

2ND EDITION

Semiconductor Devices

Physics and Technology

S. M. SZE

UMC Chair Professor

National Chiao Tung University

National Nano Device Laboratories

Hsinchu, Taiwan



JOHN WILEY & SONS, INC.

Sony Exhibit 1044
Sony v. VisionX

Acquisitions Editor *William Zobrist*
Marketing Manager *Katherine Hepburn*
Production Services Manager *Jeanine Furino*
Production Editor *Sandra Russell*
Designer *Harold Nolan*
Production Management Services *Argosy Publishing Services*

Cover Photography: A transmission-electron micrograph of a floating-gate nonvolatile semiconductor memory with a magnification of 100,000 times. (Photography courtesy of George T. T. Sheng.) For a discussion of the device, see Chapters 1, 6, and 14.

This book was typeset in *New Caledonia* by *Argosy Publishing* and printed and bound by *R. R. Donnelley and Sons, Inc. (Willard)*. The cover was printed by *The Lehigh Press*.

The paper in this book was manufactured by a mill whose forest management programs include sustained yield harvesting of its timberlands. Sustained yield harvesting principles ensure that the number of trees cut each year does not exceed the amount of new growth.

The book is printed on acid-free paper. 

Copyright © 1985, 2002 by John Wiley & Sons, Inc. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (508) 750-8400, fax (508) 750-4470. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc. 605 Third Avenue, New York, NY 10158-0012, (212) 850-6008, E-mail: PERMREQ@WILEY.COM. To order books or for customer service call 1-800-CALL-WILEY (225-5945).

Library of Congress Cataloging in Publication Data:
Sze, S. M., 1936-

Semiconductor devices, physics and technology/S.M. Sze.—2nd ed.
p. cm.

Includes bibliographical references and index.

ISBN 0-471-33372-7 (cloth: alk. paper)

1. Semiconductors. I. Title.

TK7871.85 .S9883 2001

621.3815'2—dc21

2001026003

ISBN 0-471-33372-7

Printed in the United States of America

10 9 8 7 6 5 4 3

p – n Junction

-
- ▶ 4.1 BASIC FABRICATION STEPS
 - ▶ 4.2 THERMAL EQUILIBRIUM CONDITION
 - ▶ 4.3 DEPLETION REGION
 - ▶ 4.4 DEPLETION CAPACITANCE
 - ▶ 4.5 CURRENT-VOLTAGE CHARACTERISTICS
 - ▶ 4.6 CHARGE STORAGE AND TRANSIENT BEHAVIOR
 - ▶ 4.7 JUNCTION BREAKDOWN
 - ▶ 4.8 HETEROJUNCTION
 - ▶ SUMMARY
-

In the preceding chapters we considered the carrier concentrations and transport phenomena in homogeneous semiconductor materials. In this chapter we discuss the behavior of single-crystal semiconductor material containing both p - and n -type regions that form a p – n junction. Most modern p – n junctions are made by planar technology to be described in Section 4.1.

A p – n junction serves an important role both in modern electronic applications and in understanding other semiconductor devices. It is used extensively in rectification, switching, and other operations in electronic circuits. It is a key building block for the bipolar transistor and thyristor (Chapter 5), as well as for metal-oxide-semiconductor field-effect transistors (MOSFETs) (Chapter 6). Given proper biasing conditions or when exposed to light, p – n junction also functions as either microwave (Chapter 8) or photonic device (Chapter 9).

We also consider a related device—the heterojunction, which is a junction formed between two dissimilar semiconductors. It has many unique features that are not readily available from the conventional p – n junctions. The heterojunction is an important building block for heterojunction bipolar transistors (Chapter 5), modulation doped field-effect transistors (Chapter 7), quantum-effect devices (Chapter 8), and photonic devices (Chapter 9).

Specifically, we cover the following topics:

- The formation of a p – n junction physically and electrically.
- The behavior of the junction depletion layer under voltage biases.
- The current transport in a p – n junction and the influence of the generation and recombination processes.
- The charge storage in a p – n junction and its influence on the transient behavior.

- The avalanche multiplication in a *p-n* junction and its impact on the maximum reverse voltage.
- The heterojunction and its basic characteristics.

▶ 4.1 BASIC FABRICATION STEPS

Today, planar technology is used extensively for integrated circuit (IC) fabrication. Figures 1 and 2 show the major steps of a planar process. These steps include oxidation, lithography,

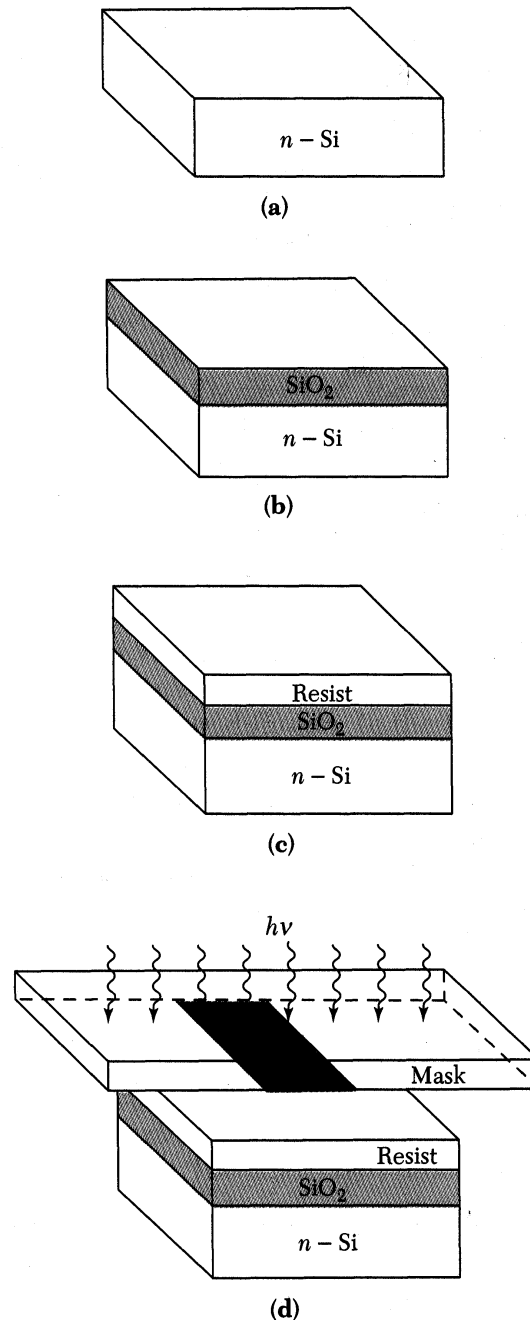


Fig. 1 (a) A bare *n*-type Si wafer. (b) An oxidized Si wafer by dry or wet oxidation. (c) Application of resist. (d) Resist exposure through the mask.

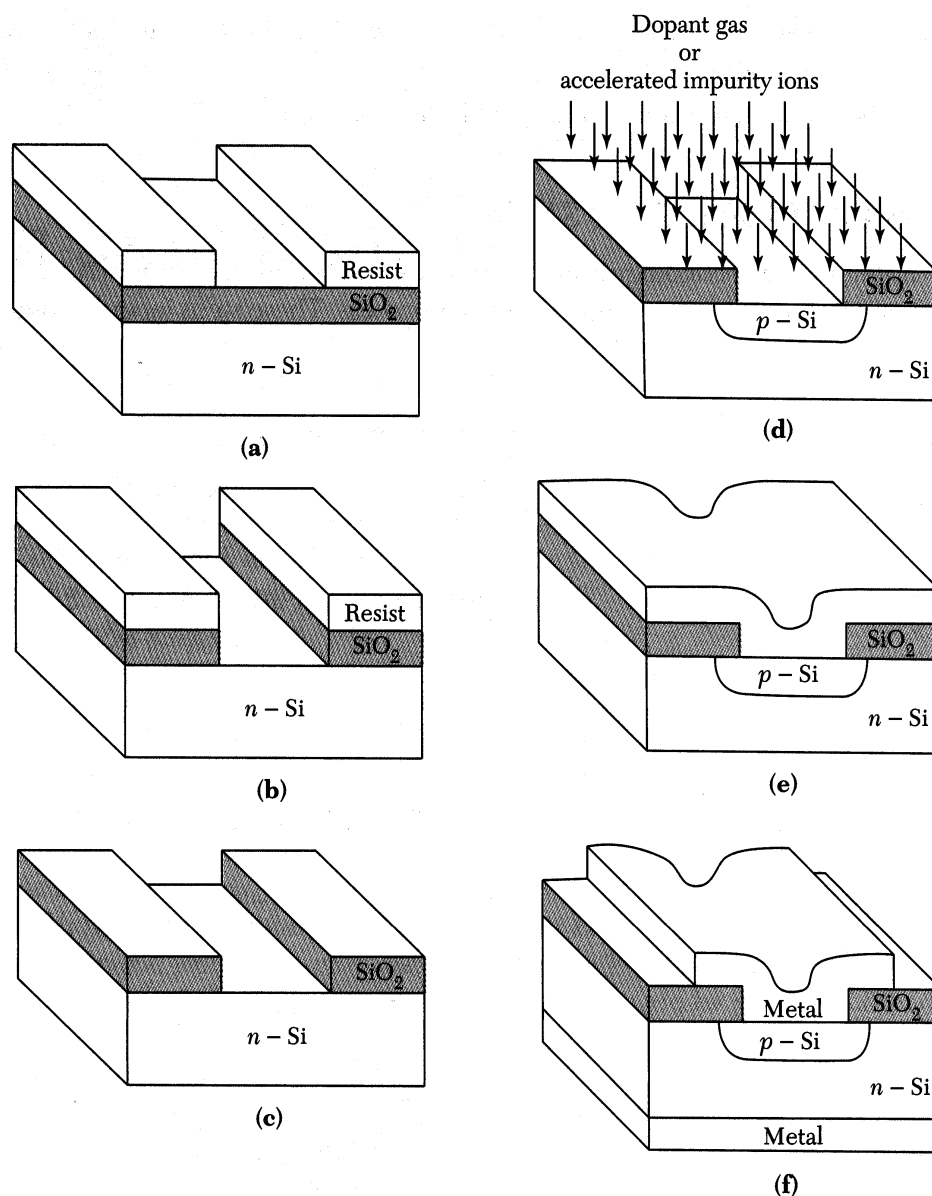


Fig. 2 (a) The wafer after the development. (b) The wafer after SiO_2 removal. (c) The final result after a complete lithography process. (d) A *p-n* junction is formed in the diffusion or implantation process. (e) The wafer after metallization. (f) A *p-n* junction after the complete processes.

ion implanatation, and metallization. We describe these steps briefly in this section. More detailed discussions can be found in Chapters 10–14.

4.1.1 Oxidation

The development of a high-quality silicon dioxide (SiO_2) has helped to establish the dominance of Si in the production of commercial ICs. Generally, SiO_2 functions as an insulator in a number of device structures or as a barrier to diffusion or implantation during device fabrication. In the fabrication of a *p-n* junction (Fig. 1), the SiO_2 film is used to define the junction area.

There are two SiO_2 growth methods, dry and wet oxidation, depending on whether dry oxygen or water vapor is used. Dry oxidation is usually used to form thin oxides in a

device structure because of its good Si-SiO₂ interface characteristics, whereas wet oxidation is used for thicker layers because of its higher growth rate. Figure 1a shows a section of a bare Si wafer ready for oxidation. After the oxidation process, a SiO₂ layer is formed all over the wafer surface. For simplicity, Fig. 1b shows only the upper surface of an oxidized wafer.

4.1.2 Lithography

Another technology, called photolithography, is used to define the geometry of the p - n junction. After the formation of SiO₂, the wafer is coated with an ultraviolet (UV)-light-sensitive material called a photoresist, which is spun on the wafer surface by a high-speed spinner. After spinner (Fig. 1c), the wafer is baked at about 80°–100°C to drive the solvent out of the resist and to harden the resist for improved adhesion. Figure 1d shows the next step, which is to expose the wafer through a patterned mask using an UV-light source. The exposed region of the photoresist-coated wafer undergoes a chemical reaction depending on the type of resist. The area exposed to light become polymerized and difficult to remove in an etchant.* The polymerized region remains when the wafer is placed in a developer, whereas the unexposed region (under the opaque area) dissolves and washes away. Figure 2a shows the wafer after the development. The wafer is again baked to 120°–180°C for 20 minutes to enhance the adhesion and improve the resistance to the subsequent etching process. Then, an etch using buffered hydrofluoric acid (HF) removes the unprotected SiO₂ surface (Fig. 2b). Last, the resist is stripped away by a chemical solution or an oxygen plasma system. Figure 2c shows the final result of a region without oxide (a window) after the lithography process. The wafer is now ready for forming the p - n junction by a diffusion or ion-implantation process.

4.1.3 Diffusion and Ion Implantation

In the diffusion method, the semiconductor surface not protected by the oxide is exposed to a source with a high concentration of opposite-type impurity. The impurity moves into the semiconductor crystal by solid-state diffusion. In the ion-implantation method, the intended impurity is introduced into the semiconductor by accelerating the impurity ions to a high-energy level and then implanting the ions in the semiconductor. The SiO₂ layer serves as barrier to impurity diffusion or ion implantation. After the diffusion or implantation process, the p - n junction is formed as shown in Fig. 2d. Due to lateral diffusion of impurities or lateral straggle of implanted ions, the width of the p region is slightly wider than the window opening.

4.1.4 Metallization

After diffusion or ion-implantation process, a metallization process is used to form ohmic contacts and interconnections (Fig. 2e). Metal films can be formed by physical vapor deposition and chemical vapor deposition. The lithography process is again used to define the front contact which is shown in Fig. 2f. A similar metallization step is done on the back contact without using a lithography process. Normally a low-temperature ($\leq 500^\circ\text{C}$) anneal would also be performed to promote low-resistance contacts between the metal layers

* This is a negative photoresist. We can also use positive photoresist. Lithography and photoresists are considered in detail in Chapter 12.

and the semiconductor. With the completion of the metallization, the *p-n* junctions become functional.

► 4.2 THERMAL EQUILIBRIUM CONDITION

The most important characteristic of *p-n* junctions is that they rectify, that is, they allow current to flow easily in only one direction. Figure 3 shows the current-voltage characteristics of a typical silicon *p-n* junction. When we apply “forward bias” to the junction (i.e., positive voltage on the *p*-side), the current increases rapidly as the voltage increases. However, when we apply a “reverse bias,” virtually no current flows initially. As the reverse bias is increased the current remains very small until a critical voltage is reached, at which point the current suddenly increases. This sudden increase in current is referred to as the junction breakdown. The applied forward voltage is usually less than 1 V, but the reverse critical voltage, or breakdown voltage, can vary from just a few volts to many thousands of volts depending on the doping concentration and other device parameters.

4.2.1 Band Diagram

In Fig. 4a, we see two regions of *p*- and *n*-type semiconductor materials that are uniformly doped and physically separated before the junction is formed. Note that the Fermi level E_F is near the valence band edge in the *p*-type material and near the conduction band edge in the *n*-type material. While *p*-type material contains a large concentration of holes with few electrons, the opposite is true for *n*-type material.

When the *p*- and *n*-type semiconductors are joined together, the large carrier concentration gradients at the junction cause carrier diffusion. Holes from the *p*-side diffuse into the *n*-side, and electrons from the *n*-side diffuse into the *p*-side. As holes continue to leave the *p*-side, some of the negative acceptor ions (N_A^-) near the junction are left uncompensated, since the acceptors are fixed in the semiconductor lattice, whereas the holes are mobile. Similarly, some of the positive donor ions (N_D^+) near the junction are left uncompensated as the electrons leave the *n*-side. Consequently, a negative space charge

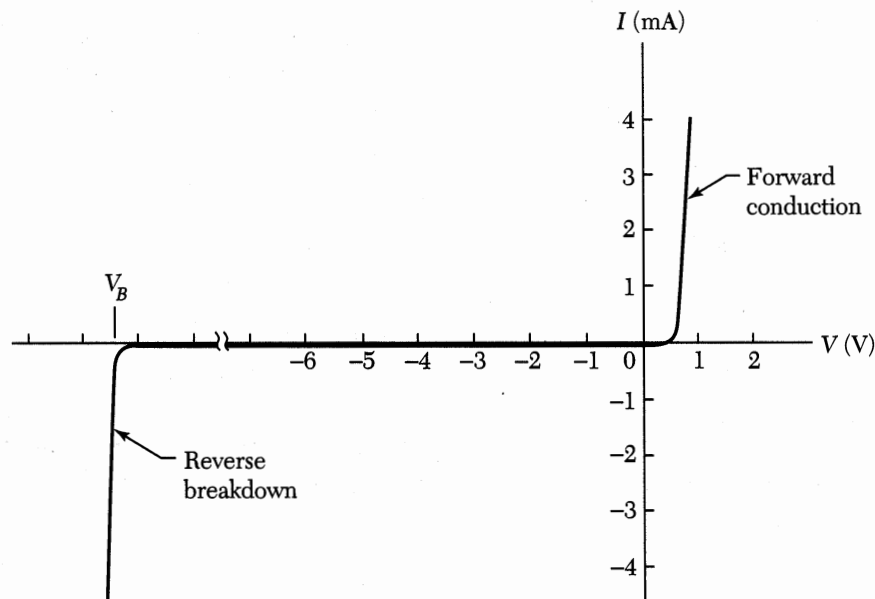


Fig. 3 Current-voltage characteristics of a typical silicon *p-n* junction.

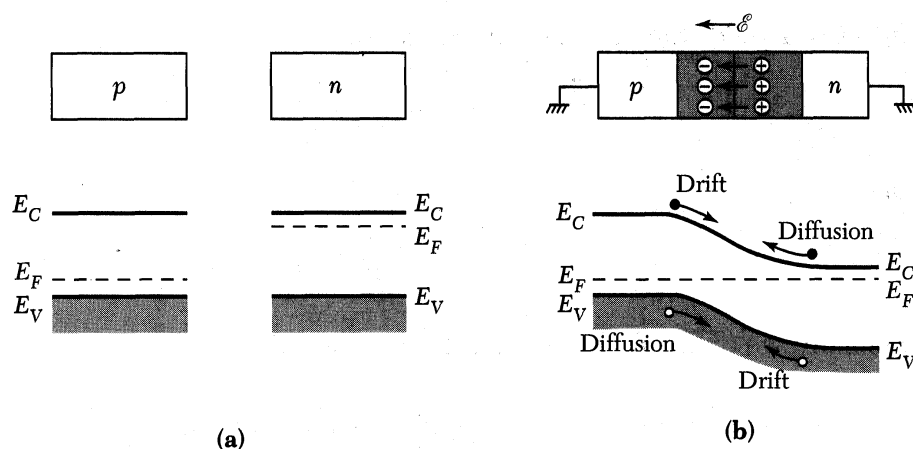


Fig. 4 (a) Uniformly doped *p*-type and *n*-type semiconductors before the junction is formed. (b) The electric field in the depletion region and the energy band diagram of a *p-n* junction in thermal equilibrium.

forms near the *p*-side of the junction and a positive space charge forms near the *n*-side. This space charge region creates an electric field that is directed from the positive charge toward the negative charge, as indicated in the upper illustration of Fig. 4b.

The electric field is in the direction opposite to the diffusion current for each type of charge carrier. The lower illustration of Fig. 4b shows that the hole diffusion current flows from left to right, whereas the hole drift current due to the electric field flows from right to left. The electron diffusion current also flows from left to right, whereas the electron drift current flows in the opposite direction. Note that because of their negative charge, electrons diffuse from right to left, opposite the direction of electron current.

4.2.2 Equilibrium Fermi Levels

At thermal equilibrium, that is, the steady-state condition at a given temperature without any external excitations, the individual electron and hole current flowing across the junctions are identically zero. Thus, for each type of carrier the drift current due to the electric field must exactly cancel the diffusion current due to the concentration gradient. From Eq. 32 in Chapter 3,

$$\begin{aligned}
 J_p &= J_p(\text{drift}) + J_p(\text{diffusion}) \\
 &= q\mu_p p \mathcal{E} - qD_p \frac{dp}{dx} \\
 &= q\mu_p p \left(\frac{1}{q} \frac{dE_i}{dx} \right) - kT\mu_p \frac{dp}{dx} = 0,
 \end{aligned} \tag{1}$$

where we have used Eq. 8 of Chapter 3 for the electric field and the Einstein relation $D_p = (kT/q)\mu_p$. Substituting the expression for hole concentration

$$p = n_i e^{(E_i - E_F)/kT} \tag{2}$$

and its derivative

$$\frac{dp}{dx} = \frac{p}{kT} \left(\frac{dE_i}{dx} - \frac{dE_F}{dx} \right) \tag{3}$$

into Eq. 1 yields the net hole current density

$$J_p = \mu_p p \frac{dE_F}{dx} = 0 \quad (4)$$

or

$$\frac{dE_F}{dx} = 0. \quad (5)$$

Similarly, we obtain for the net electron current density

$$\begin{aligned} J_n &= J_n(\text{drift}) + J_n(\text{diffusion}) \\ &= q\mu_n n \mathcal{E} + qD_n \frac{dn}{dx} \\ &= \mu_n n \frac{dE_F}{dx} = 0. \end{aligned} \quad (6)$$

Thus, for the condition of zero net electron and hole currents, the Fermi level must be constant (i.e., independent of x) throughout the sample as illustrated in the energy band diagram of Fig. 4*b*.

The constant Fermi level required at thermal equilibrium results in a unique space charge distribution at the junction. We repeat the one-dimensional *p-n* junction and the corresponding equilibrium energy band diagram in Figs. 5*a* and 5*b*, respectively. The unique space charge distribution and the electrostatic potential ψ are given by Poisson's equation:

$$\frac{d^2\psi}{dx^2} \equiv -\frac{d\mathcal{E}}{dx} = -\frac{\rho_s}{\epsilon_s} = -\frac{q}{\epsilon_s} (N_D - N_A + p - n). \quad (7)$$

Here we assume that all donors and acceptors are ionized.

