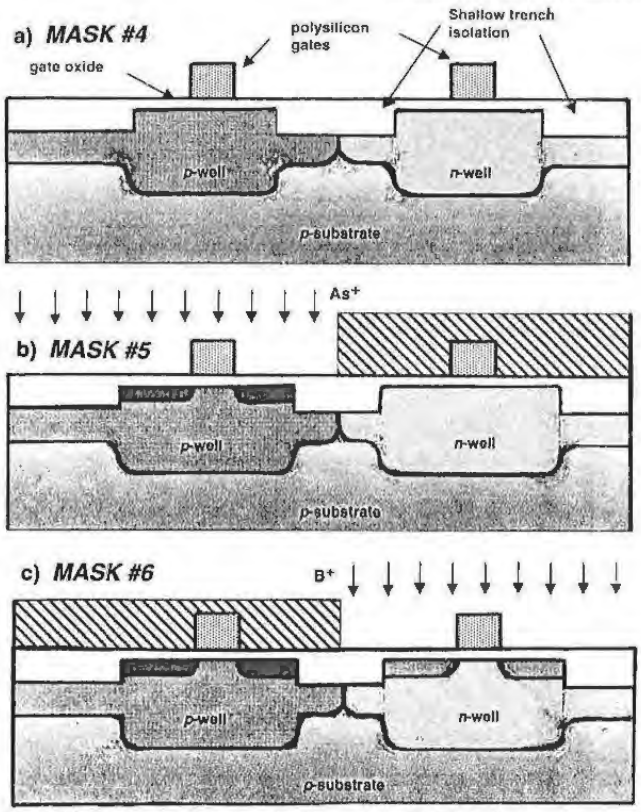


Fig. 16-26 (a) *Mask #4* is used to define the polysilicon gates; (b) *Mask #5* is used to cover the *n*-well regions, so that the NMOS SDEs can be selectively doped with an As implant; (c) *Mask #6* is used to cover the *p*-well regions, so that the PMOS SDEs can be selectively doped with a B implant.

reduces the sheet resistance of the path between the metal contact and the channel edge. The spacers formed on the sidewalls of the poly lines (in the course of fabricating the two parts of the source/drain regions) are now used once again in the salicide formation sequence. Two metals have been used to form such salicides, Ti and Co. They both form silicides with resistivities of 10–15 $\mu\Omega$-cm. Ti-salicide structures were the first ones implemented, but Co has become more popular as the dimensions have been scaled toward 0.18 μm.[23]

The sequence to form the salicide begins with the removal of any oxide on the source/drain regions and the poly lines (Fig. 16-29). Then, Ti or Co is blanket deposited by PVD (sputtering) onto the wafer. A lower temperature thermal step (~650°C) is next used to react the metal and the Si exposed on the substrate and the poly lines. (Use of an RTP step, in which oxygen is

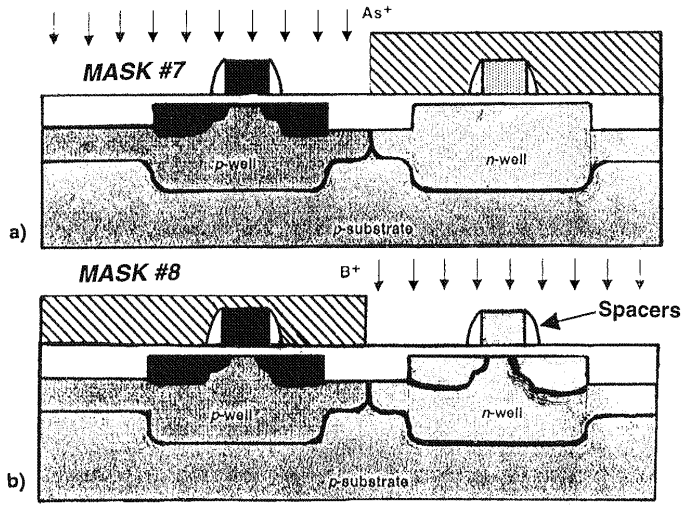


Fig. 16-27 (a) *Mask #7* is used to cover the *p*-well regions, so that the heavily doped NMOS S/D regions can be selectively doped with an As implant; (b) *Mask #8* is used to cover the *n*-well regions, so that the heavily doped PMOS S/D regions can be selectively doped with a B implant.