In regions far away from the metallurgical junction, charge neutrality is maintained and the total space charge density is zero. For these neutral regions we can simplify Eq. 7 to

$$\frac{d^2\psi}{dx^2} = 0 \quad (8)$$

and

$$N_D - N_A + p - n = 0. \quad (9)$$

For a *p*-type neutral region, we assume $N_D = 0$ and $p \gg n$. The electrostatic potential of the *p*-type neutral region with respect to the Fermi level, designated as ψ_p in Fig. 5*b*, can be obtained by setting $N_D = n = 0$ in Eq. 9 and by substituting the result ($p = N_A$) into Eq. 2:

$$\psi_p \equiv -\frac{1}{q} (E_i - E_F) \Big|_{x \leq -x_p} = -\frac{kT}{q} \ln \left(\frac{N_A}{n_i} \right). \quad (10)$$

Similarly, we obtain the electrostatic potential of the *n*-type neutral region with respect to the Fermi level:

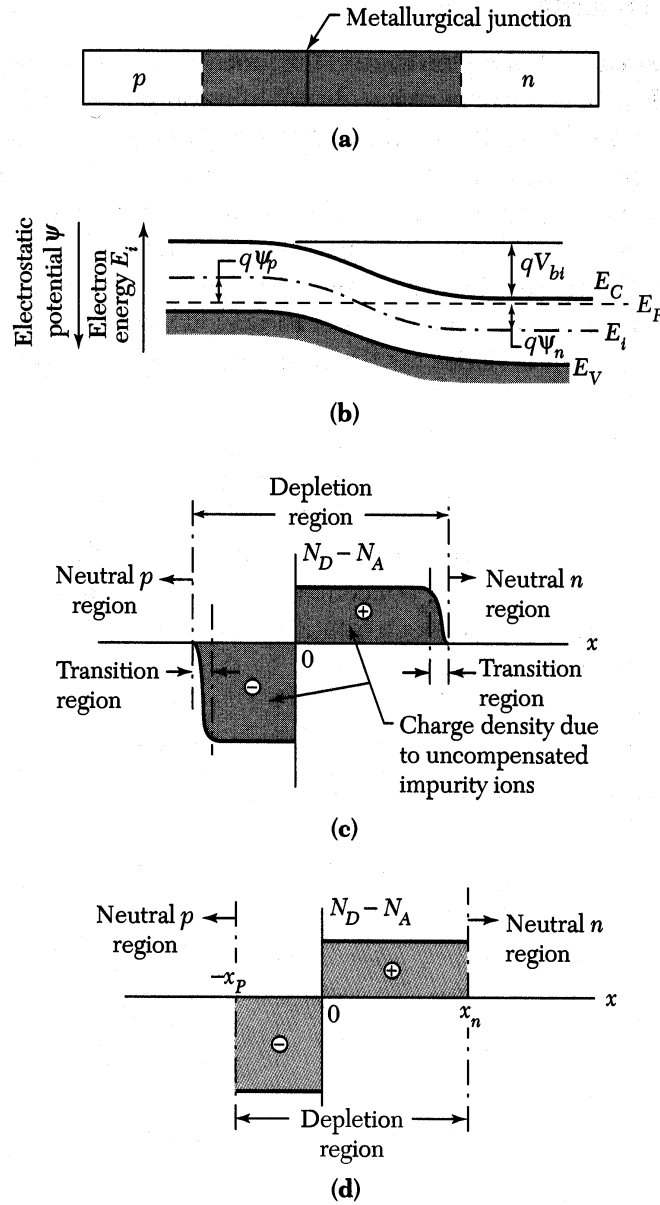


Fig. 5 (a) A *p-n* junction with abrupt doping changes at the metallurgical junction. (b) Energy band diagram of an abrupt junction at thermal equilibrium. (c) Space charge distribution. (d) Rectangular approximation of the space charge distribution.

$$\psi_n \equiv -\frac{1}{q}(E_i - E_F) \Big|_{x \geq x_n} = -\frac{kT}{q} \ln \left(\frac{N_D}{n_i} \right). \quad (11)$$

The total electrostatic potential difference between the *p*-side and the *n*-side neutral regions at thermal equilibrium is called the *built-in potential* V_{bi} :

$$V_{bi} = \psi_n - \psi_p = \frac{kT}{q} \ln \left(\frac{N_A N_D}{n_i^2} \right). \quad (12)$$

4.2.3 Space Charge

Moving from a neutral region toward the junction, we encounter a narrow transition region, shown in Fig. 5c. Here the space charge of impurity ions is partially compensated by the mobile carriers. Beyond the transition region we enter the completely depleted region where the mobile carrier densities are zero. This is called the *depletion region* (also called the *space-charge region*). For typical *p-n* junctions in silicon and gallium arsenide, the width of each transition region is small compared with the width of the depletion region. Therefore, we can neglect the transition region and represent the depletion region by the rectangular distribution shown in Fig. 5d, where x_p and x_n denote the depletion layer widths of the *p*- and *n*-sides for the completely depleted region with $p = n = 0$. Equation 7 becomes

$$\frac{d^2\psi}{dx^2} = \frac{q}{\epsilon_s} (N_A - N_D). \quad (13)$$

The magnitudes of $|\psi_p|$ and ψ_n as calculated from Eqs. 10 and 11 are plotted in Fig. 6 as a function of the doping concentration of silicon and gallium arsenide. For a given doping concentration, the electrostatic potential of gallium arsenide is higher because of its smaller intrinsic concentration n_i .

► EXAMPLE 1

Calculate the built-in potential for a silicon *p-n* junction with $N_A = 10^{18} \text{ cm}^{-3}$ and $N_D = 10^{15} \text{ cm}^{-3}$ at 300 K.

SOLUTION From Eq. 12 we obtain

$$V_{bi} = (0.0259) \ln \left[\frac{10^{18} \times 10^{15}}{(9.65 \times 10^9)^2} \right] = 0.774 \text{ V}.$$

Also from Fig. 6,

$$V_{bi} = \psi_n + |\psi_p| = 0.30 \text{ V} + 0.47 \text{ V} = 0.77 \text{ V}.$$

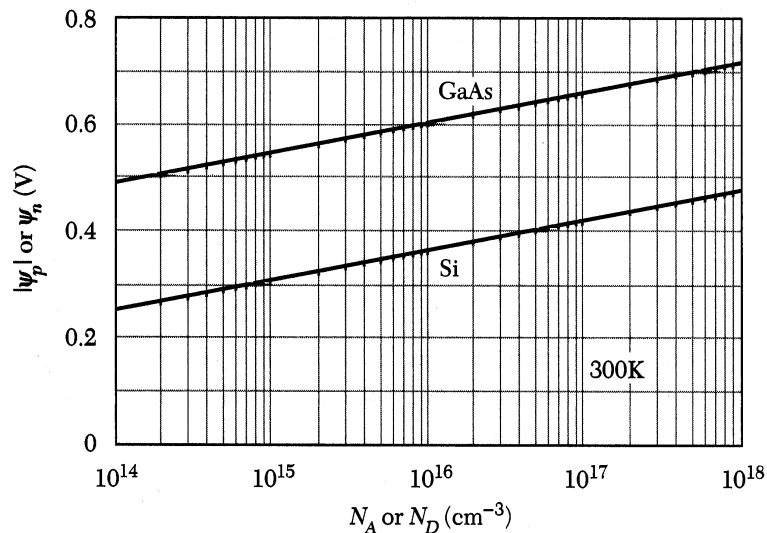


Fig. 6 Built-in potentials on the *p*-side and *n*-side of abrupt junctions in Si and GaAs as a function of impurity concentration.

4.3 DEPLETION REGION

To solve Poisson's equation, Eq.13, we must know the impurity distribution. In this section we consider two important cases—the abrupt junction and the linearly graded junction. Figure 7a shows an *abrupt junction*, that is, a *p-n* junction formed by shallow diffusion or low-energy ion implantation. The impurity distribution of the junction can be approximated by an abrupt transition of doping concentration between the *n*- and *p*-type regions. Figure 7b shows a linearly graded junction. For either deep diffusions or high-energy ion implantations, the impurity profiles may be approximated by linearly graded junctions, that is, the impurity distribution varies linearly across the junction. We consider the depletion regions of both types of junction.

4.3.1 Abrupt Junction

The space charge distribution of an abrupt junction is shown in Fig. 8a. In the depletion region, free carriers are totally depleted so that Poisson's equation, Eq.13, simplifies to

$$\frac{d^2\psi}{dx^2} = +\frac{qN_A}{\epsilon_s} \quad \text{for} \quad -x_p \leq x < 0, \quad (14a)$$

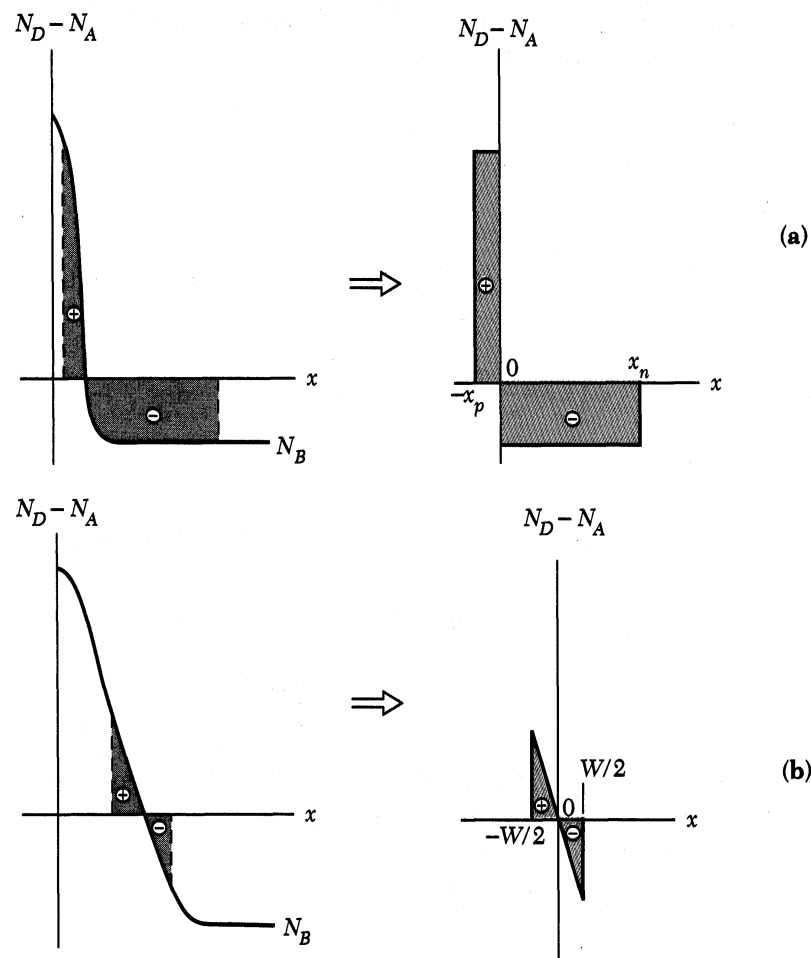


Fig. 7 Approximate doping profiles. (a) Abrupt junction. (b) Linearly graded junction.

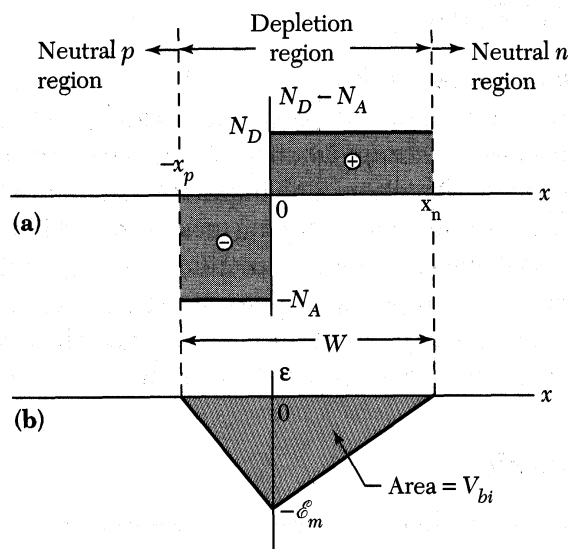


Fig. 8 (a) Space charge distribution in the depletion region at thermal equilibrium. (b) Electric-field distribution. The shaded area corresponds to the built-in potential.

$$\frac{d^2\psi}{dx^2} = -\frac{qN_D}{\epsilon_s} \quad \text{for } 0 < x \leq x_n. \quad (14b)$$

The overall space charge neutrality of the semiconductor requires that the total negative space charge per unit area in the *p*-side must precisely equal the total positive space charge per unit area in the *n*-side:

$$N_A x_p = N_D x_n. \quad (15)$$

The total depletion layer width W is given by

$$W = x_p + x_n. \quad (16)$$

The electric field shown in Fig. 8b is obtained by integrating Eqs. 14a and 14b, which gives

$$\mathcal{E}(x) = -\frac{d\psi}{dx} = -\frac{qN_A(x+x_p)}{\epsilon_s} \quad \text{for } -x_p \leq x < 0 \quad (17a)$$

and

$$\mathcal{E}(x) = -\mathcal{E}_m + \frac{qN_D x}{\epsilon_s} = \frac{qN_D}{\epsilon_s} (x - x_n) \quad \text{for } 0 < x \leq x_n, \quad (17b)$$

where $\mathcal{E}_m$ is the maximum field that exists at $x = 0$ and is given by

$$\mathcal{E}_m = \frac{qN_D x_n}{\epsilon_s} = \frac{qN_A x_p}{\epsilon_s}. \quad (18)$$

Integrating Eqs. 17a and 17b over the depletion region gives the total potential variation, namely, the built-in potential V_{bi} :

$$\begin{aligned} V_{bi} &= -\int_{-x_p}^{x_n} \mathcal{E}(x) dx = -\int_{-x_p}^0 \mathcal{E}(x) dx \Big|_{p\text{-side}} - \int_0^{x_n} \mathcal{E}(x) dx \Big|_{n\text{-side}} \\ &= \frac{qN_A x_p^2}{2\epsilon_s} + \frac{qN_D x_n^2}{2\epsilon_s} = \frac{1}{2} \mathcal{E}_m W. \end{aligned} \quad (19)$$

Therefore, the area of the field triangle in Fig. 8b corresponds to the built-in potential.

Combining Eqs. 15 to 19 gives the total depletion layer width as a function of the built-in potential,

$$W = \sqrt{\frac{2\epsilon_s}{q} \left(\frac{N_A + N_D}{N_A N_D} \right) V_{bi}}. \quad (20)$$

When the impurity concentration on one side of an abrupt junction is much higher than that of the other side, the junction is called a *one-sided abrupt junction* (Fig. 9a). Figure 9b shows the space charge distribution of a one-sided abrupt p^+n junction, where $N_A \gg N_D$. In this case, the depletion layer width of the p -side is much smaller than the n -side (i.e., $x_p \ll x_n$), and the expression for W can be simplified to

$$W \cong x_n = \sqrt{\frac{2\epsilon_s V_{bi}}{qN_D}}. \quad (21)$$

The expression for the electric-field distribution is the same as Eq. 17b:

$$\mathcal{E}(x) = -\mathcal{E}_m + \frac{qN_B x}{\epsilon_s}, \quad (22)$$

where N_B is the lightly doped bulk concentration (i.e., N_D for a p^+n junction). The field decreases to zero at $x = W$. Therefore,

$$\mathcal{E}_m = \frac{qN_B W}{\epsilon_s} \quad (23)$$

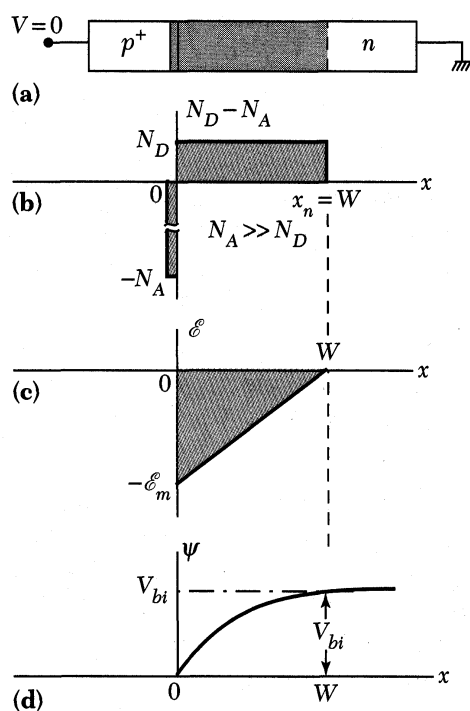


Fig. 9 (a) One-sided abrupt junction (with $N_A \gg N_D$) in thermal equilibrium. (b) Space charge distribution. (c) Electric-field distribution. (d) Potential distribution with distance, where V_{bi} is the built-in potential.

and

$$\mathcal{E}(x) = \frac{qN_B}{\epsilon_s}(-W + x) = -\mathcal{E}_m \left(1 - \frac{x}{W}\right), \quad (24)$$

which is shown in Fig. 9c.

Integrating Poisson's equation once more gives the potential distribution

$$\psi(x) = -\int_0^x \mathcal{E} dx = \mathcal{E}_m \left(x - \frac{x^2}{2W}\right) + \text{constant}. \quad (25)$$

With zero potential in the neutral *p*-region as a reference, or $\psi(0) = 0$, and employing Eq. 19,

$$\psi(x) = \frac{V_{bi}x}{W} \left(2 - \frac{x}{W}\right). \quad (26)$$

The potential distribution is shown in Fig. 9d.

► EXAMPLE 2

For a silicon one-sided abrupt junction with $N_A = 10^{19} \text{ cm}^{-3}$ and $N_D = 10^{16} \text{ cm}^{-3}$, calculate the depletion layer width and the maximum field at zero bias ($T = 300 \text{ K}$).

SOLUTION From Eqs. 12, 21, and 23, we obtain

$$\begin{aligned} V_{bi} &= 0.0259 \ln \left[\frac{10^{19} \times 10^{16}}{(9.65 \times 10^9)^2} \right] = 0.895 \text{ V}, \\ W &\equiv \sqrt{\frac{2\epsilon_s V_{bi}}{qN_D}} = 3.41 \times 10^{-5} = 0.343 \text{ } \mu\text{m}, \\ \mathcal{E}_m &= \frac{qN_B W}{\epsilon_s} = 0.52 \times 10^4 \text{ V/cm}. \end{aligned}$$

The previous discussions are for a *p-n* junction at thermal equilibrium without external bias. The equilibrium energy band diagram, shown again in Fig. 10a, illustrates that the total electrostatic potential across the junction is V_{bi} . The corresponding potential energy difference from the *p*-side to the *n*-side is qV_{bi} . If we apply a positive voltage V_F to the *p*-side with respect to the *n*-side, the *p-n* junction becomes forward-biased, as shown in Fig. 10b. The total electrostatic potential across the junction decreases by V_F , that is, it is replaced with $V_{bi} - V_F$. Thus, forward bias reduces the depletion layer width.

By contrast, as shown in Fig. 10c, if we apply positive voltage V_R to the *n*-side with respect to the *p*-side, the *p-n* junction now becomes reverse-biased and the total electrostatic potential across the junction increases by V_R , that is, it is replaced by $V_{bi} + V_R$. Here, we find that reverse bias increases the depletion layer width. Substituting these voltage values in Eq. 21 yields the depletion layer widths as a function of the applied voltage for a one-sided abrupt junction:

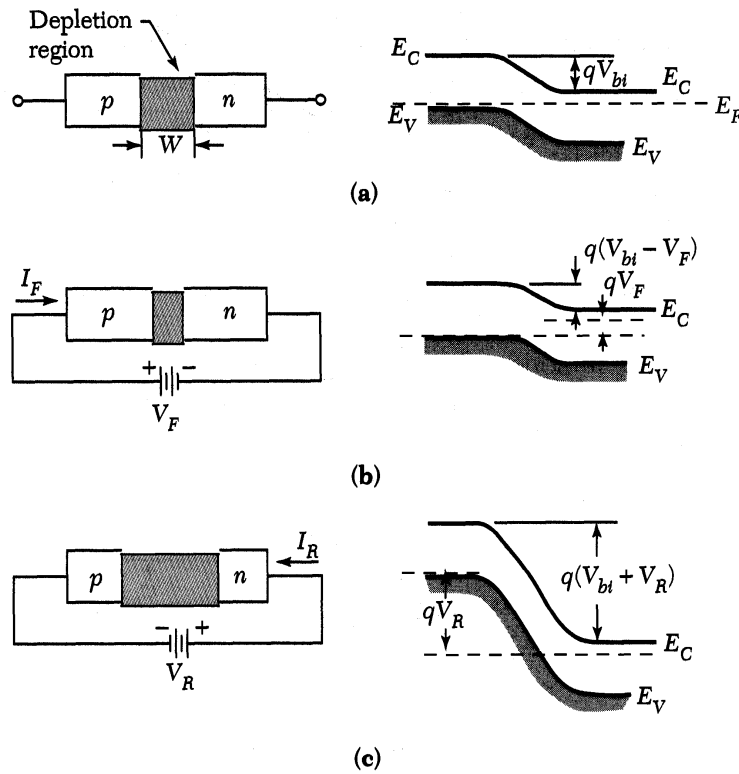


Fig. 10 Schematic representations of depletion layer width and energy band diagrams of a *p-n* junction under various biasing conditions. (a) Thermal-equilibrium condition. (b) Forward-bias condition. (c) Reverse-bias condition.

$$W = \sqrt{\frac{2\epsilon_s(V_{bi} - V)}{qN_B}}, \quad (27)$$

where N_B is the lightly doped bulk concentration, and V is positive for forward bias and negative for reverse bias. Note that the depletion layer width W varies as the square root of the total electrostatic potential difference across the junction.

4.3.2 Linearly Graded Junction

We first consider the case of thermal equilibrium. The impurity distribution for a linearly graded junction is shown in Fig. 11a. The Poisson equation for the case is

$$\frac{d^2\psi}{dx^2} = \frac{-d\mathcal{E}}{dx} = \frac{-\rho_s}{\epsilon_s} = \frac{-q}{\epsilon_s}ax \quad -\frac{W}{2} \leq x \leq \frac{W}{2}, \quad (28)$$

where a is the impurity gradient (in cm^{-4}) and W is the depletion-layer width.

We have assumed that mobile carriers are negligible in the depletion region. By integrating Eq. 28 once with the boundary conditions that the electric field is zero at $\pm W/2$, we obtain the electric-field distribution shown in Fig. 11b

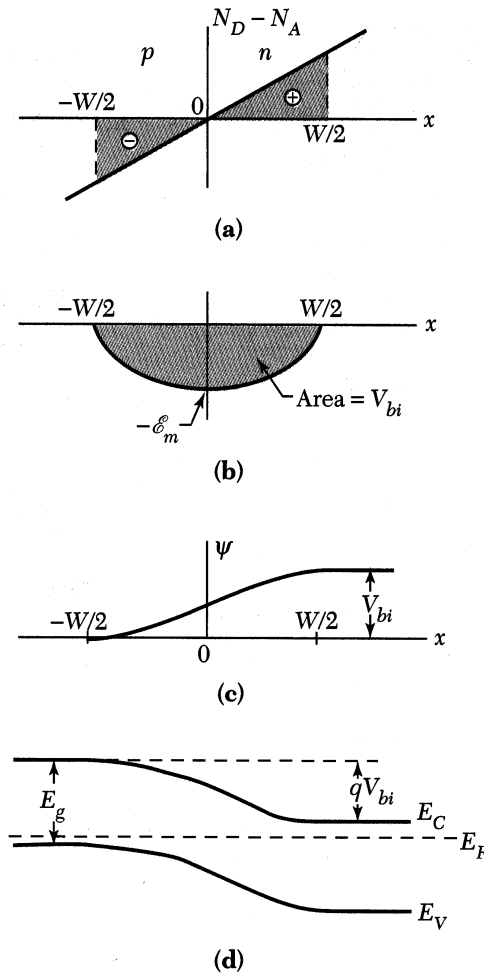


Fig. 11 Linearly graded junction in thermal equilibrium. (a) Impurity distribution. (b) Electric-field distribution. (c) Potential distribution with distance. (d) Energy band diagram.

$$\mathcal{E}(x) = -\frac{qa}{\epsilon_s} \left[\frac{(W/2)^2 - x^2}{2} \right]. \quad (29)$$

The maximum field at $x = 0$ is

$$\mathcal{E}_m = \frac{qaW^2}{8\epsilon_s}. \quad (29a)$$

Integrating Eq. 28 once again yields both the potential distribution and the corresponding energy band diagram shown in Figs. 11c and 11d, respectively. The built-in potential and the depletion layer width are given by

$$V_{bi} = \frac{qaW^3}{12\epsilon_s} \quad (30)$$

and

$$W = \left(\frac{12\epsilon_s V_{bi}}{qa} \right)^{1/3}. \quad (31)$$

Since the values of the impurity concentrations at the edges of the depletion region ($-W/2$ and $W/2$) are the same and both are equal to $aW/2$, the built-in potential for a linearly graded junction may be expressed in a form similar to Eq. 12*:

$$V_{bi} = \frac{kT}{q} \ln \left[\frac{(aW/2)(aW/2)}{n_i^2} \right] = \frac{2kT}{q} \ln \left(\frac{aW}{2n_i} \right). \quad (32)$$

Solving the transcendental equation that results when W is eliminated from Eqs. 31 and 32 yields the built-in potential as a function of a . The results for silicon and gallium arsenide linearly graded junctions are shown in Fig. 12.

When either forward or reverse bias is applied to the linearly graded junction, the variations of the depletion layer width and the energy band diagram will be similar to those shown in Fig. 10 for abrupt junctions. However, the depletion layer width will vary as $(V_{bi} - V)^{1/3}$, where V is positive for forward bias and negative for reverse bias.

EXAMPLE 3

For a silicon linearly graded junction with a impurity gradient of 10^{20} cm^{-4} , the depletion-layer width is $0.5 \text{ } \mu\text{m}$. Calculate the maximum field and built-in voltage ($T = 300 \text{ K}$):

SOLUTION From Eq. 29a and Eq. 32, we obtain

$$\mathcal{E}_m = \frac{qaW^2}{8\epsilon_s} = \frac{1.6 \times 10^{-19} \times 10^{20} \times (0.5 \times 10^{-4})^2}{8 \times 11.9 \times 8.85 \times 10^{-14}} = 4.75 \times 10^3 \text{ V/cm},$$

$$V_{bi} = \frac{2kT}{q} \ln \left(\frac{aW}{2n_i} \right) = 2 \times 0.0259 \ln \left(\frac{10^{20} \times 0.5 \times 10^{-4}}{2 \times 9.65 \times 10^9} \right) = 0.645 \text{ V}.$$

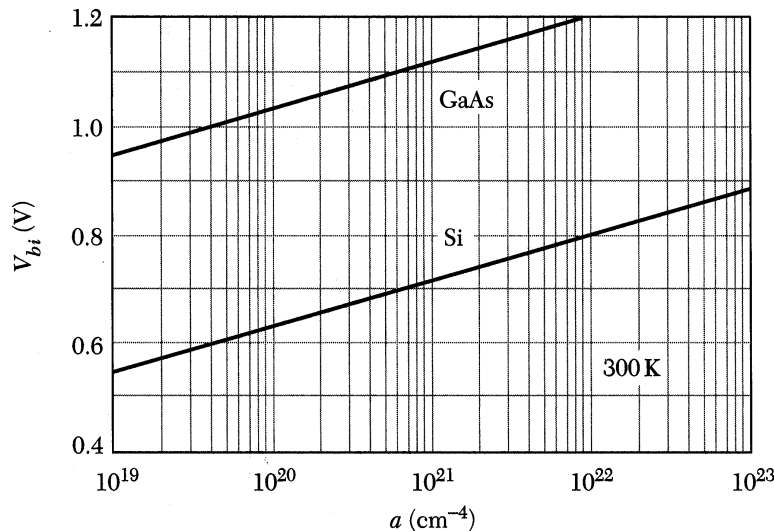


Fig. 12 Built-in potential for a linearly graded junction in Si and GaAs as a function of impurity gradient.

* Based on an accurate numerical technique, the built-in potential is given by $V_{bi} = \frac{2}{3} \frac{kT}{q} \ln \left(\frac{a^2 \epsilon_s kT / q}{8qn_i^3} \right)$.

For a given impurity gradient, the V_{bi} is smaller than that calculated from Eq. 32 by about 0.05 – 0.1 V.

► 4.4 DEPLETION CAPACITANCE

The junction depletion layer capacitance per unit area is defined as $C_j = dQ/dV$, where dQ is the incremental change in depletion layer charge per unit area for an incremental change in the applied voltage dV .*

Figure 13 illustrates the depletion capacitance of a *p-n* junction with an arbitrary impurity distribution. The charge and electric-field distributions indicated by the solid lines correspond to a voltage V applied to the *n*-side. If this voltage is increased by an amount dV , the charge and field distributions will expand to those regions bounded by the dashed lines. In Fig. 13*b*, the incremental charge dQ corresponds to the colored area between the two charge distribution curves on either side of the depletion region. The incremental space charges on the *n*- and *p*-sides of the depletion region are equal but with opposite charge polarity, thus maintaining overall charge neutrality. This incremental charge dQ causes an increase in the electric field by an amount $d\mathcal{E} = dQ/\epsilon_s$ (from Poisson's equation). The corresponding change in the applied voltage dV , represented by the cross-hatched area in Fig. 13*c* is approximately $Wd\mathcal{E}$, which equals WdQ/ϵ_s . Therefore, the depletion capacitance per unit area is given by

$$C_j \equiv \frac{dQ}{dV} = \frac{dQ}{W \frac{dQ}{\epsilon_s}} = \frac{\epsilon_s}{W} \quad (33)$$

or

$$C_j = \frac{\epsilon_s}{W} \text{ F/cm}^2. \quad (33a)$$

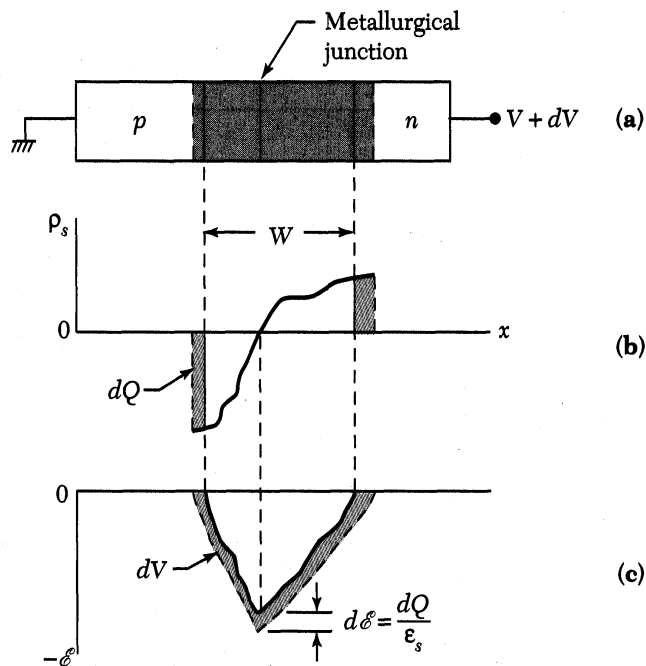


Fig. 13 (a) *p-n* junction with an arbitrary impurity profile under reverse bias. (b) Change in space charge distribution due to change in applied bias. (c) Corresponding change in electric-field distribution.

*The capacitance is also referred to as the transition region capacitance.

4.4.1 Capacitance-Voltage Characteristics

Equation 33 for the depletion capacitance per unit area is the same as the standard expression for a parallel-plate capacitor where the spacing between the two plates represents the depletion-layer width. The equation is valid for any arbitrary impurity distribution.

In deriving Eq. 33 we have assumed that only the variation of the space charge in the depletion region contributes to the capacitance. This certainly is a good assumption for the reverse-bias condition. For forward biases, however, a large current can flow across the junction corresponding to a large number of mobile carriers present within the neutral region. The incremental change of these mobile carriers with respect to the biasing voltage contributes an additional term, called the diffusion capacitance, which is considered in Section 4.6.

For a one-sided abrupt junction, we obtain, from Eqs. 27 and 33,

$$C_j = \frac{\epsilon_s}{W} = \sqrt{\frac{q\epsilon_s N_B}{2(V_{bi} - V)}} \quad (34)$$

or

$$\boxed{\frac{1}{C_j^2} = \frac{2(V_{bi} - V)}{q\epsilon_s N_B}} \quad (35)$$

It is clear from Eq. 35 that a plot of $1/C_j^2$ versus V produces a straight line for a one-sided abrupt junction. The slope gives the impurity concentration N_B of the substrate, and the intercept (at $1/C_j^2 = 0$) gives V_{bi} .

EXAMPLE 4

For a silicon one-sided abrupt junction with $N_A = 2 \times 10^{19} \text{ cm}^{-3}$ and $N_D = 8 \times 10^{15} \text{ cm}^{-3}$, calculate the junction capacitance at zero bias and reversed bias of 4 V ($T = 300 \text{ K}$).

SOLUTION From Eqs. 12, 27, and 34, we obtain at zero bias

$$V_{bi} = 0.0259 \ln \frac{2 \times 10^{19} \times 8 \times 10^{15}}{(9.65 \times 10^9)^2} = 0.906 \text{ V},$$

$$W|_{V=0} \equiv \sqrt{\frac{2\epsilon_s V_{bi}}{qN_D}} = \sqrt{\frac{2 \times 11.9 \times 8.85 \times 10^{-14} \times 0.906}{1.6 \times 10^{-19} \times 8 \times 10^{15}}} = 3.86 \times 10^{-5} = 0.386 \text{ } \mu\text{m},$$

$$C_j|_{V=0} = \frac{\epsilon_s}{W|_{V=0}} = \sqrt{\frac{q\epsilon_s N_B}{2V_{bi}}} = 2.728 \times 10^{-8} \text{ F / cm}^2.$$

From Eqs. 27 and 34, we can obtain at reverse bias of 4 V:

$$W|_{V=-4} \equiv \sqrt{\frac{2\epsilon_s (V_{bi} - V)}{qN_D}} = \sqrt{\frac{2 \times 11.9 \times 8.85 \times 10^{-14} \times (0.906 + 4)}{1.6 \times 10^{-19} \times 8 \times 10^{15}}} = 8.99 \times 10^{-5} = 0.899 \text{ } \mu\text{m},$$

$$C_j|_{V=-4} = \frac{\epsilon_s}{W|_{V=-4}} = \sqrt{\frac{q\epsilon_s N_B}{2(V_{bi} - V)}} = 1.172 \times 10^{-8} \text{ F / cm}^2.$$

4.4.2 Evaluation of Impurity Distribution

The capacitance-voltage characteristics can be used to evaluate an arbitrary impurity distribution. We consider the case of p^+-n junction with a doping profile on the n -side, as shown in Fig. 14*b*. As before, the incremental change in depletion layer charge per unit area dQ for an incremental change in the applied voltage dV is given by $qN(W) dW$ (i.e., the shaded area in Fig. 14*b*). The corresponding change in applied voltage (shaded area in Fig. 14*c*) is

$$dV \cong (d\mathcal{E}) W = \left(\frac{dQ}{\epsilon_s} \right) W = \frac{qN(W) dW^2}{2\epsilon_s}. \quad (36)$$

By substituting W from Eq. 33, we obtain an expression for the impurity concentration at the edge of the depletion region:

$$N(W) = \frac{2}{q\epsilon_s} \left[\frac{1}{d(1/C_j^2)/dV} \right]. \quad (37)$$

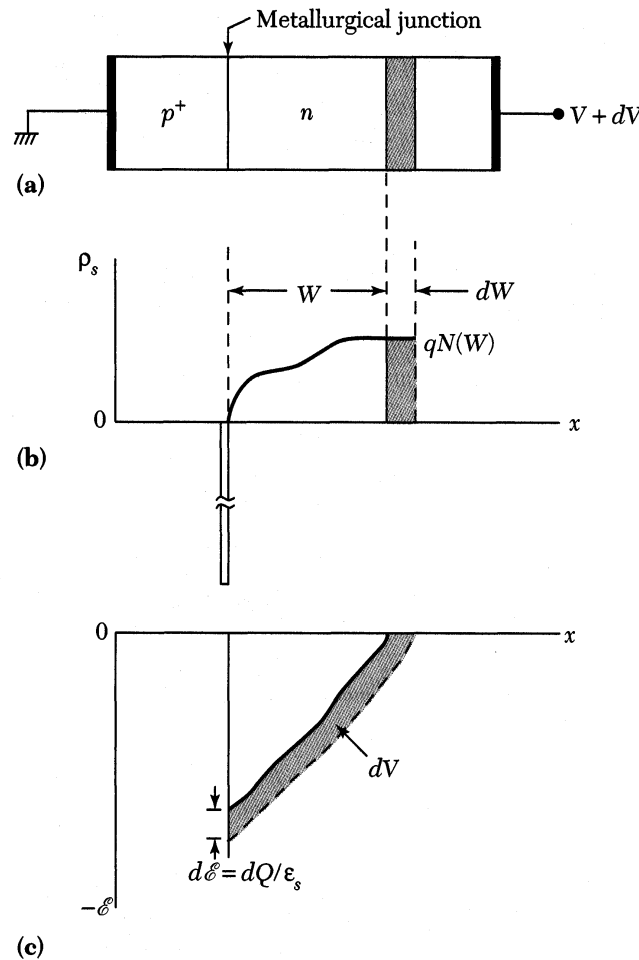


Fig. 14 (a) p^+-n junction with an arbitrary impurity distribution. (b) Change in space charge distribution in the lightly doped side due to a change in applied bias. (c) Corresponding change in the electric-field distribution.

Thus, we can measure the capacitance per unit area versus reverse-bias voltage and plot $1/C_j^2$ versus V . The slope of the plot, that is, $d(1/C_j^2)/dV$, yields $N(W)$. Simultaneously, W is obtained from Eq. 33. A series of such calculations produces a complete impurity profile. This approach is referred to as the C - V method for measuring impurity profiles.

For a linearly graded junction, the depletion layer capacitance is obtained from Eqs. 31 and 33:

$$C_j = \frac{\epsilon_s}{W} = \left[\frac{qa\epsilon_s^2}{12(V_{bi} - V)} \right]^{1/3} \text{ F/cm}^2. \quad (38)$$

For such a junction we can plot $1/C^3$ versus V and obtain the impurity gradient and V_{bi} from the slope and the intercept, respectively.

4.4.3 Varactor

Many circuit applications employ the voltage-variable properties of reverse-biased p - n junctions. A p - n junction designed for such a purpose is called a *varactor*, which is a shortened form of variable reactor. As previously derived, the reverse-biased depletion capacitance is given by

$$C_j \propto (V_{bi} + V_R)^{-n} \quad (39)$$

or

$$C_j \propto (V_R)^{-n} \quad \text{for } V_R \gg V_{bi}, \quad (39a)$$

where $n = 1/2$ for a linearly graded junction and $n = 3/2$ for an abrupt junction. Thus, the voltage sensitivity of C (i.e., variation of C with V_R) is greater for an abrupt junction than for a linearly graded junction. We can further increase the voltage sensitivity by using a hyperabrupt junction having an exponent n (Eq. 39) greater than $1/2$.

Figure 15 shows three p^+ - n doping profiles with the donor distribution $N_D(x)$ given by $B(x/x_0)^m$, where B and x_0 are constants, $m = 1$ for a linearly graded junction, $m = 0$ for an abrupt junction, and $m = -3/2$ for a hyperabrupt junction. The hyperabrupt profile can

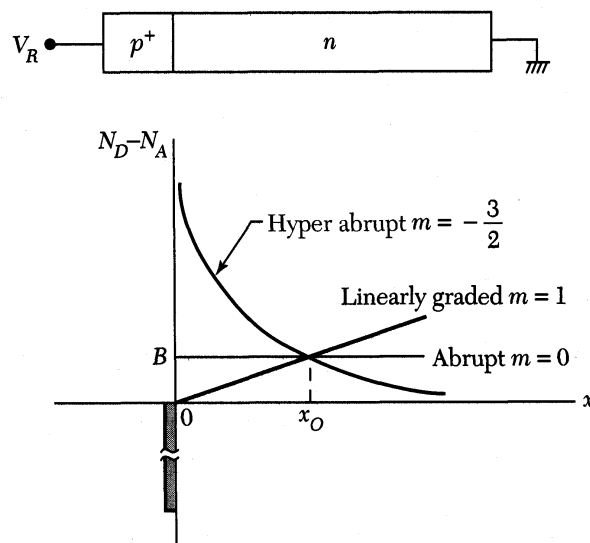


Fig. 15 Impurity profiles for hyperabrupt, one-sided abrupt, and one-sided linearly graded junctions.

be achieved by epitaxial growth techniques discussed in Chapter 10. To obtain the capacitance-voltage relationship, we solve Poisson's equation:

$$\frac{d^2\psi}{dx^2} = -B\left(\frac{x}{x_0}\right)^m. \quad (40)$$

Integrating Eq. 40 twice with appropriate boundary conditions gives the dependence of the depletion layer width on the reverse bias:

$$W \propto (V_R)^{1/(m+2)} \quad (41)$$

Therefore,

$$C_j = \frac{\epsilon_s}{W} \propto (V_R)^{-1/(m+2)} \quad (42)$$

Comparing Eq. 42 with Eq. 39a yields $n = 1/(m + 2)$. For hyperabrupt junctions with $n > 1/2$, m must be a negative number.

By choosing different values for m , we can obtain a wide variety of C_j -versus- V_R dependencies for specific applications. One interesting example, shown in Fig. 15, is the case for $m = -3/2$. For this case, $n = 2$. When this varactor is connected to an inductor L in a resonant circuit, the resonant frequency varies linearly with the voltage applied to the varactor:

$$\omega_r = \frac{1}{\sqrt{LC_j}} \propto \frac{1}{\sqrt{V_R^{-n}}} = V_R \quad \text{for } n = 2. \quad (43)$$

► 4.5 CURRENT-VOLTAGE CHARACTERISTICS

A voltage applied to a *p-n* junction will disturb the precise balance between the diffusion current and drift current of electrons and holes. Under forward bias, the applied voltage reduces the electrostatic potential across the depletion region, as shown in the middle of Fig. 16. The drift current is reduced in comparison to the diffusion current. We have an enhanced hole diffusion from the *p*-side to the *n*-side and electron diffusion from the *n*-side to the *p*-side. Therefore, minority carrier injections occur, that is, electrons are injected into the *p*-side, whereas holes are injected into the *n*-side. Under reverse bias, the applied voltage increases the electrostatic potential across the depletion region as shown in the middle of Fig. 16*b*. This greatly reduces the diffusion currents, resulting in a small reverse current. In this section, we first consider the ideal current-voltage characteristics. We then discuss departures from these ideal characteristics due to generation and recombination and other effects.

4.5.1 Ideal Characteristics

We now derive the ideal current-voltage characteristics based on the following assumptions: (a) The depletion region has abrupt boundaries and, outside the boundaries, the semiconductor is assumed to be neutral; (b) the carrier densities at the boundaries are related by the electrostatic potential difference across the junction. (c) the low-injection condition, that is, the injected minority carrier densities, are small compared with the majority carrier densities; in other words, the majority carrier densities are changed negligibly at the boundaries of neutral regions by the applied bias; and (d) neither genera-

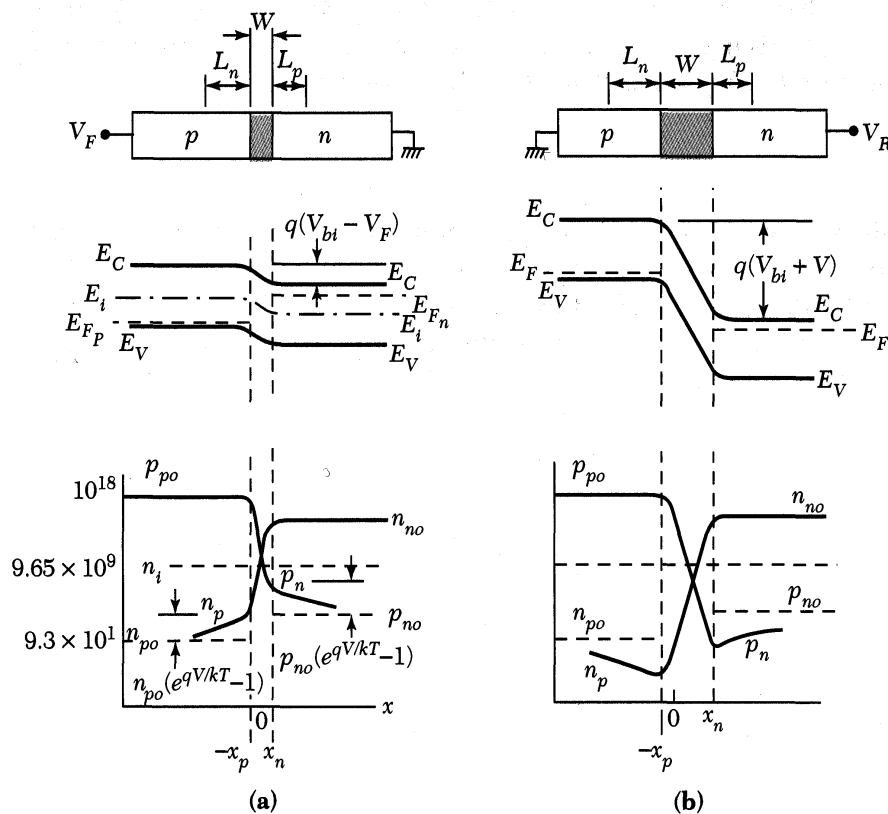


Fig. 16 Depletion region, energy band diagram and carrier distribution. (a) Forward bias. (b) Reverse bias.

tion nor recombination current exists in the depletion region, and the electron and hole currents are constant throughout the depletion region. Departures from these idealized assumptions are considered in the next section.