excluded from the ambient during the anneal, is a key aspect of this process.) This forms a metastable C49 phase of $TiSi_2$. The Ti that lies on either the field oxide or the nitride spacers is, in principle, not converted to $TiSi_2$ during this step. The C49, however, is a metastable, high resistivity material, not yet suitable for the interconnection task. The wafer is then etched in a bath of H_2O:H_2O_2:NH_4OH that selectively removes only the unreacted Ti, leaving behind the C49 $TiSi_2$. A second thermal step using RTP is carried out, this time at a temperature greater than 750°C. This converts the C49-phase $TiSi_2$ to the more stable, lower resistivity (~10 $\mu\Omega$-cm) C54-phase $TiSi_2$. The thermal steps are carried out by RTP and in an N_2 atmosphere. Details of this sequence are found in Vol. 2, Sect. 3.9.1.1.

When the polysilicon lines become smaller than about 0.5 μm, the formation of $TiSi_2$ becomes difficult. The transformation of the C49 phase to the C54 phase becomes the limiting factor. On such small lines, the density of C54 grains is low, and only a small fraction of the $TiSi_2$ consists of the C54 phase, because it can grow only from the few nuclei available. This problem could be solved by heating the $TiSi_2$ above 750°C to accelerate the nucleation of the C54 phase. Unfortunately, thin $TiSi_2$ agglomerates at high temperatures, rendering this solution unworkable. Since $CoSi_2$ does not exhibit this problem, it has started to replace $TiSi_2$. $CoSi_2$ transformation occurs at a lower temperature (600 to 700°C). Thus, full formation of low-resistance $CoSi_2$ is achieved before the silicide agglomerates. The Co is also transformed to $CoSi_2$ by a two-step process. In the first, Co and Si are reacted to form CoSi at 450°C. After selectively etching the unreacted Co away, a second thermal step at 700°C reacts the CoSi with more Si to form $CoSi_2$. An even lower resistance cobalt-salicide structure can be obtained by

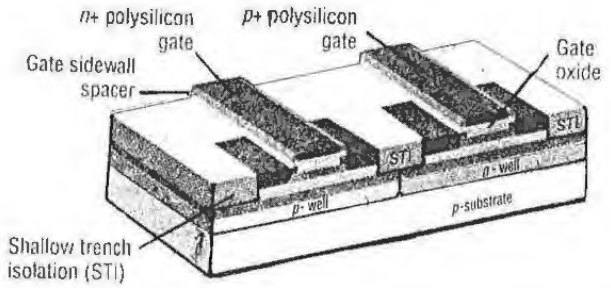


Fig. 16-28 Perspective drawing of a 0.25-μm CMOS structure just prior to the salicide formation process. It depicts the STI isolation structures, the wells, the gates, and the S/D regions. Courtesy of Eaton Corp.

depositing a 20-nm cap layer of TiN onto the Co film before the first thermal step. This TiN cap cuts the sheet resistance from ~10 Ω/sq to about 5 Ω/sq by protecting the Co from oxidizing during the first silicidation anneal.[25]

16.3.8 Premetal Dielectric Deposition and Contact Formation: The premetal dielectric layer is deposited next, much as was done in the first flow. To fill the narrow gaps between adjacent gate structures without voids, a number of dielectric processes may be used including such "sandwich" films as: (1) (PECVD-TEOS/Ozone-TEOS/PECVD-TEOS); or (2) (PECVD-TEOS/SOG/PECVD-TEOS), or an HDP-CVD dielectric film.[26] However, for planarization purposes, instead of performing a flow and reflow, this layer is polished by CMP to provide a flat topography. *Mask #9* is used to pattern the contact hole openings in the premetal dielectric layer. The contact dry-etch process produces vertical-walled contacts, which must be filled (Fig. 16-30). These contact holes can be filled in various ways, including with CVD-W. In that case, a liner film is deposited first (e.g., TiN deposited by sputtering or CVD).