At thermal equilibrium, the majority carrier density in the neutral regions is essentially equal to the doping concentration. We use the subscripts *n* and *p* to denote the semiconductor type and the subscript *o* to specify the condition of thermal equilibrium. Hence, *n_{no}* and *n_{po}* are the equilibrium electron densities in the *n*- and *p*-sides, respectively. The expression for the built-in potential in Eq. 12 can be rewritten as

$$V_{bi} = \frac{kT}{q} \ln \frac{p_{po} n_{no}}{n_i^2} = \frac{kT}{q} \ln \frac{n_{no}}{n_{po}}, \quad (44)$$

where the mass action law $p_{po} n_{po} = n_i^2$ has been used. Rearranging Eq. 44 gives

$$n_{no} = n_{po} e^{qV_{bi}/kT}. \quad (45)$$

Similarly, we have

$$p_{po} = p_{no} e^{qV_{bi}/kT}. \quad (46)$$

We note from Eqs. 45 and 46 that the electron density and the hole density at the two boundaries of the depletion region are related through the electrostatic potential difference *V_{bi}* at thermal equilibrium. From our second assumption we expect that the same relation holds when the electrostatic potential difference is changed by an applied voltage.

When a forward bias is applied, the electrostatic potential difference is reduced to $V_{bi} - V_F$; but when a reverse bias is applied, the electrostatic potential difference is increased to $V_{bi} + V_R$. Thus, Eq. 45 is modified to

$$n_n = n_p e^{q(V_{bi}-V)/kT}, \quad (47)$$

where n_n and n_p are the nonequilibrium electron densities at the boundaries of the depletion region in the *n*- and *p*-sides, respectively, with V positive for forward bias and negative for reverse bias. For the low-injection condition, the injected minority carrier density is much smaller than the majority carrier density; therefore, $n_n \cong n_{no}$. Substituting this condition and Eq. 45 into Eq. 47 yields the electron density at the boundary of the depletion region on the *p*-side ($x = -x_p$):

$$n_p = n_{po} e^{qV/kT} \quad (48)$$

or

$$n_p - n_{po} = n_{po} (e^{qV/kT} - 1). \quad (48a)$$

Similarly, we have

$$p_n = p_{no} e^{qV/kT} \quad (49)$$

or

$$p_n - p_{no} = p_{no} (e^{qV/kT} - 1) \quad (49a)$$

at $x = x_n$ for the *n*-type boundary. Figures 16*a* and 16*b* show band diagrams and carrier concentrations in a *p-n* junction under forward-bias and reverse-bias conditions, respectively. Note that the minority carrier densities at the boundaries ($-x_p$ and x_n) increase substantially above their equilibrium values under forward bias, whereas they decrease below their equilibrium values under reverse bias. Equations 48 and 49 define the minority carrier densities at the boundaries of depletion region. These equations are the most important boundary conditions for the ideal current-voltage characteristics.

Under our idealized assumptions, no current is generated within the depletion region; all currents come from the neutral regions. In the neutral *n*-region, there is no electric field, thus the steady-state continuity equation reduces to

$$\frac{d^2 p_n}{dx^2} - \frac{p_n - p_{no}}{D_p \tau_p} = 0. \quad (50)$$

The solution of Eq. 50 with the boundary conditions of Eq. 49 and $p_n(x = \infty) = p_{no}$ gives

$$p_n - p_{no} = p_{no} (e^{qV/kT} - 1) e^{-(x-x_n)/L_p}, \quad (51)$$

where L_p , which is equal to $\sqrt{D_p \tau_p}$, is the diffusion length of holes (minority carriers) in the *n*-region. At $x = x_n$,

$$J_p(x_n) = -qD_p \left. \frac{dp_n}{dx} \right|_{x_n} = \frac{qD_p p_{no}}{L_p} (e^{qV/kT} - 1). \quad (52)$$

Similarly, we obtain for the neutral *p*-region

$$n_p - n_{po} = n_{po} \left(e^{qV/kT} - 1 \right) e^{(x+x_p)/L_n} \quad (53)$$

and

$$J_n(-x_p) = qD_n \left. \frac{dn_p}{dx} \right|_{-x_p} = \frac{qD_n n_{po}}{L_n} \left(e^{qV/kT} - 1 \right), \quad (54)$$

where L_n , which is equal to $\sqrt{D_n \tau_n}$, is the diffusion length of electrons. The minority carrier densities (Eqs. 51 and 53) are shown in the middle of Fig. 17.

The graphs illustrate that the injected minority carriers recombine with the majority carriers as the minority carriers move away from the boundaries. The electron and hole currents are shown at the bottom of Fig. 17. The hole and electron currents at the boundaries are given by Eqs. 52 and 54, respectively. The hole diffusion current will decay

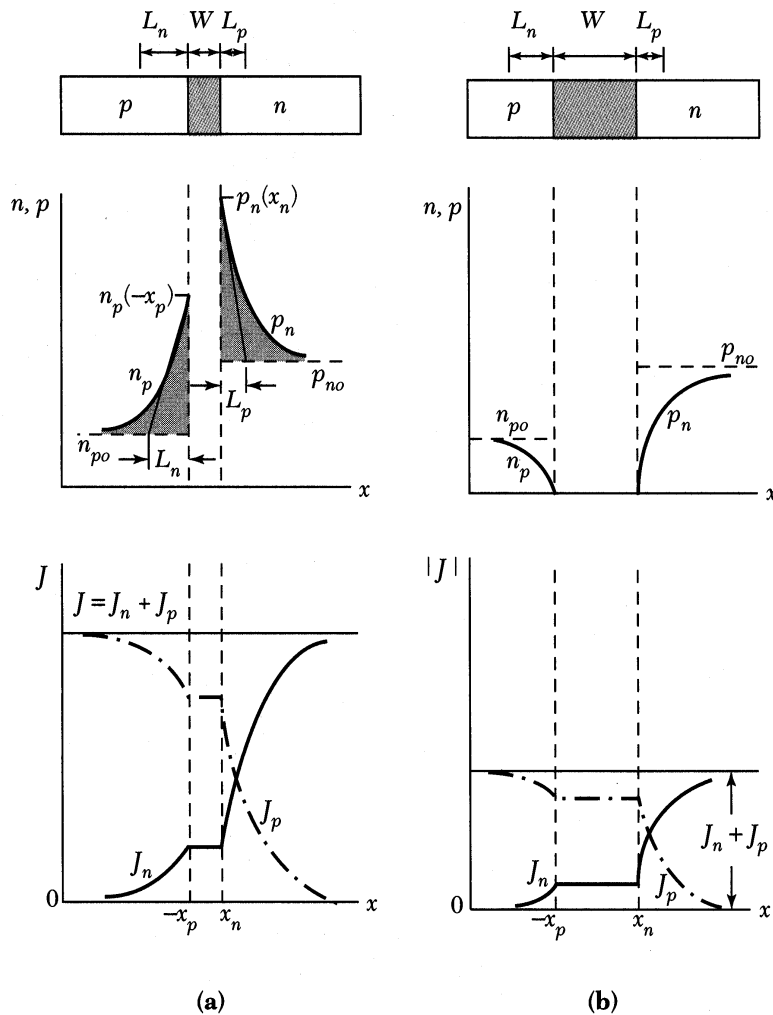


Fig. 17 Injected minority carrier distribution and electron and hole currents. (a) Forward bias. (b) Reverse bias. The figure illustrates idealized currents. For practical devices, the currents are not constant across the space charge layer.

exponentially in the *n*-region with diffusion length L_p , and the electron diffusion current will decay exponentially in the *p*-region with diffusion length L_n .

The total current is constant throughout the device and is the sum of Eqs. 52 and 54:

$$J = J_p(x_n) + J_n(-x_p) = J_s(e^{qV/kT} - 1), \quad (55)$$

$$J_s \equiv \frac{qD_p p_{n0}}{L_p} + \frac{qD_n n_{p0}}{L_n}, \quad (55a)$$

where J_s is the saturation current density. Equation 55 is the *ideal diode equation*.¹ The ideal current-voltage characteristic is shown in Figs. 18*a* and 18*b* in the Cartesian and

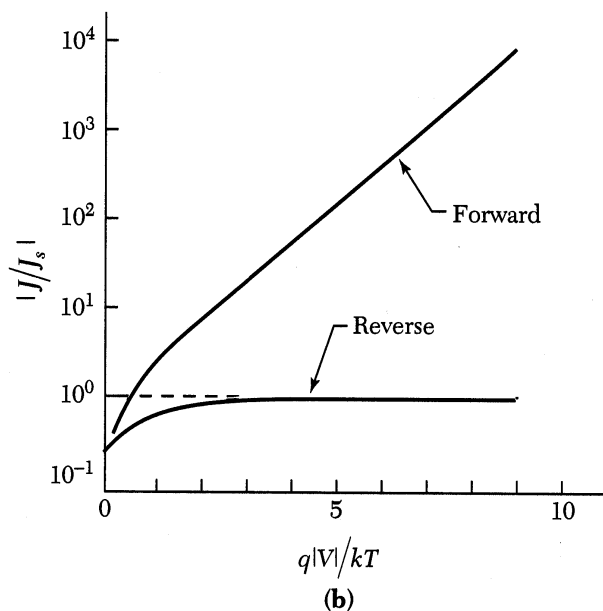
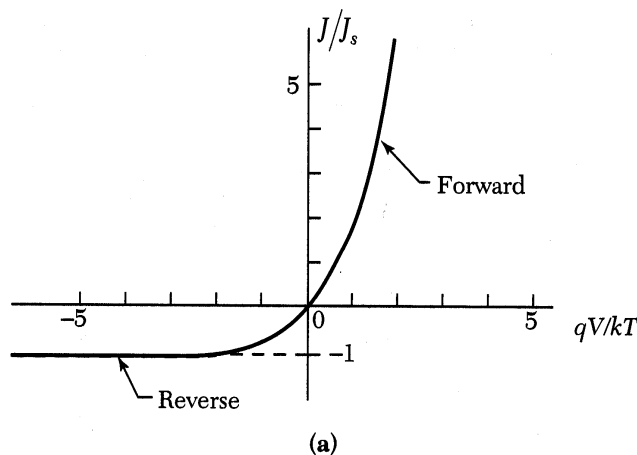


Fig. 18 Ideal current-voltage characteristics. (a) Cartesian plot. (b) Semilog plot.

semilog plots, respectively. In the forward direction with positive bias on the *p*-side, for $V \geq 3kT/q$, the rate of current increase is constant, as shown in Fig. 18*b*. At 300 K for every decade change of current, the voltage change for an ideal diode is 60 mV ($= 2.3 kT/q$). In the reverse direction, the current density saturates at $-J_s$.

EXAMPLE 5

Calculate the ideal reverse saturation current in a Si *p-n* junction diode with a cross-sectional area of $2 \times 10^{-4} \text{ cm}^2$. The parameters of the diode are

$$\begin{aligned} N_A &= 5 \times 10^{16} \text{ cm}^{-3}, & N_D &= 10^{16} \text{ cm}^{-3}, & n_i &= 9.65 \times 10^9 \text{ cm}^{-3} \\ D_n &= 21 \text{ cm}^2/\text{s}, & D_p &= 10 \text{ cm}^2/\text{s}, & \tau_p &= \tau_n = 5 \times 10^{-7} \text{ s}. \end{aligned}$$

SOLUTION From Eq. 55*a* and $L_p = \sqrt{D_p \tau_p}$, we can obtain

$$\begin{aligned} J_s &= \frac{qD_p p_{n0}}{L_p} + \frac{qD_n n_{p0}}{L_n} = qn_i^2 \left(\frac{1}{N_D} \sqrt{\frac{D_p}{\tau_p}} + \frac{1}{N_A} \sqrt{\frac{D_n}{\tau_n}} \right) \\ &= 1.6 \times 10^{-19} \times (9.65 \times 10^9)^2 \left(\frac{1}{10^{16}} \sqrt{\frac{10}{5 \times 10^{-7}}} + \frac{1}{5 \times 10^{16}} \sqrt{\frac{21}{5 \times 10^{-7}}} \right) \\ &= 8.58 \times 10^{-12} \text{ A/cm}^2. \end{aligned}$$

From the cross-sectional area $A = 2 \times 10^{-4} \text{ cm}^2$, we obtain

$$I_s = A \times J_s = 2 \times 10^{-4} \times 8.58 \times 10^{-12} = 1.72 \times 10^{-15} \text{ A}.$$

4.5.2 Generation-Recombination and High-Injection Effects

The ideal diode equation, Eq. 55, adequately describes the current-voltage characteristics of germanium *p-n* junctions at low current densities. For silicon and gallium arsenide *p-n* junctions, however, the ideal equation can only give qualitative agreement because of the generation or recombination of carriers in the depletion region.

Consider the reverse-bias condition first. Under reverse bias, carrier concentrations in the depletion region fall far below their equilibrium concentrations. The dominant generation-recombination processes discussed in Chapter 3 are those of electron and hole emissions through bandgap generation-recombination centers. The capture processes are not important because their rates are proportional to the concentration of free carriers, which is very small in the reverse-biased depletion region.

The two emission processes operate in the steady state by alternately emitting electrons and holes. The rate of electron-hole pair generation can be obtained from Eq. 48 of Chapter 3 with the conditions $p_n < n_i$ and $n_n < n_i$:

$$\begin{aligned} G = -U &= \left[\frac{\sigma_p \sigma_n v_{th} N_t}{\sigma_n \exp\left(\frac{E_t - E_i}{kT}\right) + \sigma_p \exp\left(\frac{E_i - E_t}{kT}\right)} \right] n_i \\ &\equiv \frac{n_i}{\tau_g}, \end{aligned} \tag{56}$$

where τ_g , the generation lifetime, is the reciprocal of the expression in the square brackets. We can arrive at an important conclusion about electron-hole generation from this

expression. Let us consider a simple case where $\sigma_n = \sigma_p = \sigma_o$. For this case, Eq. 56 reduces to

$$G = \frac{\sigma_o v_{th} N_t n_i}{2 \cosh\left(\frac{E_t - E_i}{kT}\right)} \quad (57)$$

The generation rate reaches a maximum value at $E_t = E_i$ and falls off exponentially as E_t moves in either direction away from the middle of the bandgap. Thus, only those centers with an energy level of E_t near the intrinsic Fermi level can contribute significantly to the generation rate.

The current due to generation in the depletion region is

$$J_{gen} = \int_0^W qGdx \cong qGW = \frac{qn_i W}{\tau_g} \quad (58)$$

where W is the depletion layer width. The total reverse current for a p^+-n junction, that is, for $N_A \gg N_D$ and for $V_R > 3kT/q$, can be approximated by the sum of both the diffusion current in the neutral regions and the generation current in the depletion region:

$$J_R \cong q \sqrt{\frac{D_p}{\tau_p} \frac{n_i^2}{N_D}} + \frac{qn_i W}{\tau_g} \quad (59)$$

For semiconductors with large values of n_i , such as germanium, the diffusion current dominates at room temperature, and the reverse current follows the ideal diode equation. But if n_i is small, such as for silicon and gallium arsenide, the generation current in the depletion region may dominate.

► EXAMPLE 6

Consider the Si p - n junction diode in Example 5 and assume $\tau_g = \tau_p = \tau_n$, calculate the generation current density for a reverse bias of 4 V.

SOLUTION From Eq. 20, we obtain

$$\begin{aligned} W &= \sqrt{\frac{2\epsilon_s}{q} \left(\frac{N_A + N_D}{N_A N_D} \right) (V_{bi} + V)} = \sqrt{\frac{2\epsilon_s}{q} \left(\frac{N_A + N_D}{N_A N_D} \right) \left(\frac{kT}{q} \ln \frac{N_A N_D}{n_i^2} + V \right)} \\ &= \sqrt{\frac{2 \times 11.9 \times 8.85 \times 10^{-14}}{1.6 \times 10^{-19}} \left(\frac{5 \times 10^{16} + 10^{16}}{5 \times 10^{16} \times 10^{16}} \right) \left(0.0259 \ln \frac{5 \times 10^{16} \times 10^{16}}{(9.65 \times 10^9)^2} + V \right)} \\ &= 3.97 \times \sqrt{0.758 + V} \times 10^{-5} \text{ cm.} \end{aligned}$$

Hence the generation current density is

$$\begin{aligned} J_{gen} &= \frac{qn_i W}{\tau_g} = \frac{1.6 \times 10^{-19} \times 9.65 \times 10^9}{5 \times 10^{-7}} \times 3.97 \times \sqrt{0.758 + V} \times 10^{-5} \text{ A/cm}^2 \\ &= 1.22 \times \sqrt{0.758 + V} \times 10^{-7} \text{ A/cm}^2. \end{aligned}$$

If we apply a reversed bias of 4 V, the generation current density is $2.66 \times 10^{-7} \text{ A/cm}^2$. ◀

Under forward bias, the concentrations of both electrons and holes exceed their equilibrium values. The carriers will attempt to return to their equilibrium values by recombination. Therefore, the dominant generation-recombination processes in the depletion region are the capture processes. From Eq. 49 we obtain

$$p_n n_n \cong p_{no} n_{no} e^{qV/kT} = n_i^2 e^{qV/kT}. \quad (60)$$

Substituting Eq. 60 in Eq. 48 of Chapter 3 and assuming $\sigma_n = \sigma_p = \sigma_o$ yields

$$U = \frac{\sigma_o v_{th} N_t n_i^2 (e^{qV/kT} - 1)}{n_n + p_n + 2n_i \cosh \frac{E_i - E_t}{kT}}. \quad (61)$$

In either recombination or generation, the most effective centers are those located near E_i . As practical examples, gold and copper yield effective generation-recombination centers in silicon where the values of $E_t - E_i$ are 0.02 V for gold and -0.02 eV for copper. In gallium arsenide, chromium gives an effective center with an $E_t - E_i$ value of 0.08 eV.

Equation 61 can be simplified for the case $E_t = E_i$:

$$U = \sigma_o v_{th} N_t \frac{n_i^2 (e^{qV/kT} - 1)}{n_n + p_n + 2n_i} \quad (62)$$

For a given forward bias, U reaches its maximum value at a location in the depletion region either where the denominator $n_n + p_n + 2n_i$ is a minimum or where the sum of the electron and hole concentrations, $n_n + p_n$, is at its minimum value. Since the product of these concentrations is a constant given by Eq. 60, the condition $d(p_n + n_n) = 0$ leads to

$$dp_n = -dn_n = \frac{p_n n_n}{p_n^2} dp_n \quad (63)$$

or

$$p_n = n_n \quad (64)$$

as the condition for the minimum. This condition exists at the location in the depletion region, where E_i is halfway between E_{Fp} and E_{Fn} , as illustrated in the middle of Fig. 16a. Here, the carrier concentrations are

$$p_n = n_n = n_i e^{qV/2kT} \quad (65)$$

and, therefore,

$$U_{\max} = \sigma_o v_{th} N_t \frac{n_i^2 (e^{qV/kT} - 1)}{2n_i (e^{qV/2kT} + 1)}. \quad (66)$$

For $V > 3kT/q$,

$$U_{\max} \cong \frac{1}{2} \sigma_o v_{th} N_t n_i e^{qV/2kT}. \quad (67)$$

The recombination current is then

$$J_{\text{rec}} = \int_0^W q U dx \cong \frac{qW}{2} \sigma_o v_{th} N_t n_i e^{qV/2kT} = \frac{qW n_i}{2\tau_r} e^{qV/2kT}, \quad (68)$$

where τ_r is the effective recombination lifetime given by $1/(\sigma_o v_{th} N_t)$. The total forward current can be approximated by the sum of Eqs. 55 and 68. For $p_{no} \gg n_{po}$ and $V > 3kT/q$ we have

$$J_F = q \sqrt{\frac{D_p}{\tau_p} \frac{n_i^2}{N_D}} e^{qV/kT} + \frac{qWn_i}{2\tau_r} e^{qV/2kT}. \quad (69)$$

In general, the experimental results can be represented empirically by

$$J_F \approx \exp\left(\frac{qV}{\eta kT}\right), \quad (70)$$

where the factor η is called the *ideality factor*. When the ideal diffusion current dominates, $\eta = 1$; whereas when the recombination current dominates, $\eta = 2$. When both currents are comparable, η has a value of between 1 and 2.

Figure 19 shows the measured forward characteristics of a silicon and gallium arsenide *p-n* junction at room temperature.² At low current levels, recombination current dominates and $\eta = 2$. At higher current levels, diffusion current dominates and η approaches 1.

At even higher current levels, we notice that the current departs from the ideal $\eta = 1$ situation and increases more gradually with forward voltage. This phenomenon is associated with two effects: series resistance and high injection. We first consider the series resistance effect. At both low- and medium-current levels, the IR drop across the neutral regions is usually small compared with kT/q (26 mV at 300 K), where I is the forward current and R is the series resistance. For example, for a silicon diode with $R =$

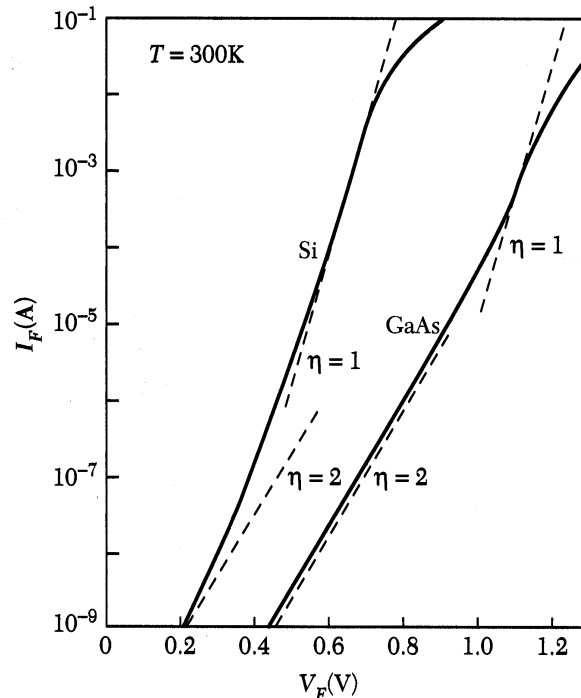


Fig. 19 Comparison of the forward current-voltage characteristics of Si and GaAs diodes² at 300 K. Dashed lines indicate slopes of different ideality factors η .

1.5 ohms, the IR drop at 1 mA is only 1.5 mV. However, at 100 mA, the IR drop becomes 0.15 V, which is six times larger than kT/q . This IR drop reduces the bias across the depletion region; therefore, the current becomes

$$I \cong I_s \exp \left[\frac{q(V - IR)}{kT} \right] = \frac{I_s \exp(qV/kT)}{\exp \left[\frac{q(IR)}{kT} \right]} \quad (71)$$

and the ideal diffusion current is reduced by the factor $\exp[q(IR)/kT]$.

At high-current densities, the injected minority carrier density is comparable to the majority concentration, that is at the n -side of the junction $p_n(x=x_n) \cong n_n$. This is the high-injection condition. By substituting the high-injection condition in Eq. 60, we obtain $p_n(x=x_n) \cong n_i \exp(qV/2kT)$. Using this as a boundary condition, the current becomes roughly proportional to $\exp(qV/2kT)$. Thus, the current increases at a slower rate under the high-injection condition.

4.5.3 Temperature Effect

Operating temperature has a profound effect on device performance. In both the forward-bias and reverse-bias conditions, the magnitudes of the diffusion and the recombination-generation currents depend strongly on temperature. We consider the forward-bias case first. The ratio of hole diffusion current to the recombination is given by

$$\frac{I_{\text{diffusion}}}{I_{\text{recombination}}} = 2 \frac{n_i}{N_D} \frac{L_p}{W} \frac{\tau_r}{\tau_p} e^{qV/2kT} \approx \exp \left(-\frac{E_g - qV}{2kT} \right). \quad (72)$$

This ratio depends on both the temperature and the semiconductor bandgap. Figure 20a shows the temperature dependence of the forward characteristics of a silicon diode. At

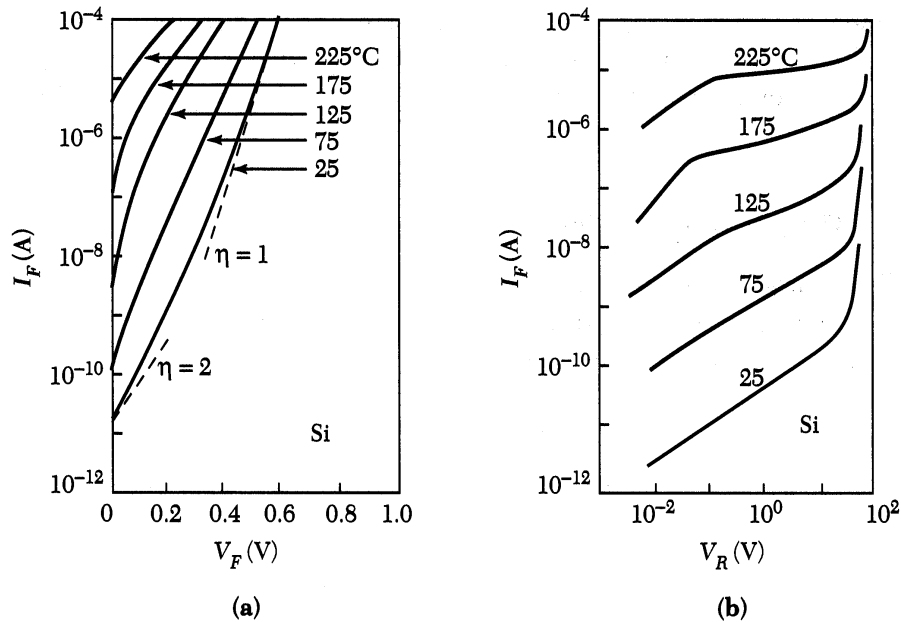


Fig. 20 Temperature dependence of the current-voltage characteristics of a Si diode². (a) Forward bias. (b) Reverse bias.

room temperature for small forward voltages, the recombination current generally dominates, whereas at higher forward voltages the diffusion current usually dominates. At a given forward bias, as the temperature increases, the diffusion current will increase more rapidly than the recombination current. Therefore, the ideal diode equation will be followed over a wide range of forward biases as the temperature increases.

The temperature dependence of the saturation current density J_s (Eq. 55a) for a one-sided p^+-n junction in which diffusion current dominates is given by

$$J_s \equiv \frac{qD_p p_{no}}{L_p} \approx n_i^2 \approx \exp\left(-\frac{E_g}{kT}\right). \quad (73)$$

Thus, the activation energy obtained from the slope of a plot of J_s versus $1/T$ corresponds to the energy bandgap E_g .

In the reverse-bias condition for a p^+-n junction, the ratio of the diffusion current to the generation current is

$$\frac{I_{\text{diffusion}}}{I_{\text{generation}}} = \frac{n_i L_p \tau_g}{N_D W \tau_p}. \quad (74)$$

This ratio is proportional to the intrinsic carrier density n_i . As the temperature increases, the diffusion current eventually dominates. Figure 20b shows the effects of temperature on the reverse characteristics of a silicon diode. At low temperatures, the generation current dominates and the reverse current varies as $\sqrt{V_R}$ in accordance with Eq. 58 for an abrupt junction (i.e., $W \sim \sqrt{V_R}$). As the temperature increases beyond 175°C, the current demonstrates a saturation tendency for $V_R \geq 3kT/q$, at which point the diffusion current becomes dominant.

► 4.6 CHARGE STORAGE AND TRANSIENT BEHAVIOR

Under forward bias, electrons are injected from the n -region into the p -region and holes are injected from the p -region into the n -region. Once injected across the junction, the minority carriers recombine with the majority carriers and decay exponentially with distance, as shown in Fig. 17a. These minority-carrier distributions lead to current flow and to charge storage in the p - n junction. We consider the stored charge, its effect on junction capacitance, and the transient behavior of the p - n junction due to sudden changes of bias.

4.6.1 Minority-Carrier Storage

The charge of injected minority carriers per unit area stored in the neutral n -region can be found by integrating the excess holes in the neutral region, shown as the shaded area in the middle of Fig. 17a, using Eq. 51:

$$\begin{aligned} Q_p &= q \int_{x_n}^{\infty} (p_n - p_{no}) dx, \\ &= q \int_{x_n}^{\infty} p_{no} \left(e^{qV/kT} - 1 \right) e^{-(x-x_n)/L_p} dx, \\ &= qL_p p_{no} \left(e^{qV/kT} - 1 \right). \end{aligned} \quad (75)$$

A similar expression can be obtained for the stored electrons in the neutral p -region. The number of stored minority carriers depends on both the diffusion length and the charge

density at the boundary of the depletion region. We can express the stored charge in terms of the injected current. From Eqs. 52 and 75, we have

$$Q_p = \frac{L_p^2}{D_p} J_p(x_n) = \tau_p J_p(x_n) \quad (76)$$

Equation 76 states that the amount of stored charge is the product of the current and lifetime of the minority carriers. This is because the injected holes diffuse farther into the n -region before recombining if their lifetime is longer; thus, more holes are stored.

EXAMPLE 7

For an ideal abrupt silicon p^+-n junction with $N_D = 8 \times 10^{15} \text{ cm}^{-3}$, calculate the stored minority carriers per unit area in the neutral n -region when a forward bias of 1V is applied. The diffusion length of the holes is 5 μm .

SOLUTION From Eq. 75, we obtain

$$\begin{aligned} Q_p &= qL_p p_{no} \left(e^{qV/kT} - 1 \right) = 1.6 \times 10^{-19} \times 5 \times 10^{-4} \text{ cm} \times \frac{(9.65 \times 10^9)^2}{8 \times 10^{15}} \times \left(e^{\frac{1}{0.0259}} - 1 \right) \\ &= 4.69 \times 10^{-2} \text{ C/cm}^2. \end{aligned}$$

4.6.2 Diffusion Capacitance

The depletion-layer capacitance considered previously accounts for most of the junction capacitance when the junction is reverse biased. When the junction is forward biased, there is an additional significant contribution to junction capacitance from the rearrangement of the stored charges in the neutral regions. This is called the *diffusion capacitance*, denoted C_d , a term derived from the ideal-diode case in which minority carriers move across the neutral region by diffusion.

The diffusion capacitance of the stored holes in the neutral n -region is obtained by applying the definition $C_d = AdQ_p/dV$ to Eq. 75:

$$C_d = \frac{Aq^2 L_p p_{no}}{kT} e^{qV/kT}, \quad (77)$$

where A is the device cross-section area. We may add the contribution to C_d of the stored electrons in the neutral p -region in cases of significant storage. For a p^+-n junction, however, $n_{po} \ll p_{no}$, and the contribution to C_d of the stored electrons becomes insignificant. Under reverse bias (i.e., V is negative), Eq. 77 shows that C_d is inconsequential because of negligible minority-carrier storage.

In many applications we prefer to represent a p - n junction by an equivalent circuit. In addition to diffusion capacitance C_d and depletion capacitance C_j , we must include conductance to account for the current through the device. In the ideal diode the conductance can be obtained from Eq. 55:

$$G = \frac{AdJ}{dV} = \frac{qA}{kT} J_s e^{qV/kT} = \frac{qA}{kT} (J + J_s) \cong \frac{qI}{kT}. \quad (78)$$

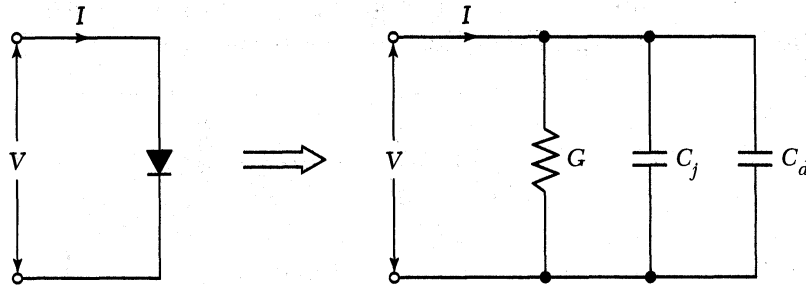


Fig. 21 Small-signal equivalent circuit of a *p-n* junction.

The diode equivalent circuit is shown in Fig. 21, where C_j stands for the total depletion capacitance (i.e., the result in Eq. 33 times the device area A). For low-voltage, sinusoidal excitation of a diode that is biased quiescently (i.e., at *dc*), the circuit shown in Fig. 21 provides adequate accuracy. Therefore, we refer to it as the diode small-signal equivalent circuit.

4.6.3 Transient Behavior

For switching applications, the forward-to-reverse-bias transition must be nearly abrupt and the transient time short. Figure 22*a* shows a simple circuit where a forward current I_F flows through a *p-n* junction. At time $t = 0$, switch S is suddenly thrown to the right and an initial reverse current $I_R \cong V/R$ flows. The transient time t_{off} , plotted in Fig. 22*b*, is the time required for the current to reach 10% of the initial reverse current I_R .

The transient time may be estimated as follows. Under the forward-bias condition, the stored minority carriers in the *n*-region for a p^+n junction is given by Eq. 76:

$$Q_p = \tau_p J_p = \tau_p \frac{I_F}{A}, \quad (79)$$

where I_F is the total forward current and A is the device area. If the average current flowing during the turn-off period is $I_{R,\text{ave}}$, the turn-off time is the length of time required to remove the total stored charge Q_p :

$$t_{\text{off}} \cong \frac{Q_p A}{I_{R,\text{ave}}} = \tau_p \left(\frac{I_F}{I_{R,\text{ave}}} \right). \quad (80)$$

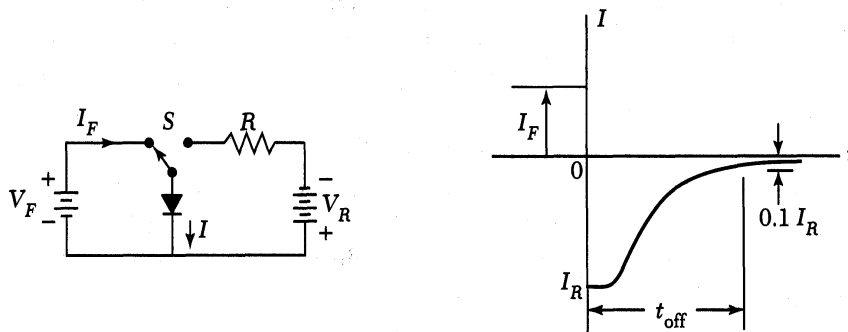


Fig. 22 Transient behavior of a *p-n* junction. (a) Basic switching circuit. (b) Transient response of the current switched from forward bias to reverse bias.

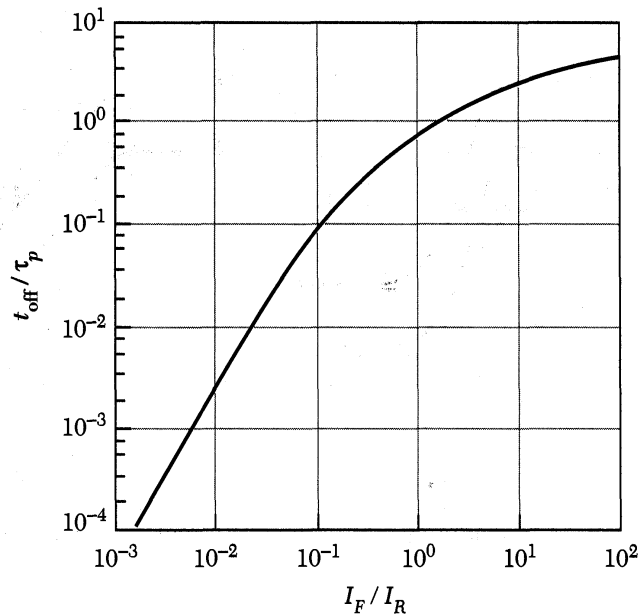


Fig. 23 Normalized transient time versus the ratio of forward current to reverse current.³

Thus the turn-off time depends on both the ratio of forward to reverse currents and the lifetime of the minority carriers. The result of a more precise turn-off time calculation³ accounting for the time-dependent minority-carrier diffusion problem is shown in Fig. 23. For fast-switching devices, we must reduce the lifetime of the minority carriers. Therefore, recombination-generation centers that have energy levels located near mid-bandgap, such as gold in silicon, are usually introduced.

4.7 JUNCTION BREAKDOWN

When a sufficiently large reverse voltage is applied to a *p-n* junction, the junction breaks down and conducts a very large current. Although the breakdown process is not inherently destructive, the maximum current must be limited by an external circuit to avoid excessive junction heating. Two important breakdown mechanisms are the tunneling effect and avalanche multiplication. We consider the first mechanism briefly and then discuss avalanche multiplication in detail, because avalanche breakdown imposes an upper limit on the reverse bias for most diodes. Avalanche breakdown also limits the collector voltage of a bipolar transistor (Chapter 5) and the drain voltage of a MOSFET (Chapter 6). In addition, the avalanche multiplication mechanisms can generate microwave power, as in an IMPATT diode (Chapter 8), and detect optical signals, as in an avalanche photodetector (Chapter 9).

4.7.1 Tunneling Effect

When a high electric field is applied to a *p-n* junction in the reverse direction, a valence electron can make a transition from the valence band to the conduction band, as shown in Fig. 24a. This process, in which an electron penetrates through the energy bandgap, is called tunneling.

The tunneling process is discussed in Chapter 3. Tunneling occurs only if the electric field is very high. The typical field for silicon and gallium arsenide is about 10^6 V/cm or higher. To achieve such a high field, the doping concentrations for both *p*- and *n*-regions

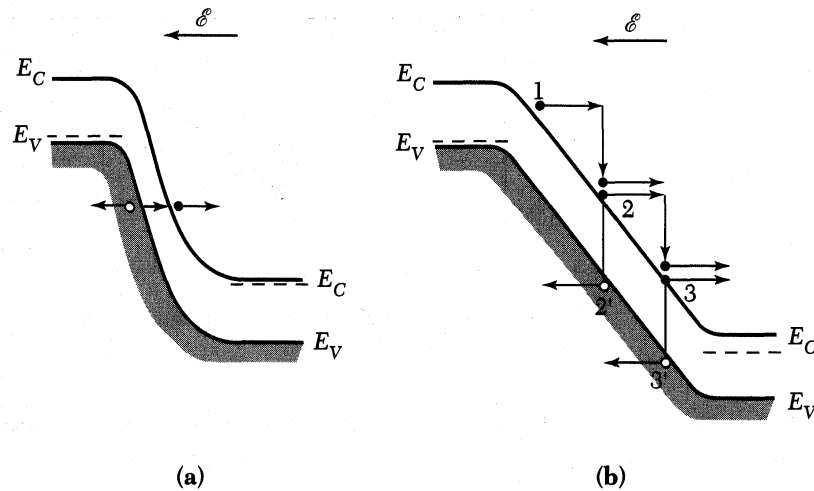


Fig. 24 Energy band diagrams under junction-breakdown conditions. (a) Tunneling effect. (b) Avalanche multiplication.

must be quite high ($>5 \times 10^{17} \text{ cm}^{-3}$). The breakdown mechanisms for silicon and gallium arsenide junctions with breakdown voltages of less than about $4E_g/q$, where E_g is the bandgap, are the result of the tunneling effect. For junctions with breakdown voltages in excess of $6E_g/q$, the breakdown mechanism is the result of avalanche multiplication. At voltages between 4 and $6E_g/q$, the breakdown is due to a mixture of both avalanche multiplication and tunneling.⁴

4.7.2 Avalanche Multiplication

The avalanche multiplication process is illustrated in Fig. 24b. The *p-n* junction such as a p^+-n one-sided abrupt junction with a doping concentration of $N_D \cong 10^{17} \text{ cm}^{-3}$ or less is under reverse bias. This figure is essentially the same as Fig. 25 in Chapter 3. A thermally generated electron in the depletion region (designated by 1) gains kinetic energy from the electric field. If the field is sufficiently high, the electron can gain enough kinetic energy that on collision with an atom, it can break the lattice bonds, creating an electron-hole pair (2 and 2'). These newly created electron and hole both acquire kinetic energy from the field and create additional electron-hole pairs (e.g., 3 and 3'). These in turn continue the process, creating other electron-hole pairs. This process is therefore called *avalanche multiplication*.

To derive the breakdown condition, we assume that a current I_{no} is incident at the left-hand side of the depletion region of width W , as shown in Fig. 25. If the electric field in the depletion region is high enough to initiate the avalanche multiplication process, the electron current I_n will increase with distance through the depletion region to reach a value $M_n I_{no}$ at W , where M_n , the multiplication factor, is defined as

$$M_n \equiv \frac{I_n(W)}{I_{no}}. \quad (81)$$

Similarly, the hole current I_p increases from $x = W$ to $x = 0$. The total current $I = (I_p + I_n)$ is constant at steady state. The incremental electron current at x equals the number of electron-hole pairs generated per second in the distance dx :

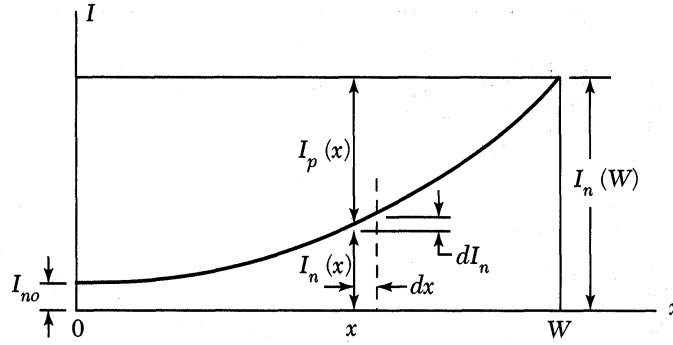


Fig. 25 Depletion region in a *p-n* junction with multiplication of an incident current.

$$d\left(\frac{I_n}{q}\right) = \left(\frac{I_n}{q}\right)(\alpha_n dx) + \left(\frac{I_p}{q}\right)(\alpha_p dx) \quad (82)$$

or

$$\frac{dI_n}{dx} + (\alpha_p - \alpha_n)I_n = \alpha_p I, \quad (82a)$$

where α_n and α_p are the electron and hole ionization rates, respectively. If we use the simplified assumption that $\alpha_n = \alpha_p = \alpha$, the solution of Eq. 82a is

$$\frac{I_n(W) - I_n(0)}{I} = \int_0^W \alpha dx. \quad (83)$$

From Eqs. 81 and 83, we have

$$1 - \frac{1}{M_n} = \int_0^W \alpha dx. \quad (83a)$$

The avalanche breakdown voltage is defined as the voltage where M_n approaches infinity. Hence, the breakdown condition is given by

$$\int_0^W \alpha dx = 1. \quad (84)$$

From both the breakdown condition described above and the field dependence of the ionization rates, we may calculate the critical field (i.e., the maximum electric field at breakdown) at which the avalanche process takes place. Using measured α_n and α_p (Fig. 26 in Chapter 3) the critical field $\mathcal{E}_c$ are calculated for silicon and gallium arsenide one-sided abrupt junctions and shown in Fig. 26 as functions of the impurity concentration of the substrate. Also indicated is the critical field for the tunneling effect. It is evident that tunneling occurs only in semiconductors having high doping concentrations.

With the critical field determined, we may calculate the breakdown voltages. As discussed previously, voltages in the depletion region are determined from the solution of Poisson's equation:

$$V_B(\text{breakdown voltage}) = \frac{\mathcal{E}_c W}{2} = \frac{\epsilon_s \mathcal{E}_c^2}{2q} (N_B)^{-1} \quad (85)$$

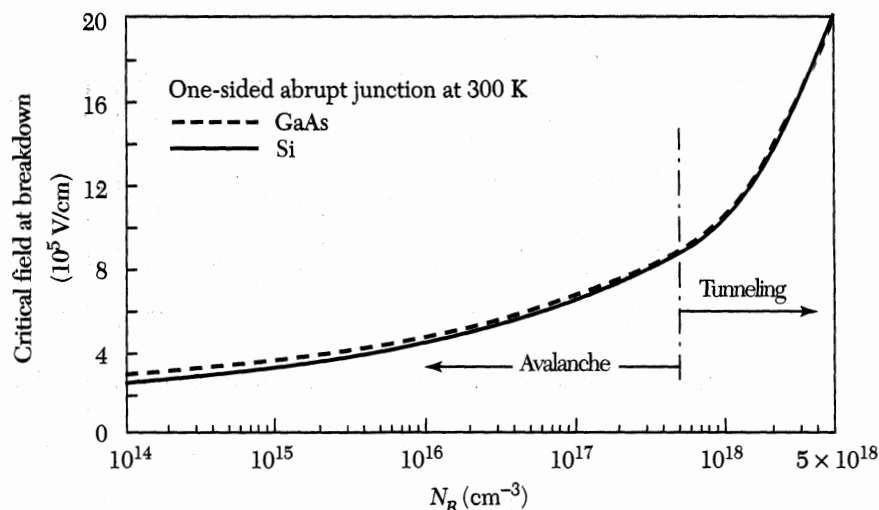


Fig. 26 Critical field at breakdown versus background doping for Si and GaAs one-sided abrupt junctions.⁵

for one-sided abrupt junctions and

$$V_B = \frac{2\mathcal{E}_c W}{3} = \frac{4\mathcal{E}_c^{3/2}}{3} \left(\frac{2\epsilon_s}{q} \right)^{1/2} (a)^{-1/2} \quad (86)$$

for linearly graded junctions, where N_B is the background doping of the lightly doped side, ϵ_s is the semiconductor permittivity, and a is the impurity gradient. Since the critical field is a slowly varying function of either N_B or a , the breakdown voltage, as a first-order approximation, varies as N_B^{-1} for abrupt junctions and as $a^{-1/2}$ for linearly graded junctions.

Figure 27 shows the calculated avalanche breakdown voltages for silicon and gallium arsenide junctions.⁵ The dash-dot line (to the right) at high dopings or high-impurity gradients indicates the onset of the tunneling effect. Gallium arsenide has higher breakdown voltages than silicon for a given N_B or a , mainly because of its larger bandgap. The larger the bandgap, the larger the critical field must be for sufficient kinetic energy to be gained between collisions. As Eqs. 85 and 86 demonstrate, the larger critical field, in turn, gives rise to higher breakdown voltage.