16.3.9 Interconnect Structure Formation: The technologies for fabricating the interconnect structures used in 0.25-μm CMOS were discussed in Chap. 15. A variety of possible materials and processing approaches can be used to implement a particular multilevel interconnect structure. Figures 16-31 and 15-61 show examples of the kinds of multilevel interconnect structures that are used in deep-submicron ICs.

A number of different materials and technologies can be incorporated when establishing the fabrication sequence for such structures. The exact combination chosen will depend on the IC manufacturer and the application that the chip will be serving. The following are some materials and processes that are used to manufacture such structures: (a) Ti/TiN diffusion barriers; (b) Al:Cu films deposited by sputtering; (c) TaN/Cu seed layers (deposited by sputtering) for Cu interconnects; (d) electroplated Cu films; (e) CMP of the dielectric and metal layers; (f) W-plugs (deposited by CVD); (g) low-k intermetal dielectrics; (h) second- or third-generation spin on-glasses (SOG) for gap filling; (i) use of high-density plasmas (HDP) to deposit gap-filling

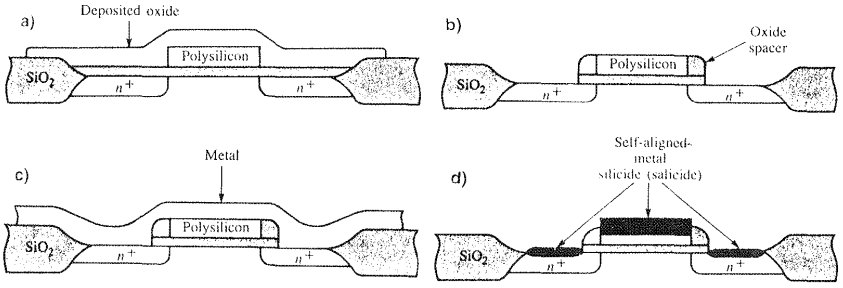


Fig. 16-29 *Salicide* (self-aligned silicide) processing steps.[24] (a) Oxide (or nitride) is deposited following polysilicon definition; (b) structure after etching back, leaving sidewall spacers on the polysilicon; (c) a metal film (Ti or Co) is deposited over the structure and heated. Where the metal is in contact with the silicon substrate, or the polysilicon gate, silicides form; (d) the unreacted metal is selectively etched away, leaving silicide automatically aligned to the gate and source/drain regions. (© 1984 IEEE).

dielectric layers; and (j) single damascene or dual-damascene interconnect structures. Some interconnect structures may use Al films for some interconnect levels, and Cu films for others.[27]

An example of a 0.18-μm CMOS technology using six levels of metals was described in 1998.[28] Both contacts and vias were filled with CVD-W plugs (preceded in each case with a PVD-Ti/CVD-TiN liner film). To make the plugs, the CVD-W was blanket-deposited and then etched back by CMP. The metal interconnect lines consisted of multi-layer films of Ti/TiN/-Al:Cu/TiN to achieve low resistance, good electromigration resistance, and low via resistance. The intermetal dielectric is HDP oxide doped with 5.5% fluorine (which reduces the dielectric constant to 3.5 from the 4.2 for undoped-CVD-SiO_2 films) and is planarized by CMP.

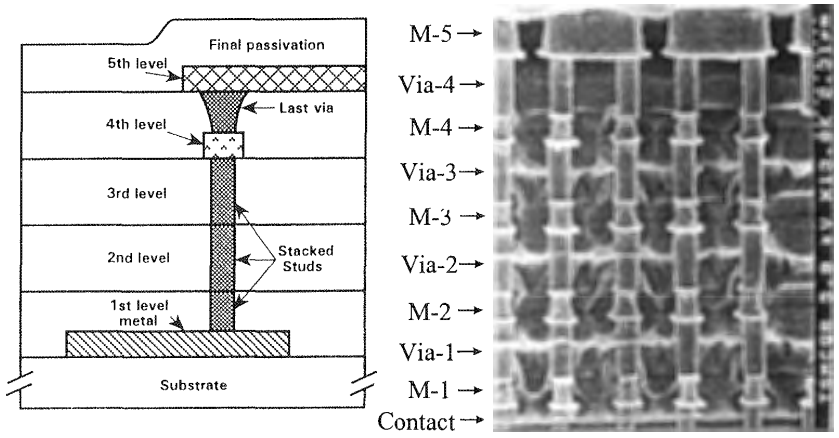


Fig. 16-30 Stacked vias in a 5-metal interconnect, showing filled vertical-walled vias.[15] (© 1998 IEEE).