The inset of Fig. 28 shows the space-charge distribution of a diffused junction with a linear gradient near the surface and a constant doping inside the semiconductor. The breakdown voltage lies between the two limiting cases of abrupt junction and linearly graded junction considered previously.⁶ For large a and low N_B , the breakdown voltage of the diffused junctions is given by the abrupt junction results shown on the bottom line in Fig. 28, whereas for small a and high N_B , V_B is given by the linearly graded junction results indicated by the parallel lines in Fig. 28.

► EXAMPLE 8

Calculate the breakdown voltage for a Si one-sided p^+n abrupt junctions with $N_D = 5 \times 10^{16} \text{ cm}^{-3}$.

SOLUTION From Fig. 26, we see that the critical field at breakdown for a Si one-sided abrupt junction is about $5.7 \times 10^5 \text{ V/cm}$. Then from Eq. 85, we obtain

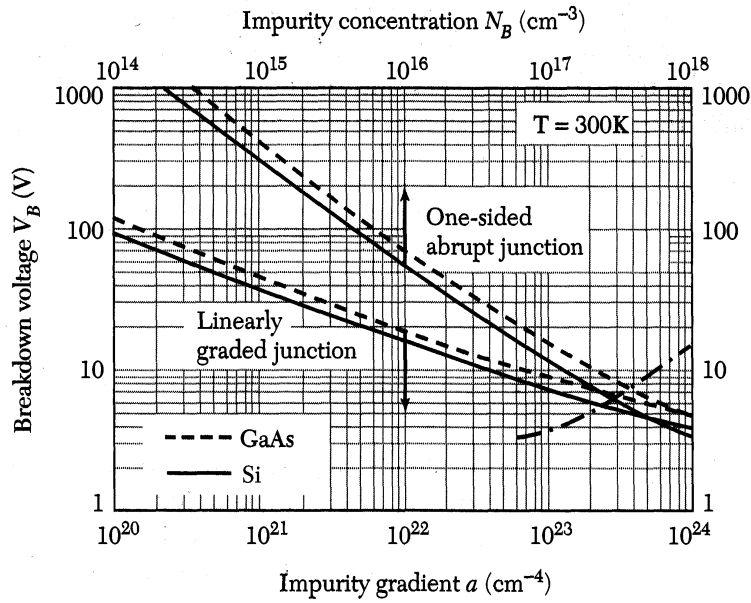


Fig. 27 Avalanche breakdown voltage versus impurity concentration for one-sided abrupt junction and avalanche breakdown voltage versus impurity gradient for linearly graded junction in Si and GaAs. Dash-dot line indicates the onset of the tunneling mechanism.⁵

$$\begin{aligned}
 V_B(\text{breakdown voltage}) &= \frac{\mathcal{E}_c W}{2} = \frac{\epsilon_s \mathcal{E}_c^2}{2q} (N_B)^{-1}, \\
 &= \frac{11.9 \times 8.85 \times 10^{-14} \times (5.7 \times 10^5)^2}{2 \times 1.6 \times 10^{-19}} (5 \times 10^{16})^{-1} \\
 &= 21.4 \text{ V}.
 \end{aligned}$$

In Figs. 27 and 28 we assume that the semiconductor layer is thick enough to support the reverse-biased depletion layer width W_m at breakdown. If the semiconductor

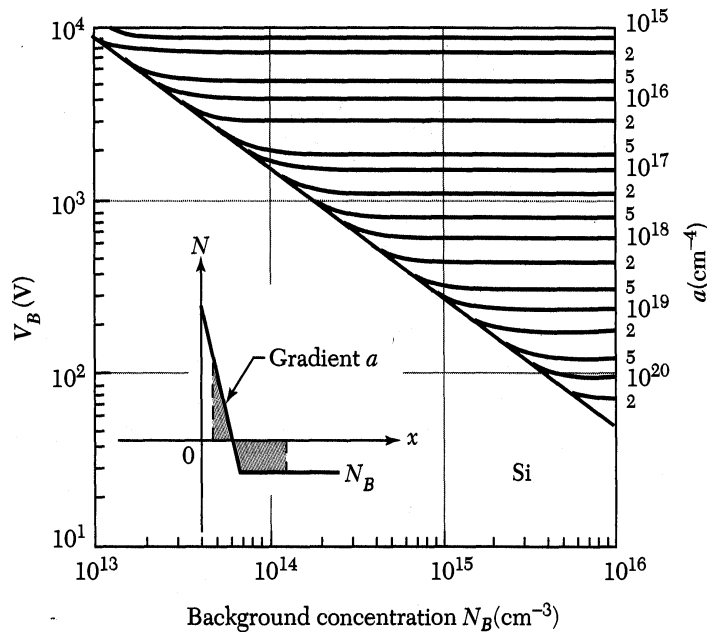


Fig. 28 Breakdown voltage for diffused junctions. Inset shows the space charge distribution.⁶

layer W is smaller than W_m , as shown in the inset of Fig. 29, the device will be punched through; that is, the depletion layer will reach the $n-n^+$ interface prior to breakdown. Increase the reverse bias further and the device will break down. The critical field $\mathcal{E}_c$ is essentially the same as that shown in Fig. 26. Therefore, the breakdown voltage V'_B for the punch-through diode is

$$\begin{aligned} \frac{V'_B}{V_B} &= \frac{\text{shaded area in Fig. 29 inset}}{(\mathcal{E}_c W_m) / 2}, \\ &= \left(\frac{W}{W_m} \right) \left(2 - \frac{W}{W_m} \right). \end{aligned} \quad (87)$$

Punch-through occurs when the doping concentration N_B becomes sufficiently low, as in a $p^+-\pi-n^+$ or p^+-v-n^+ diode, where π stands for a lightly doped p -type and v stands for a lightly doped n -type semiconductor. The breakdown voltages for such diodes calculated from Eqs. 85 and 87 are shown in Fig. 29. For a given thickness, the breakdown voltage approaches a constant value as the doping decreases.

► EXAMPLE 9

For a GaAs p^+-n one-sided abrupt junction with $N_D = 8 \times 10^{14} \text{ cm}^{-3}$, calculate the depletion width at breakdown. If the n -type region of this structure is reduced to $20 \mu\text{m}$, calculate the breakdown voltage.

SOLUTION From Fig. 27, we can find the breakdown voltage (V_B) is about 500 V, which is much larger than the built-in voltage (V_{bi}). And from Eq. 27, we obtain

$$W = \sqrt{\frac{2\epsilon_s(V_{bi} - V)}{qN_B}} \cong \sqrt{\frac{2 \times 12.4 \times 8.85 \times 10^{-14} \times 500}{1.6 \times 10^{-19} \times 8 \times 10^{14}}} = 2.93 \times 10^{-3} = 29.3 \mu\text{m}.$$

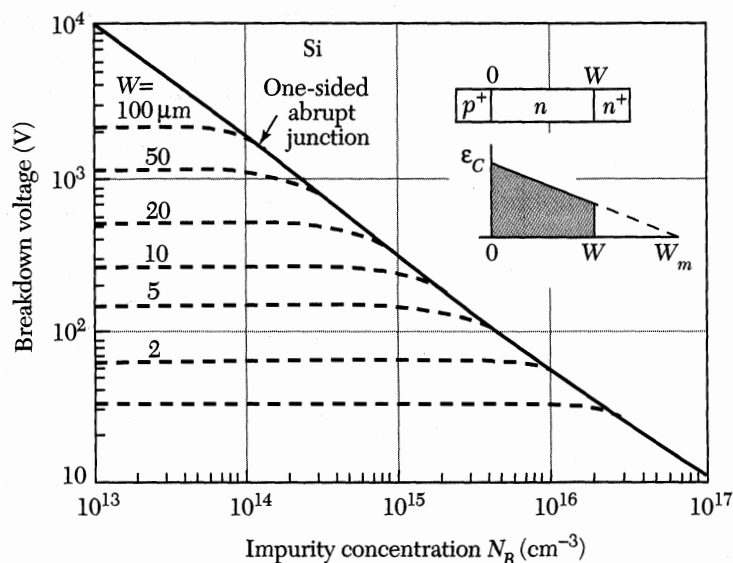


Fig. 29 Breakdown voltage for $p^+-\pi-n^+$ and p^+-v-n^+ junctions. W is the thickness of the lightly doped p -type (π) or the lightly doped n -type (v) region.

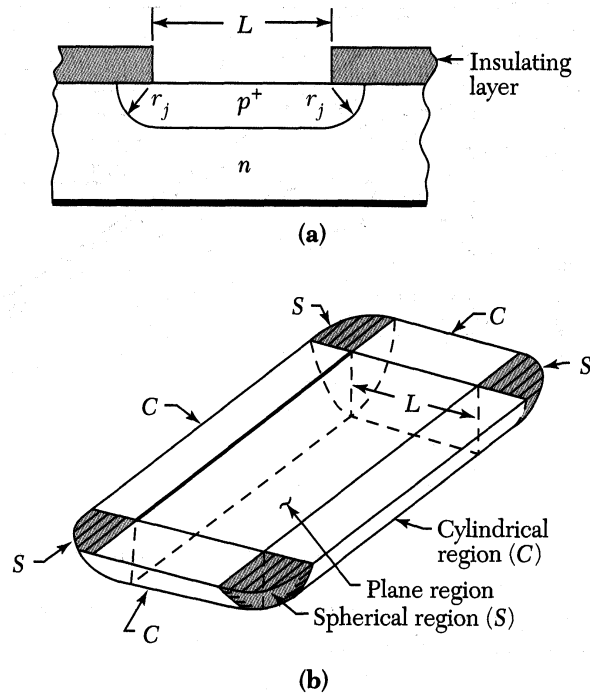


Fig. 30 (a) Planar diffusion process that forms junction curvature near the edge of the diffusion mask, where r_j is the radius of curvature. (b) Cylindrical and spherical regions formed by diffusion through a rectangular mask.

When the n -type region reduces to $20\text{ }\mu\text{m}$, the punch-through will occur first. From Eq. 87, we can obtain

$$\frac{V'_B}{V_B} = \frac{\text{shaded area in Fig. 29 inset}}{(\mathcal{E}_c W_m) / 2} = \left(\frac{W}{W_m} \right) \left(2 - \frac{W}{W_m} \right),$$

$$V'_B = V_B \left(\frac{W}{W_m} \right) \left(2 - \frac{W}{W_m} \right) = 500 \times \left(\frac{20}{29.3} \right) \left(2 - \frac{20}{29.3} \right) = 449\text{ V.}$$

Another important consideration of breakdown voltage is the junction curvature effect.⁷ When a p - n junction is formed by diffusion through a window in the insulating layer on a semiconductor, the impurities diffuse downward and sideways (see Chapter 13). Hence, the junction has a plane (or flat) region with nearly cylindrical edges, as shown in Fig. 30a. If the diffusion mask contains sharp corners, the corner of the junction will acquire the roughly spherical shape shown in Fig. 30b. Because the spherical or cylindrical regions of the junction have a higher field intensity, they determine the avalanche breakdown voltage. The calculated results for silicon one-sided abrupt junctions are shown in Fig. 31. The solid line represents the plane junctions considered previously. Note that as the junction radius r_j becomes smaller, the breakdown voltage decreases dramatically, especially for spherical junctions at low impurity concentrations.

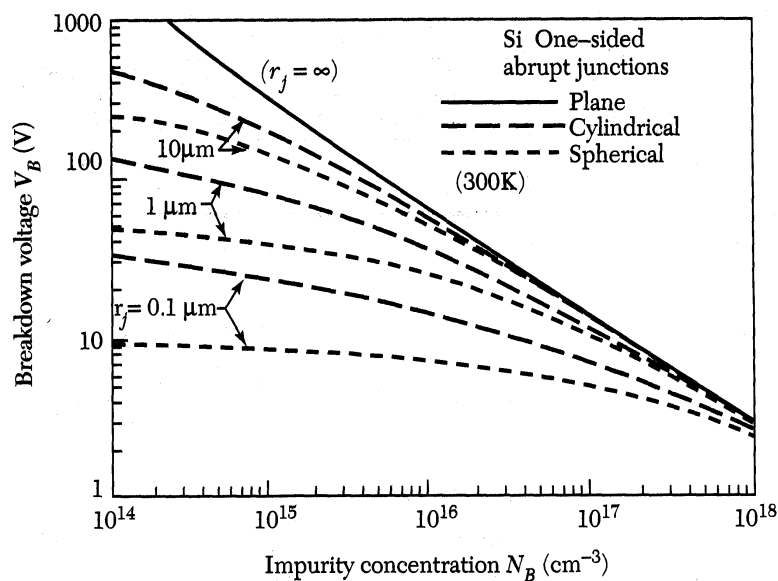


Fig. 31 Breakdown voltage versus impurity concentration for one-sided abrupt doping profile with cylindrical and spherical junction geometries,⁷ where r_j is the radius of curvature indicated in Fig. 30.

► 4.8 HETEROJUNCTION

A heterojunction is defined as a junction formed between two dissimilar semiconductors. Figure 32a shows the energy band diagram of two isolated pieces of semiconductors prior to the formation of a heterojunction. The two semiconductors are assumed to have different energy bandgaps E_g , different dielectric permittivities ϵ_s , different work function $q\phi_s$, and different electron affinities $q\chi$. The work function is defined as the energy required to remove an electron from Fermi level E_F to a position just outside the material (the vacuum level). The electron affinity is the energy required to remove an electron from the bottom of the conduction band E_C to the vacuum level. The difference in energy of the conduction band edges in the two semiconductors is represented by ΔE_C , and the difference in energy in the valence band edges is represented by ΔE_V . From Fig. 32a, ΔE_C and ΔE_V can be expressed by

$$\Delta E_C = q(\chi_2 - \chi_1) \quad (88a)$$

and

$$\Delta E_V = E_{g1} + q\chi_1 - (E_{g2} + q\chi_2) = \Delta E_g - \Delta E_C, \quad (88b)$$

where ΔE_g is the energy band difference and $\Delta E_g = E_{g1} - E_{g2}$.

Figure 32b shows the equilibrium band diagram of an ideal abrupt heterojunction formed between these semiconductors.⁸ In this diagram it is assumed that there is a negligible number of traps or generation-recombination centers at the interface of the two dissimilar semiconductors. Note that this assumption is valid only when heterojunctions are formed between semiconductors with closely matched lattice constants. Therefore, we must choose lattice-matched materials to satisfy the assumption.[§] For example, the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ materials, with x from 0 to 1, is the most important material for heterojunc-

[§]The lattice-mismatched epitaxy, also called the strained-layer epitaxy, is considered in Section 10.6.

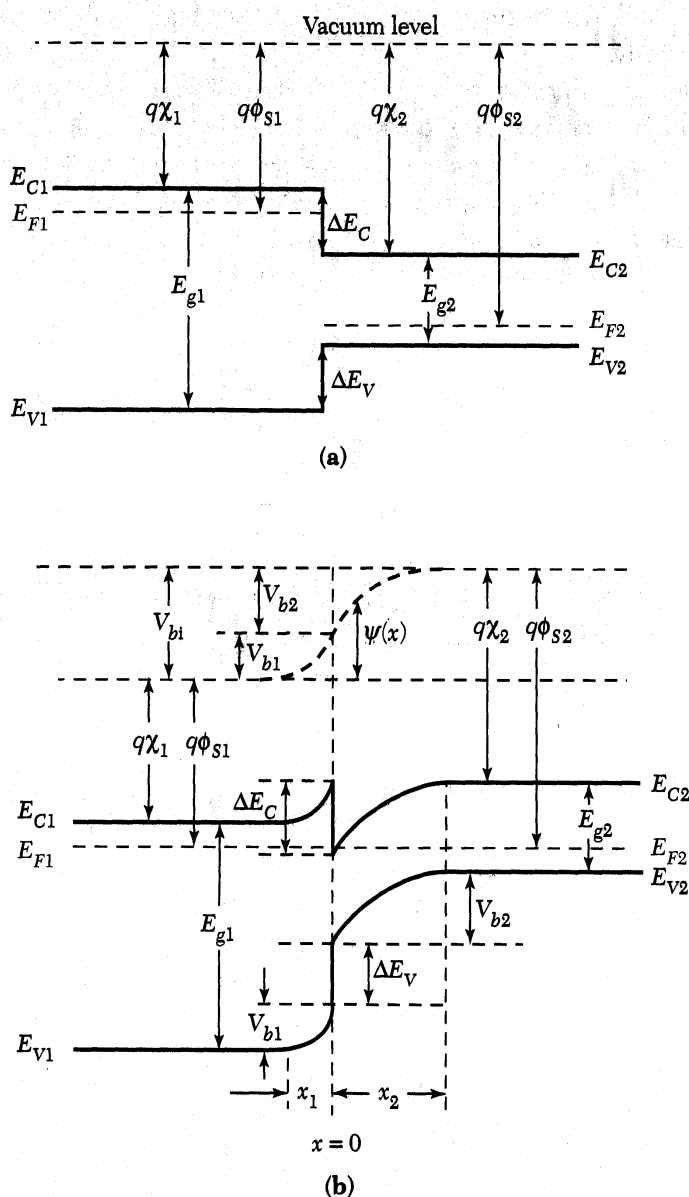


Fig. 32 (a) Energy band diagram of two isolated semiconductors. (b) Energy band diagram of an ideal *n-p* heterojunction at thermal equilibrium.

tions. When $x = 0$, we have GaAs, with a bandgap of 1.42 eV and a lattice constant of 5.6533 Å at 300 K. When $x = 1$, we have AlAs, with a bandgap of 2.17 eV and a lattice constant of 5.6605 Å. The bandgap for the ternary $\text{Al}_x\text{Ga}_{1-x}\text{As}$ increases with x ; however, the lattice constant remains essentially a constant. Even for the extreme cases where $x = 0$ and $x = 1$, the lattice constant mismatch is only 0.1%.

There are two basic requirements in the construction of the energy band diagram: (a) the Fermi level must be the same on both sides of the interface in thermal equilibrium, and (b) the vacuum level must be continuous and parallel to the band edges. Because of these requirements, the discontinuity in conduction band edges ΔE_C and valence band edges ΔE_V will be unaffected by doping as long as the bandgap E_g and electron affinity $q\chi$ are not functions of doping (i.e., as in nondegenerate semiconductors). The total built-in potential V_{bi} can be expressed by

$$V_{bi} = V_{b1} + V_{b2}, \quad (89)$$

where V_{b1} and V_{b2} are the electrostatic potentials at equilibrium in semiconductors 1 and 2, respectively.

Under the conditions that the potential and the *free-carrier flux density* (defined as the rate of free-carrier flow through a unit area) are continuous at the heterointerface, we can derive the depletion widths and capacitance from the Poisson equation using the conventional depletion approximation. One boundary condition is the continuity of electric displacement, that is, $\epsilon_1 \mathcal{E}_1 = \epsilon_2 \mathcal{E}_2$, where $\mathcal{E}_1$ and $\mathcal{E}_2$ are the electric fields at the interface ($x = 0$) in semiconductors 1 and 2, respectively. V_{b1} and V_{b2} are given by

$$V_{b1} = \frac{\epsilon_2 N_2 (V_{bi} - V)}{\epsilon_1 N_1 + \epsilon_2 N_2}, \quad (90a)$$

$$V_{b2} = \frac{\epsilon_1 N_1 (V_{bi} - V)}{\epsilon_1 N_1 + \epsilon_2 N_2}, \quad (90b)$$

where N_1 and N_2 are the doping concentrations in semiconductors 1 and 2, respectively. The depletion widths x_1 and x_2 can be obtained by

$$x_1 = \sqrt{\frac{2\epsilon_1 \epsilon_2 N_2 (V_{bi} - V)}{q N_1 (\epsilon_1 N_1 + \epsilon_2 N_2)}} \quad (91a)$$

and

$$x_2 = \sqrt{\frac{2\epsilon_1 \epsilon_2 N_1 (V_{bi} - V)}{q N_2 (\epsilon_1 N_1 + \epsilon_2 N_2)}} \quad (91b)$$

► EXAMPLE 10

Consider an ideal abrupt heterojunction with a built-in potential of 1.6 V. The impurity concentrations in semiconductor 1 and 2 are 1×10^{16} donors/cm³ and 3×10^{19} acceptors/cm³, and the dielectric constants are 12 and 13, respectively. Find the electrostatic potential and depletion width in each material at thermal equilibrium.

SOLUTION From Eq. 90, the electrostatic potentials of a heterojunction at thermal equilibrium or $V = 0$ are

$$V_{b1} = \frac{13 \times (3 \times 10^{19}) \times 1.6}{12 \times (1 \times 10^{16}) + 13 \times (3 \times 10^{19})} = 1.6 \text{ V}$$

and

$$V_{b2} = \frac{12 \times (1 \times 10^{16}) \times 1.6}{12 \times (1 \times 10^{16}) + 13 \times (3 \times 10^{19})} = 4.9 \times 10^{-4} \text{ V.}$$

The depletion widths can be calculated by Eq. 91:

$$x_1 = \sqrt{\frac{2 \times 12 \times 13 \times (8.85 \times 10^{-14}) \times (3 \times 10^{19}) \times 1.6}{(1.6 \times 10^{-19}) \times (1 \times 10^{16}) \times (12 \cdot 1 \times 10^{16} + 13 \cdot 3 \times 10^{19})}} = 4.608 \times 10^{-5} \text{ cm,}$$

$$x_2 = \sqrt{\frac{2 \times 12 \times 13 \times (8.85 \times 10^{-14}) \times (1 \times 10^{16}) \times 1.6}{(1.6 \times 10^{-19}) \times (3 \times 10^{19}) \times (12 \cdot 1 \times 10^{16} + 13 \cdot 3 \times 10^{19})}} = 1.536 \times 10^{-8} \text{ cm.}$$

We see that most of the built-in potential is in the semiconductor with a lower doping concentration. The depletion width there is also much wider. ◀

SUMMARY

A *p-n* junction is formed when a *p*-type and an *n*-type semiconductor are brought into intimate contact. The *p-n* junction, in addition to being a device that is used in many applications, is the basic building block for other semiconductor devices. Therefore, an understanding of junction theory serves as the foundation to understanding other semiconductor devices.

Most modern *p-n* junctions are fabricated using “planar technology.” This technology includes the thermal oxidation process to grow an oxide layer on the semiconductor surface, the lithography process to open a window in the oxide, the diffusion or ion implantation process to form a *p-n* junction in the window area, and the metallization process to provide contacts to connect the junction to other circuit elements. Planar technology is covered in detail in Chapters 10–14.

When a *p-n* junction is formed, there are uncompensated negative ions (N_A^-) on the *p*-side and uncompensated positive ions (N_D^+) on the *n*-side. Therefore, a depletion region (i.e., depletion of mobile carriers) is formed at the junction. This region, in turn, creates an electric field. At thermal equilibrium, the drift current due to the electric field is exactly balanced by the diffusion current due to concentration gradients of the mobile carriers on the two sides of the junction. When a positive voltage is applied to the *p*-side with respect to the *n*-side, a large current will flow through the junction. However, when a negative voltage is applied, virtually no current flows. This “rectifying” behavior is the most important characteristic of *p-n* junctions.

The basic equations presented in Chapters 2 and 3 have been used to develop the ideal static and dynamic behaviors of *p-n* junctions. We derived expressions for the depletion region, the depletion capacitance, and the ideal current-voltage characteristics of *p-n* junctions. However, practical devices depart from these ideal characteristics because of carrier generation and recombination in the depletion layers, high injection under forward bias, and series-resistance effects. The theory and methods of calculating the effects of these departures from the ideal are discussed in detail. We also considered other factors that influence *p-n* junctions, such as minority-carrier storage, diffusion capacitance, and transient behavior in high-frequency and switching applications.

A limiting factor in the operation of *p-n* junctions is junction breakdown—especially that due to avalanche multiplication. When a sufficiently large reverse voltage is applied to a *p-n* junction, the junction breaks down and conducts a very large current. Therefore, the breakdown voltage imposes an upper limit on the reverse bias for *p-n* junctions. We derived equations for the breakdown condition of the *p-n* junction and have shown the effect of device geometry and doping on the breakdown voltage.

A related device is the heterojunction formed between two dissimilar semiconductors. We obtained expressions for its electrostatic potentials and depletion widths. These expressions are simplified to that for a conventional *p-n* junction when these two semiconductors become identical.

11.4.3 Copper Metallization

It is well known that both high conductivity wiring and low-dielectric-constant insulators are required to lower the RC time delay of the interconnect network. Copper is the obvious choice for a new interconnection metallization because it has higher conductivity and higher electromigration resistance than aluminum. Copper can be deposited by PVD, CVD, and electrochemical methods. However, the use of Cu as an alternative material to Al in ULSI circuits has drawbacks, such as its tendency to corrode under standard chip manufacture conditions, its lack of a feasible dry-etching method or a stable self-passivating oxide similar to Al_2O_3 on Al, and its poor adhesion to dielectric materials, such as SiO_2 and low- k polymers. In this section, we discuss the copper metallization techniques.

Several different techniques for fabrication of multilevel Cu interconnects have been reported.^{17,18} The first method is a conventional method to pattern the metal lines followed by dielectric deposition. The second method is to pattern the dielectric layer first and fill copper metal into trenches. This step is followed by chemical mechanical polishing, discussed in Section 11.4.5, to remove the excess metal on the top surface of dielectric and leave Cu material in the holes and trenches. This method is also known as a damascene process.

Damascene technology

The approach for fabricating a copper/low- k dielectric interconnect structure is by the “damascene” or “dual damascene.” Figure 23 shows the dual damascene sequence for an advanced Cu interconnection structure. For a typical damascene structure, trenches for metal lines are defined and etched in the interlayer dielectric (ILD) and then followed by metal deposition of TaN/Cu. The TaN layer serves as a diffusion barrier layer and prevents copper from penetrating the low- k dielectric. The excess copper metal on the surface is removed to obtain a planar structure with metal inlays in the dielectric.

For the dual damascene process, the vias and trenches in the dielectric are defined using two lithography and reactive ion etching (RIE) steps before depositing the Cu metal (Fig. 23a–c). Then a Cu chemical-mechanical polishing process is used to remove the metal on the top surface, leaving the planarized wiring and via imbedded in the insulator.¹⁹ One special benefit of dual damascene is that the via plug is now of the same material as the metal line and the risk of via electromigration failure is reduced.

► EXAMPLE 5

If we replace Al with Cu wire associated with some low- k dielectric ($k = 2.6$) instead of SiO_2 layer, what percentage of reduction of RC time constant will be achieved? (The resistivity of Al is $2.7 \mu\Omega\text{-cm}$, and the resistivity of Cu is $1.7 \mu\Omega\text{-cm}$.)

SOLUTION

$$\frac{1.7}{2.7} \times \frac{2.6}{3.9} \times 100\% = 42\%.$$

11.4.5 Chemical-Mechanical Polishing

In recent years, the development of chemical-mechanical polishing (CMP) has become increasingly important for multilevel interconnection because it is the only technology that allows global planarization (i.e., to make a flat surface across the whole wafer). It

the need to pattern the composite polycide gate structure and adds silicide to the source/drain area to reduce the contact resistance.

The silicides are promising materials for ULSI circuits because of their low resistivity and excellent thermal stability. Cobalt silicide has been widely investigated recently because of its lowest resistivity and high-temperature thermal stability. However, cobalt is sensitive to native oxide as well as an oxygen-contained environment, and a large amount of silicon is consumed during silicidation.

► EXAMPLE 7

Calculate the thickness of the cobalt silicide. For a desired sheet resistance of $0.6 \Omega/\text{sq}$, the resistivity is $18 \mu\Omega\text{-cm}$.

SOLUTION The resistivity is equal to the product of the sheet resistance and the film thickness:

$$\rho = R_s \times t.$$

Then

$$t = \frac{\rho}{R_s} = \frac{18 \times 10^{-6}}{0.6} = 3 \times 10^{-5} \text{ cm} = 300 \text{ nm}.$$

► SUMMARY

Modern semiconductor device fabrication requires the use of thin films. Currently, there are four important types of films—thermal oxides, dielectric layers, polycrystalline silicon, and metal films. The major issues related to film formation are low-temperature processing, step coverage, selective deposition, uniformity, film quality, planarization, throughput, and large-wafer capacity.

Thermal oxidation offers the best quality for the Si-SiO₂ interface and has the lowest interface trap density. Therefore, it is used to form the gate oxide and the field oxide. LPCVD of dielectrics and polysilicon offer conformal step coverage. In contrast, PVD and atmospheric-pressure CVD generally result in nonconformal step coverage. CMP offers global planarization and reduces defect density. Conformal step coverage and planarization are also required for precise pattern transfer at the deep-submicron lithography level. Pattern transfer technology is discussed in the next chapter.

To minimize the RC time delay due to parasitic resistance and capacitance, the silicide process for ohmic contacts, copper metallization for interconnects, and low-dielectric-constant materials for interlayer films are extensively used to meet the requirements of the multilevel interconnect structures of ULSI circuits. In addition, we have investigated high-dielectric-constant materials to improve the gate insulator performance and to increase the capacitance per unit area for DRAM.

► REFERENCES

1. E. H. Nicollian and J. R. Brews, *MOS Physics and Technology*, Wiley, New York, 1982.
2. B. E. Deal and A. S. Grove, "General Relationship for the Thermal Oxidation of Silicon," *J. Appl. Phys.*, **36**, 3770 (1965).
3. J. D. Meindl, et al., "Silicon Epitaxy and Oxidation," in F. Van de wiele, W. L. Engl, and P. O. Jespers, Eds., *Process and Device Modeling for Integrated Circuit Design*, Noorhoff, Leyden, 1977.

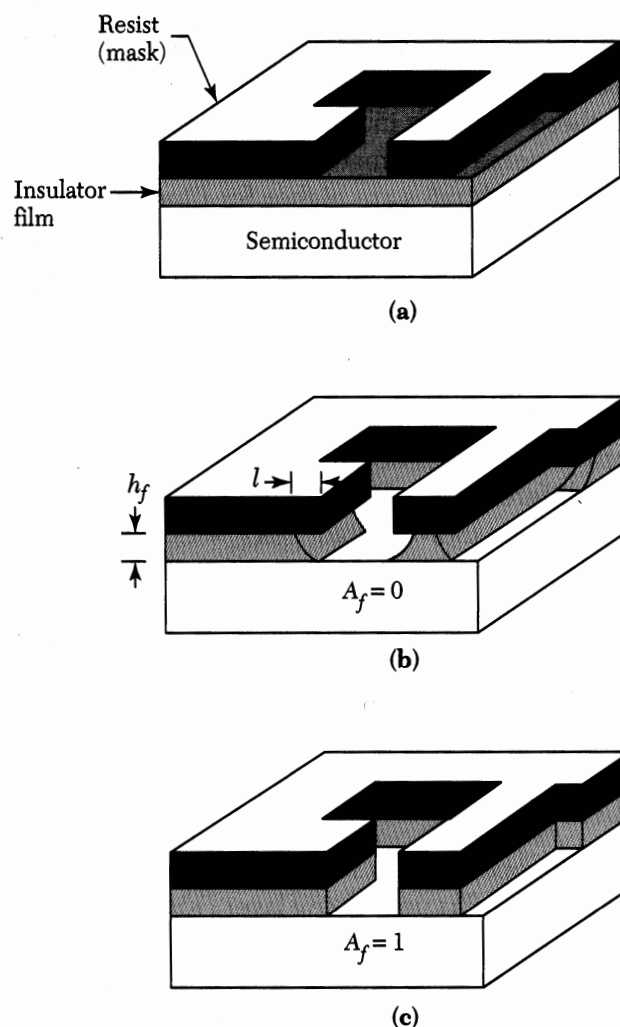


Fig. 23 Comparison of wet chemical etching and dry etching for pattern transfer. ²²

reactive ion etching (RIE), sputter etching, magnetically enhanced RIE (MERIE), reactive ion beam etching, and high-density plasma (HDP) etching.

12.4.1 Plasma Fundamentals

Plasma is a fully or partially ionized gas composed of equal numbers of positive and negative charges and a different number of unionized molecules. Plasma is produced when an electric field of sufficient magnitude is applied to a gas, causing the gas to break down and become ionized. The plasma is initiated by free electrons that are released by some means, such as field emission from a negatively biased electrode. The free electrons gain kinetic energy from the electric field. In the course of their travel through the gas, the electrons collide with gas molecules and lose their energy. The energy transferred in the collision causes the gas molecules to be ionized (i.e., to free electrons). The free electrons gain kinetic energy from the field, and the process continues. Therefore, when the applied voltage is larger than the breakdown potential, a sustained plasma is formed throughout the reaction chamber.

The electron concentrations in the plasma for dry etchings are relatively low, typically on the order of 10^9 – 10^{12} cm^{-3} . At a pressure of 1 Torr, the concentrations of gas molecules are 10^4 – 10^7 times higher than the electron concentrations. This results in an

average gas temperature in the range of 50°–100°C. Therefore, the plasma-assisted dry etching is a low-temperature process.

► EXAMPLE 4

The electron densities in a RIE system and HDP system range from 10^9 – 10^{10} and 10^{11} – 10^{12} cm^{-3} , respectively. Assuming the RIE chamber pressure is 200 mTorr and HDP chamber pressure is 5 mTorr, calculate the ionization efficiency in RIE reactors and HDP reactors at room temperature. The ionization efficiency is the ratio of the electron density to the density of molecules.

SOLUTION

$$PV = nRT$$

where P is the pressure in atm (1 atm = 760,000 mTorr), V the volume in liters, n the number of moles, R the gas constant (0.082 liter·atm/mol·K), and T the absolute temperature in K, respectively.

For a RIE system,

$$\begin{aligned} n/V = P/RT &= (200/760,000)/(0.082 \times 300) = 1.06 \times 10^{-5} \text{ (mol/liter)}, \\ &= 1.06 \times 10^{-5} \times 6.02 \times 10^{23} \div 1000, \\ &= 6.38 \times 10^{15} \text{ (cm}^{-3}\text{)}. \end{aligned}$$

$$\begin{aligned} \text{Ionization efficiency} &= (10^9 \sim 10^{10})/(6.38 \times 10^{15}), \\ &= 1.56 \times 10^{-7} \sim 1.56 \times 10^{-6}. \end{aligned}$$

For a HDP system,

$$\begin{aligned} n/V = P/RT &= (5/760,000)/(0.082 \times 300) = 2.66 \times 10^{-7} \text{ (mol/liter)}, \\ &= 2.66 \times 10^{-7} \times 6.02 \times 10^{23} \div 1000, \\ &= 1.6 \times 10^{14} \text{ (cm}^{-3}\text{)}. \end{aligned}$$

$$\begin{aligned} \text{Ionization efficiency} &= (10^{11} \sim 10^{12})/(1.6 \times 10^{14}), \\ &= 6.25 \times 10^{-4} \sim 6.25 \times 10^{-3}. \end{aligned}$$

Therefore, HDP has much higher ionization efficiency than RIE. ◀

12.4.2 Etch Mechanism, Plasma Diagnostics, and End-Point Control

Plasma etching is a process in which a solid film is removed by a chemical reaction with ground-state or excited-state neutral species. Plasma etching is often enhanced or induced by energetic ions generated in a gaseous discharge. The basic etch mechanism, plasma diagnostic, and end-point control are introduced briefly in this section.

Etch Mechanism

Plasma etching proceeds in five steps, as illustrated in Fig. 24: the etchant species is generated in the plasma; the reactant is then transported by diffusion through a stagnant gas layer to the surface. The reactant is adsorbed on the surface; a chemical reaction (along with physical effects such as ion bombardment) follows to form volatile compounds; the compounds are desorbed from the surface, diffused into the bulk gas, and pumped out by the vacuum system.²³

Plasma etching is based on the generation of plasma in a gas at low pressure. Two basic methods are used—physical methods and chemical methods. The former includes sputter etching and the latter includes pure chemical etching. In physical etching, positive

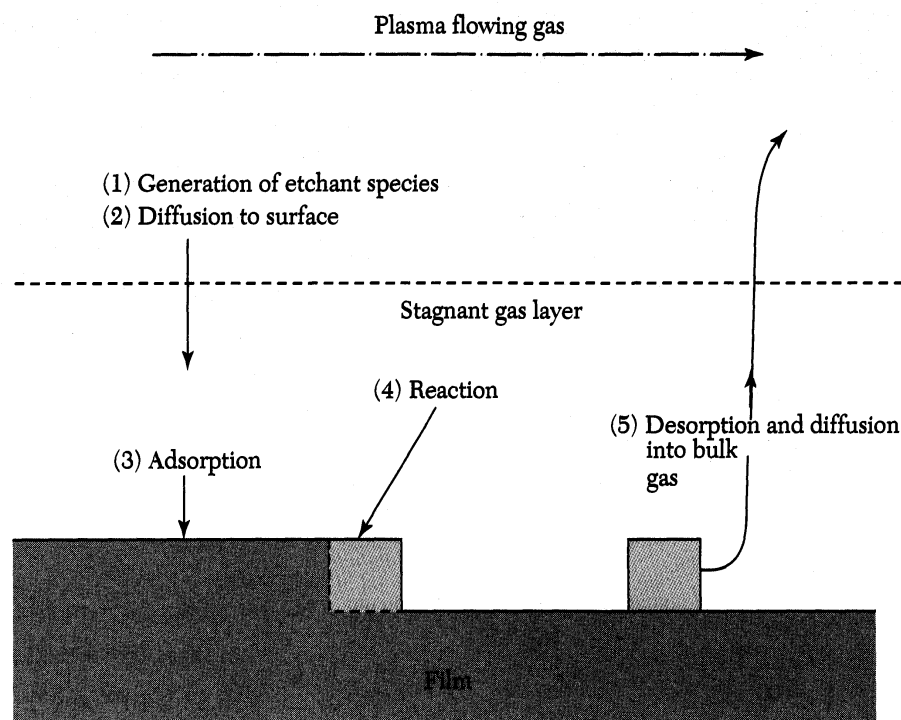


Fig. 24 Basic steps in a dry-etching processing.²³

ions bombard the surface at high speed; small amounts of negative ions formed in the plasma cannot reach the wafer surface and therefore play no direct role in plasma etching. In chemical etching, neutral reactive species generated by the plasma interact with the material surface to form volatile products. Chemical- and physical-etch mechanisms have different characteristics. Chemical etching exhibits a high etch rate, and good selectivity (i.e., the ratio of etch rates for different materials) produces low ion-bombardment-induced damage but yields isotropic profiles. Physical etching can yield anisotropic profiles, but it is associated with low etch selectivity and high bombardment-induced damage. Combinations of chemical and physical etching give anisotropic etch profiles, reasonably good selectivity, and moderate bombardment-induced damage. An example is the RIE, which uses a physical method to assist chemical etching or creates reactive ions to participate in chemical etching

Plasma Diagnostics

Most processing plasmas emit radiation in that range from infrared to ultraviolet. A simple analytical technique is to measure the intensity of these emissions versus wavelength with the aid of optical emission spectroscopy (OES). Using observed spectral peaks, it is usually possible to determine the presence of neutral and ionic species by correlating these emissions with previously determined spectral series. Relative concentrations of the species can be obtained by correlating changes in intensity with the plasma parameter. The emission signal derived from the primary etchant or by-product begins to rise or fall at the end of the etch cycle.

End-Point Control

Dry etching differs from the wet chemical etching, in that dry etching does not have enough etch selectivity to the underlying layer. Therefore, the plasma reactor must be equipped with a monitor that indicates when the etching process is to be terminated

(i.e., an end point detection system). Laser interferometry of the wafer surface is used to continuously monitor etch rates and to determine the end point. During etching, the intensity of laser light reflected off a thin film surface oscillates. This oscillation occurs because of the phase interference between the light reflected from the outer and inner interfaces of the etching layer. This layer must therefore be optically transparent or semi-transparent to observe the oscillation. Figure 25 shows a typical signal from a silicide/polycrystalline Si gate etch. The period of the oscillation is related to the change in film thickness by

$$\Delta d = \lambda / 2\bar{n} \quad (13)$$

where Δd is the change in film thickness for one period of reflected light, λ is the wavelength of the laser light, and $\bar{n}$ is the refractive index of the etching layer. For example, Δd for polysilicon is 80 nm, measured by using a helium-neon laser source for which $\lambda = 632.8$ nm.

12.4.3 Reactive Plasma-Etching Techniques and Equipment

Plasma reactor technology in the IC industry has changed dramatically since the first application of plasma processing to photoresist stripping. A reactor for plasma etching contains a vacuum chamber, pump system, power supply generators, pressure sensors, gas flow control units, and end-point detector. Table 3 shows the similarities and differences in the types of etching equipment that are commercially available. A comparison of pressure operating ranges and ion energies for different types of reactors is shown in Fig. 26. Each etch tool is designed empirically and uses a particular combination of pressure, electrode configuration and type, and source frequency to control the two primary etch mechanisms – chemical and physical. Higher etch rates and tool automation are required for most etchers used in manufacturing.

Reactive Ion Etcher (RIE)

RIE has been extensively used in the microelectronic industry. In a parallel-plate diode system, a radio frequency (rf) capacitive-coupled bottom electrode holds the wafer. This allows the grounded electrode to have a significantly larger area because it is, in fact, the chamber itself. The larger grounded area combined with the lower operating pressure

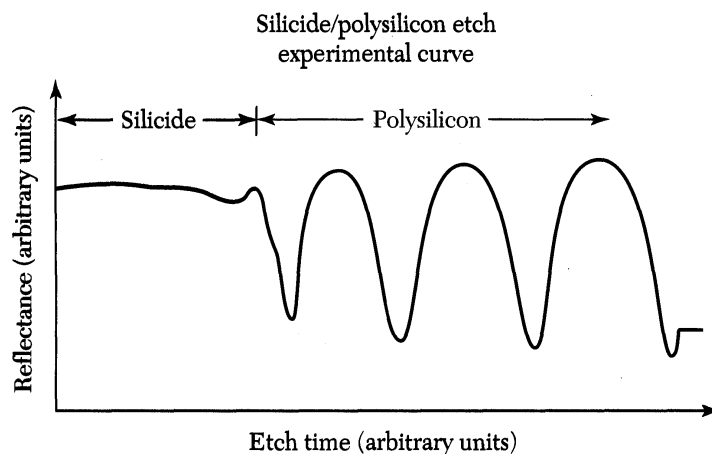
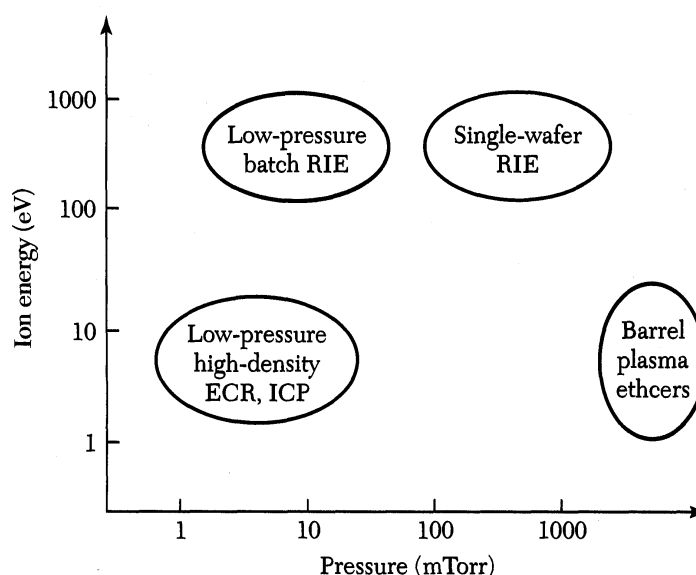


Fig. 25 The relative reflectance of the etching surface of a composite silicide/poly-Si layer. The end point of the etch is indicated by the cessation of the reflectance oscillation.

TABLE 3 Etch Mechanisms, Pressure Ranges of Plasma Reactors

Etch tool configuration	Etch mechanism	Pressure range (Torr)
Barrel etching	Chemical	0.1–10
Downstream plasma etching	Chemical	0.1–10
Reactive ion etching (RIE)	Chemical and physical	0.01–1
Magnetic enhanced RIE	Chemical and physical	0.01–1
Magnetic confinement triode RIE	Chemical and physical	0.001–0.1
Electron cyclotron resonance plasma etch	Chemical and physical	0.001–0.1
Inductively coupled plasma or transformer-coupled plasma	Chemical and physical	0.001–0.1
Surface wave coupled plasma or helicon plasma etching	Chemical and physical	0.001–0.1

**Fig. 26** Comparison of ion energy and operating pressure ranges for different types of plasma reactors.

(< 500 mTorr) causes the wafers to be subjected to a heavy bombardment of energetic ions from the plasma as a result of the large, negative self-bias at the wafer surface.

The etch selectivity of this system is relatively low compared with traditional barrel-etch systems because of strong physical sputtering. However, selectivity can be improved by choosing the proper etch chemistry, for example, by polymerizing the silicon surface with fluorocarbon polymers to obtain selectivity of SiO_2 over silicon. Alternatively, a triode-configuration RIE etch, as shown in Fig. 27, can separate plasma generation from ion transport. Ion energy is controlled through a separate bias on the wafer electrode, thereby minimizing the loss of selectivity and the ion-bombardment-induced damage observed in most traditional RIE systems.