owing
; (c) a
ith the
away,

Some
s.[27]
ed in
with a
I then
TiN/-
tance.
ectric

E).

REFERENCES

1. J.Y. Chen, *CMOS Devices and Technology for VLSI*, Prentice-Hall, Englewood Cliffs, N.J., 1990.
2. R.S. Muller and T.I. Kamins, *Device Electronics for Integrated Circuits*, 2nd. Ed., New York, John Wiley & Sons, 1986.
3. Y.P. Tsividis, *Operation and Modeling of the MOS Transistor,"* McGraw-Hill, New York, 1987.
4. F.M. Wanlass and C.T. Sah, *IEEE Internat. Solid-State Circuits Conf.*, February 1963.
5. W.C. Till and J.T. Luxton, *Integrated Circuits: Materials, Devices, and Fabrication,"* Prentice-Hall, Englewood Cliffs, N.J., 1982.
6. R. Chwang "CHMOS: An *n*-Well Bulk CMOS Technology for VLSI," *VLSI Design*, 4th Q, 1981, p. 42.
7. B. Hoeflinger and G. Zimmer, in J. Carrol Ed., *Solid-State Devices 1980, Institute of Physics Conf. Series 57*, from the 10th European Solid-State Device Research Conference.
8. L.C. Parrilllo *et al.*, "Twin-Tub CMOS -II," *Tech. Dig. IEDM*, 1982, p. 706.
9. E. Kooi and J.A. Appels, in *Semiconductor Silicon, 1973*, H.R. Huff and R. Burgess, Eds. P. 860, The Electrochemical Society Symposium Series, Princeton, NJ (1973).
10. T. Ohzone *et al.*, *IEEE Trans. Electron Dev.*, **ED-32**, 1789, (1980).
11. W.S. Chang *et al.*, "A High-Performance 0.25 μm CMOS Technology: I–Design and Characterization," *IEEE Trans. Electron Dev.*, **Vol. 39**, April 1992, p. 959.
12. W.S. Chang *et al.*, "A High-Performance 0.25 μm CMOS Technology: II – Technology," *IEEE Trons. Electron Dev.*, **Vol. 39**, April 1992, p. 967.
13. M.S.C. Luo *et al.*, "A 0.25 μm CMOS Technology with 45 Å NO-nitrided Oxide," *Tech. Dig. IEDM*, 1995, p. 691.
14. T.C. Holloway *et al.*, "0.18 μm CMOS Technology for High-Performance, Low Power, and RF Applications," *Dig. of Tech Papers, 1997 Symp. on VLSI Technology*, p. 13.
15. J.Y.C. Sun *et al.*, "Foundry Technology for the Next Decade," *Tech. Dig. IEDM* 1998, p. 321.
16. L. Rubin, "Buried Layer Substrates: Economically Enhancing Device Performance," *Solid State Technology*, June 1999, p. 95.
17. L. Peters, "Choices and Challenges for Shallow Trench Isolation," *Semiconductor International.*, April 1999, p. 69.
18. Y. Okada and P.J. Tobin, "Hot-Carrier Degradation of LDD MOSFETs with Gate Oxynitride Grown in N_2O," *IEEE Electron Dev. Letts.*, **EDL-15**, July 1994, p. 233.
19. C.T. Liu *et al.*, "High Performance 0.2 μm CMOS with 25Å Gate Oxide Grown on Nitrogen Implanted Silicon," *Tech. Dig. IEDM*, 1996, p. 499.
20. S.V. Hattangady *et al.*, "Ultrathin Nitrogen-Profile-Engineered Gate Dielectric Films," *Tech. Dig. IEDM*, 1996, p. 495.
21. P. Packan et al., "Modeling Solid Source Boron Diffusion for Advanced Transistor Applications," *Tech. Dig. IEDM*, 1998, p. 505.
22. J. Schmitz et al., "Ultra-Shallow Junction Formation by Outdiffusion From Implanted Oxide," *Tech. Dig. IEDM*, 1998, p. 1009.
23. J.A. Kittl *et al.*, "Salicides and Alternative Technologies for Future ICs," *Solid State Technology*, June 1999, p. 81.
24. C.Y. Ting, "Silicide for Contacts and Interconnects," *Tech. Dig. IEDM*, 1984, p. 110.
25. T. Yamazaki *et al.*, "21 psec Switching 0.1 μm-CMOS at Room Temperature using High-Performance Co Salicide Process," *Tech. Dig. IEDM*, 1993, p. 906.
26. S. Nag *et al.*, "Low-Temperature Pre-Metal Dielectrics for Future ICs," *Solid State Technology*,

September 1998, p. 69.
27. M. Iragashi *et al.*, "The Best Combination of Aluminum and Copper Interconnects for a High Performance 0.18 μm CMOS Logic Device," *Tech. Dig. IEDM*, 1998, p. 829.
28. S. Yang *et al.*, "A High Performance 180-nm Generation Logic Technology," *Tech. Dig. IEDM*, 1998, p. 197.

PROBLEMS

1. Assume you were asked to design a MOS capacitor to act as a test structure for monitoring the threshold voltage of a silicon-gate NMOSFET. Formulate a detailed process flow of the fabrication steps needed to manufacture such an *n*-channel MOS capacitor.

2. Compare retrograde-well CMOS and conventional-well CMOS. What are the most important differences in the process technology and the finished device structure? List four advantages that retrograde well technology provides when compared to conventional-well CMOS technology.

3. (a) Why is the <100> orientation preferred for the Si wafers used to fabricate MOSFET ICs? (b) What are the disadvantages of using a field oxide that is too thin in CMOS ICs? (c) List three functions that the BPSG interlevel dielectric layer performs in CMOS ICs?

4. The following are some of the new technologies used in ULSI CMOS technologies (i.e., technologies in which the minimum feature size is 0.25 μm, or smaller. (a) shallow trench isolation; (b) dual-doped polysilicon; (c) $TiSi_2$ salicides; (d) shallow source/drain junctions; (e) copper interconnects; (f) low-k dielectrics. List at least two advantages and two problems associated with each of these technologies.

5. In an *n*-well 1 μm CMOS IC technology, the starting material is a *p*-type Si wafer with <100> orientation. The wafer may be either lightly doped bulk *p*-type Si, or heavily *p*-doped bulk Si with a lightly *p*-doped epi layer on its surface. In either case the resistivity of the Si near the surface is about 8 Ω-cm. The well doping in the *n*-well near the Si surface is $3x10^{16}/cm^3$. The gate material of the MOSFETs is a heavily *n*-doped polysilicon film. The source and drain regions of the NMOSFETs are formed with an As implant of $1x10^{16}$ As ions/cm^2 at 50 keV. A channel implant with $8x10^{11}$ boron ions/cm^2 is performed through a gate oxide that is 15 nm thick. (a) What would the threshold voltages of the NMOSFETs and PMOSFETs be if the devices were fabricated without implanting the channel with the boron ions? (b) Calculate the threshold voltage that the NMOSFETs and PMOSFETs would exhibit as a result of implanting the boron ions into the channel? (c) Draw a doping profile along the coordinate perpendicular to the surface and passing through (i) the channel region; and (ii) the drain region of both types of FET.

6. Discuss each of the following drawbacks of LOCOS technology: (a) bird's beak encroachment; (b) oxide thinning in narrow field-oxide openings; (c) boron encroachment; (d) non-planarity; (e) process complexity; (f) sensitivity of bird's beak encroachment on pattern geometry.