Electron Cyclotron Resonance Plasma Etcher

Most parallel-plate plasma etchers, except triode RIE, do not provide the ability to control plasma parameters, such as electron energy, plasma density, and reactant density inde-

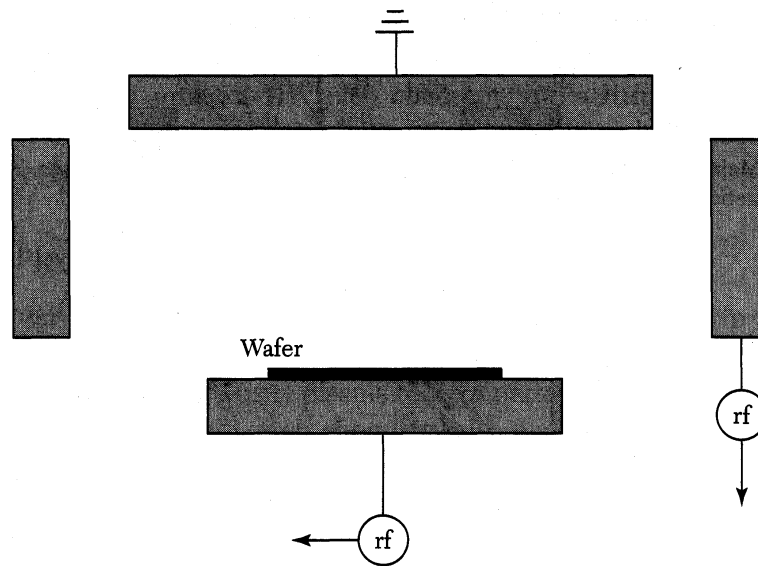


Fig. 27 Schematic of a triode reactive ion etch reactor. The ion energy is separately controlled by a bias voltage on the bottom electrode. rf, radio frequency.

pendently. As a result, ion-bombardment-induced damage becomes a serious problem. The electron cyclotron resonance (ECR) reactor combines microwave power with a static magnetic field to force electrons to circulate around the magnetic field lines at an angular frequency. When this frequency equals to the applied microwave frequency, a resonance coupling occurs between the electron energy and the applied electric field that results in a high degree of dissociation and ionization (10^{-2} for ECR compared with 10^{-6} for RIE). Figure 28 shows an ECR reaction chamber configuration. Microwave power is coupled through a microwave window into the ECR source region. The magnetic field is supplied from the magnetic coils. The ECR plasma systems can also be used in thin film deposition. The high efficiency in exciting the reactants in ECR plasmas allows the deposition of films at room temperature without the need for thermal activation.

Other High-Density Plasma Etchers

As feature sizes for ultralarge-scale integration (ULSI) continue to decrease, the limits of the conventional RIE system are being approached. In addition to the ECR system, other types of high-density plasma (HDP) sources, such as the inductively coupled plasma (ICP) source, the transformer-coupled plasma (TCP) source, and the surface wave-coupled plasma (SWP) sources, have been developed. These etchers have high plasma density (10^{11} – 10^{12} cm^{-3}), and low processing pressure (< 20 mTorr). In addition, they allow the wafer platen to be powered independently of the source, providing significant decoupling between the ion energy (wafer bias) and the ion flux (plasma density primarily driven by source power). The primary processing advantages of HDP sources are better critical dimension (CD) control, higher etching rates, and better selectivity.

In addition, HDP sources provide low substrate damage (because of independent biasing of the substrate and the side-electrode potentials) and high anisotropy (because of low pressure yet high active species density). However, because of their complexity and higher cost, the systems may not be used for less critical applications, such as spacer etching or planarization.²⁴ Figure 29 shows a TCP plasma reactor. A high-density, low-pressure plasma is generated by a flat spiral coil that is separated from the plasma by a dielectric plate on the top of the reactor. The wafer is located away from the coil, so it is not affected by the electromagnetic field generated by the coil. There is little plasma

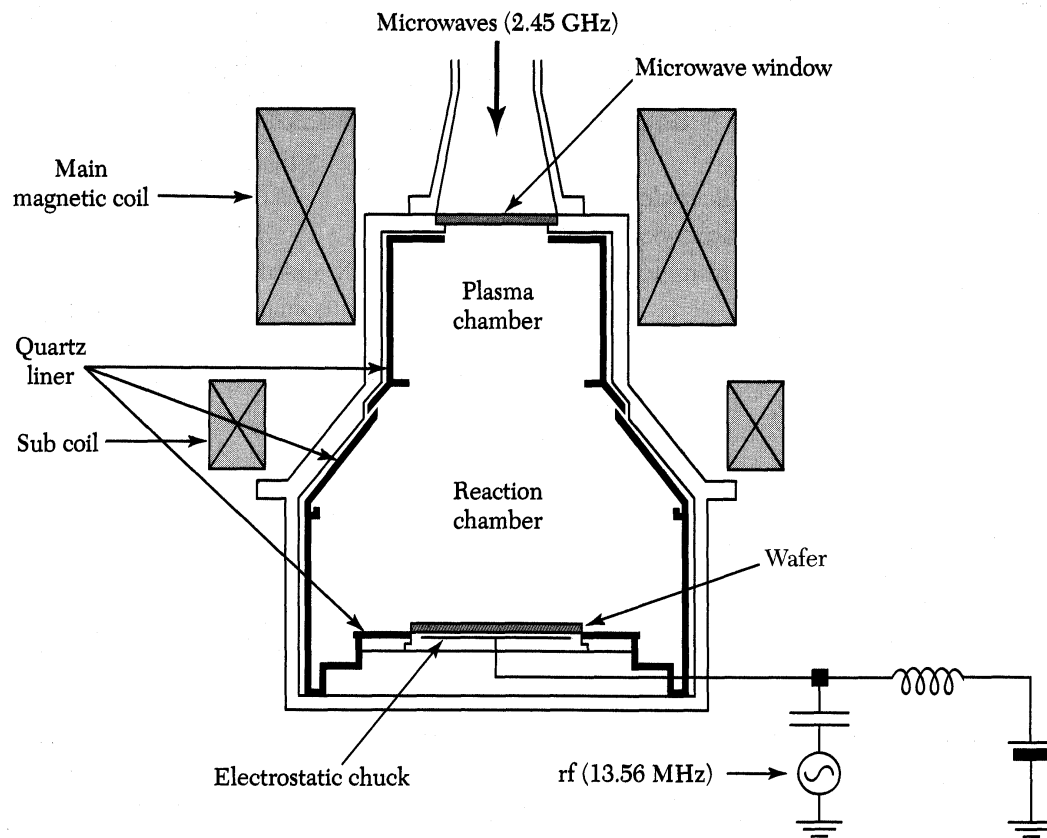


Fig. 28 Schematic of an electron cyclotron resonance reactor.²⁴

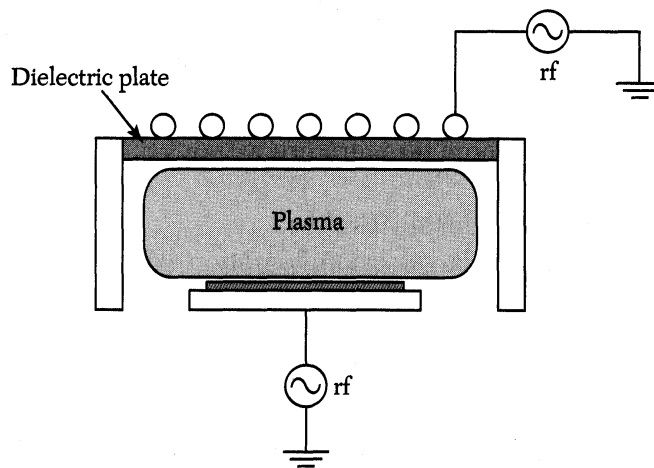


Fig. 29 Schematic of a transformer-coupled plasma reactor.

density loss because plasma is generated only a few mean free paths away from the wafer surface. Therefore, a high-density plasma and high etch rates are achieved.

Clustered Plasma Processing

Semiconductor wafers are processed in clean rooms to minimize exposure to ambient particulate contamination. As device dimensions shrink, particulate contamination becomes a more serious problem. To minimize particulate contamination, clustered plasma tools use a wafer handler to pass wafers from one process chamber to another in a vacuum

environment. The clustered plasma processing tools can also increase throughput. Figure 30 shows the multilayer metal interconnect (TiW/AlCu/TiW) etching process with clustered tools in an AlCu etch chamber, a TiW etch chamber, and a strip passivation chamber. The clustered tools provide an economic advantage by having a high chip yield because the wafer is exposed to less ambient contamination and is handled less.

12.4.4 Reactive Plasma–Etching Applications

Plasma etching has rapidly evolved from a simple and batch resist stripping to a large and single-wafer processing. The etching system continues to be improved from the conventional RIE tool to the high-density plasma tool for pattern transfer of deep-submicron devices. Besides the etching tool, etch chemistry also plays a critical role in the performance of etch process. Table 4 lists some etch chemistries for different etch processes. Developing an etch process usually means optimizing etch rate, selectivity, profile control, critical dimension, damage, etc., by adjusting a large number of process parameters.

Silicon Trench Etching

As device feature size decreases, a corresponding decrease is needed in the wafer surface area occupied by the isolation between circuit elements and the storage capacitor of a DRAM cell. This surface area can be reduced by etching trenches into the silicon substrate and filling them with suitable dielectric or conductive materials. Deep trenches, usually with a depth larger than 5 μm , are used mainly for forming storage capacitors. Shallow trenches, usually with a depth less than 1 μm , are often used for isolation.

Chlorine-based and bromine-based chemistries have a high silicon etch rate and high etch selectivity to the silicon dioxide mask. The combination of $\text{HBr} + \text{NF}_3 + \text{SF}_6 + \text{O}_2$ gas mixtures is used to form a trench capacitor with a depth of $\sim 7 \mu\text{m}$. It is also used for shallow trench isolation etching. Aspect ratio dependent etching (i.e., variation in etch rate with aspect ratio) is often observed in submicron-deep silicon trench etching caused

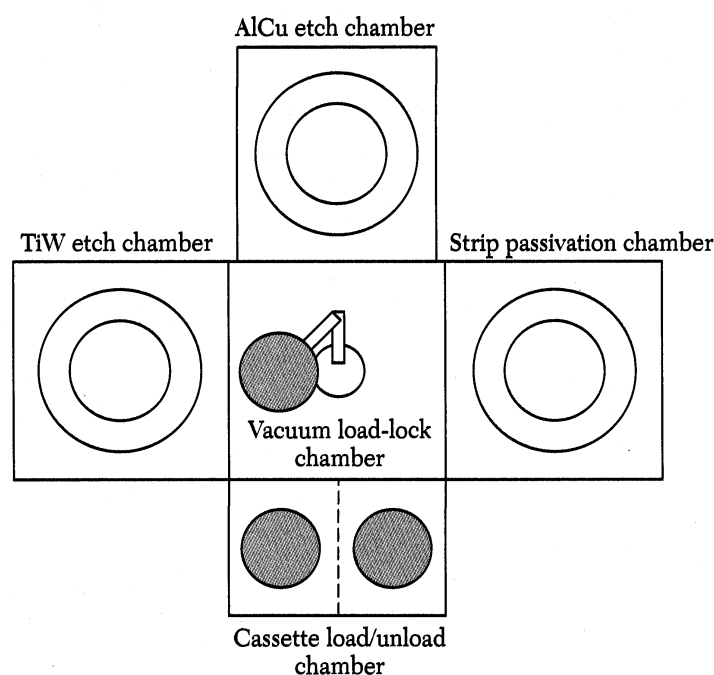


Fig. 30 Cluster reactive ion etch tool for multilayer metal (TiW/AlCu/TiW) interconnect etching.²

TABLE 4 Etch Chemistries of Different Etch Processes

Material being etched	Etching chemistry
Deep Si trench	HBr/NF ₃ /O ₂ /SF ₆
Shallow Si trench	HBr/Cl ₂ /O ₂
Poly Si	HBr/Cl ₂ /O ₂ , HBr/O ₂ , BCl ₃ /Cl ₂ , SF ₆
Al	BCl ₃ /Cl ₂ , SiCl ₄ /Cl ₂ , HBr/Cl ₂
AlSiCu	BCl ₃ /Cl ₂ /N ₂
W	SF ₆ only, NF ₃ /Cl ₂
TiW	SF ₆ only
WSi ₂ , TiSi ₂ , CoSi ₂	CCl ₂ F ₂ /NF ₃ , CF ₄ /Cl ₂ , Cl ₂ /N ₂ /C ₂ F ₆
SiO ₂	CF ₄ /CHF ₃ /Ar, C ₂ F ₆ , C ₃ F ₈ , C ₄ F ₈ /CO, C ₅ F ₈ , CH ₂ F ₂
Si ₃ N ₄	CHF ₃ /O ₂ , CH ₂ F ₂ , CH ₂ CHF ₂

by limited ion and neutral transport within the trench. Figure 31 shows the dependence of average silicon trench etch rate on aspect ratio. Trenches with large aspect ratios are etched more slowly than trenches with small aspect ratios.

Polysilicon and Polycide Gate Etching

Polysilicon or polycide (i.e., low-resistance metal silicides over polysilicon) is usually used as a gate material for MOS devices. Anisotropic etching and high etch selectivity to the gate oxide are the most important requirements for gate etching. For example, the selectivity required in 1G DRAM is more than 150 (i.e., the ratio of etch rates for polycide and gate oxide is 150:1). Achieving high selectivity and etch anisotropy at the same time is difficult for most ion-enhanced etching processes. Therefore, multistep processing is used in which different etch steps in the process are optimized for etch anisotropy and selectivity. On the other hand, the trend in plasma technology for anisotropic etching and high selectivity is to utilize a low-pressure, high-density plasma using a relatively low power. Most chlorine-based and bromine-based chemistries can be used for gate etching to achieve the required etch anisotropy and selectivity.

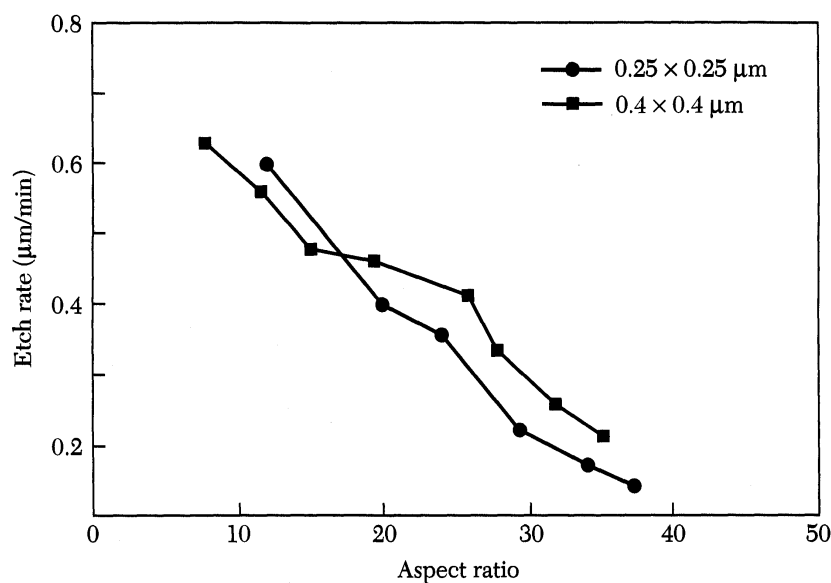


Fig. 31 Dependence of average silicon trench etch rate on aspect ratio.²

Dielectric Etching

The patterning of dielectrics, especially silicon dioxide and silicon nitride, is a key process in the manufacture of modern semiconductor devices. Because of their higher bonding energies, dielectric etching requires aggressive ion-enhanced, fluorine-based plasma chemistry. Vertical profiles are achieved by sidewall passivation, typically by introducing a carbon-containing fluorine species to the plasma (e.g., CF_4 , CHF_3 , C_4F_8). High ion-bombardment energies are required to remove this polymer layer from the oxide, as well as to mix the reactive species into the oxide surface to form SiF_x products.

A low-pressure, high-density plasma is advantageous for aspect-ratio-dependent etching. However, the HDP generates high-temperature electrons and subsequently generates a high degree of dissociation of ions and radicals. It generates far more active radicals and ions than RIE or MERIE plasmas. In particular, a high F concentration worsens the selectivity to silicon. Various methods were tried to enhance the selectivities in the high-density plasma. First, a parent gas with a high C/F ratio, such as C_2F_6 , C_4F_8 , or C_5F_8 , has been tried. Also other methods to scavenge F radicals have been developed.²⁵

Interconnect Metal Etching

Etching of a metallization layer is a very important step in IC fabrication. Aluminum, copper, and tungsten are the most popular materials used for interconnection. These materials usually require anisotropic etching. The reaction of aluminum with fluorine results in nonvolatile AlF_3 , which has a vapor pressure of only 1 Torr at 1240°C . Chlorine-based (e.g., Cl_2/BCl_3 mixture) chemistry has been widely used for aluminum etching. Chlorine has a very high chemical etch rate with aluminum and tends to produce an undercut during etching. Carbon-containing gas (e.g., CHF_3) or N_2 is added to form sidewall passivation during aluminum etching to obtain anisotropic etching.

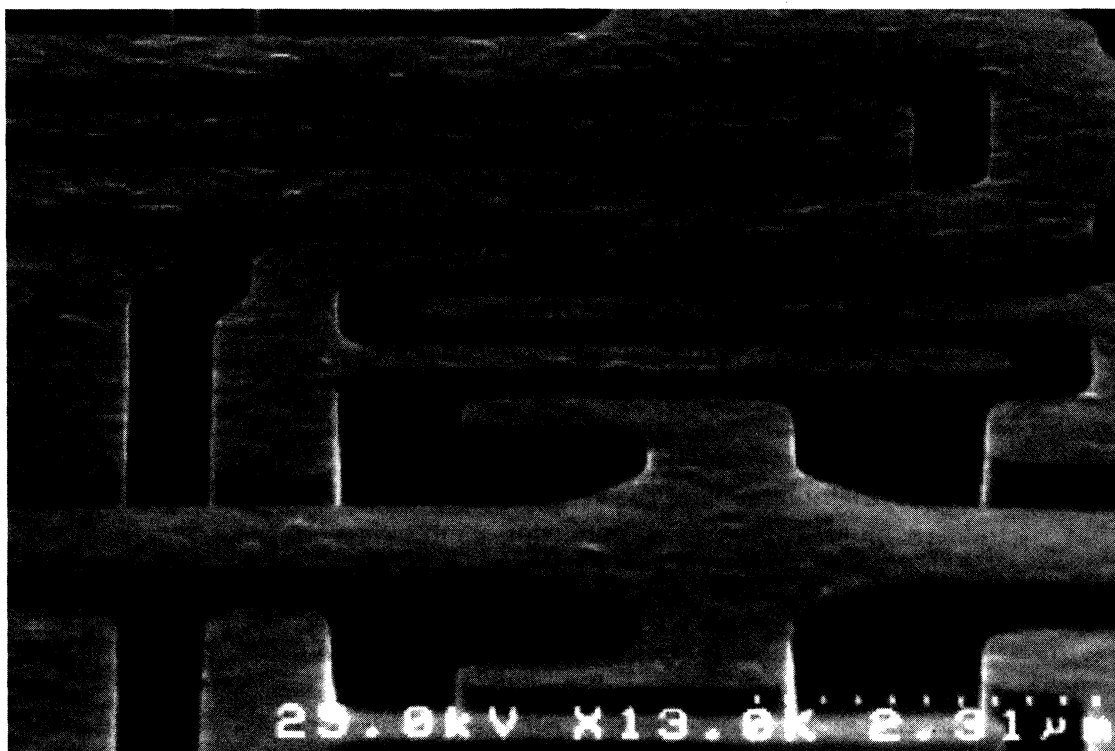


Fig. 32 0.35 μm TiN/Al/TiN lines and spaces on a wafer maintained at ambient for 72 hours after a microwave strip are not corroded.

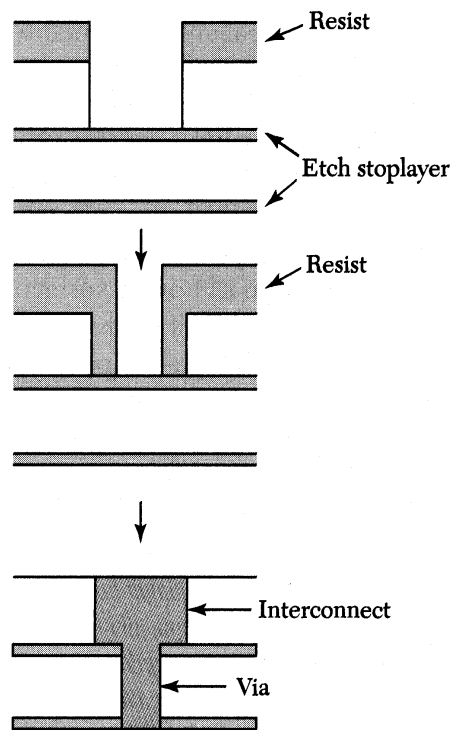


Fig. 33 Various process sequences for dual damascene process.

Exposure to the ambient is another problem in aluminum etching. Residue chlorine on the Al sidewall and the photoresist tends to react with atmosphere water to form HCl, which corrodes aluminum. An in-situ exposure of the wafer to a CF_4 discharge to exchange Cl with F and then to an oxygen discharge to remove the resist followed by immediate immersion in deionized water can eliminate Al corrosion. Figure 32 shows $0.35\text{ }\mu\text{m}$ TiN/Al/Ti lines and spaces on a wafer maintained at ambient for 72 hours. No corrosion is present even after prolonged exposure to the ambient.

Copper has drawn much attention as a new metallization material in ULSI circuits because of its low resistivity ($\sim 1.7\text{ }\mu\text{ohm-cm}$) and superior resistance to electromigration compared with Al or Al alloys. However, because of the low volatility of copper halides, plasma etching at room temperature is difficult. Process temperatures higher than 200°C are required to etch copper films. Therefore, the damascene process is used to form Cu interconnection without dry etching. Damascene processing involves the creation of interconnect lines by first etching a trench or canal in a planar dielectric layer and then filling that trench with metal, such as aluminum or copper. In dual damascene processing (Fig. 33), a second level is involved where a series of holes (i.e., contacts or vias) are etched and filled in addition to the trench. After filling, the metal and dielectric are planarized by chemical-mechanical processing (CMP). The advantage of damascene processing is that it eliminates the need for metal etch. This is an important concern as the industry moves from aluminum to copper interconnections.

Low-pressure CVD (LPCVD) tungsten (W) has been widely used for filling contact holes and first-level metallization because of its excellent deposition conformability. Both fluorine- and chlorine-based chemistries etch W and form volatile etch products. An important tungsten etch process is the blanket W etchback to form a W plug. The blanket LPCVD W is deposited on top of a TiN barrier layer, as shown in Fig. 34. A two-step process is usually used. First, 90% of the W is etched at a high etch rate, and then the

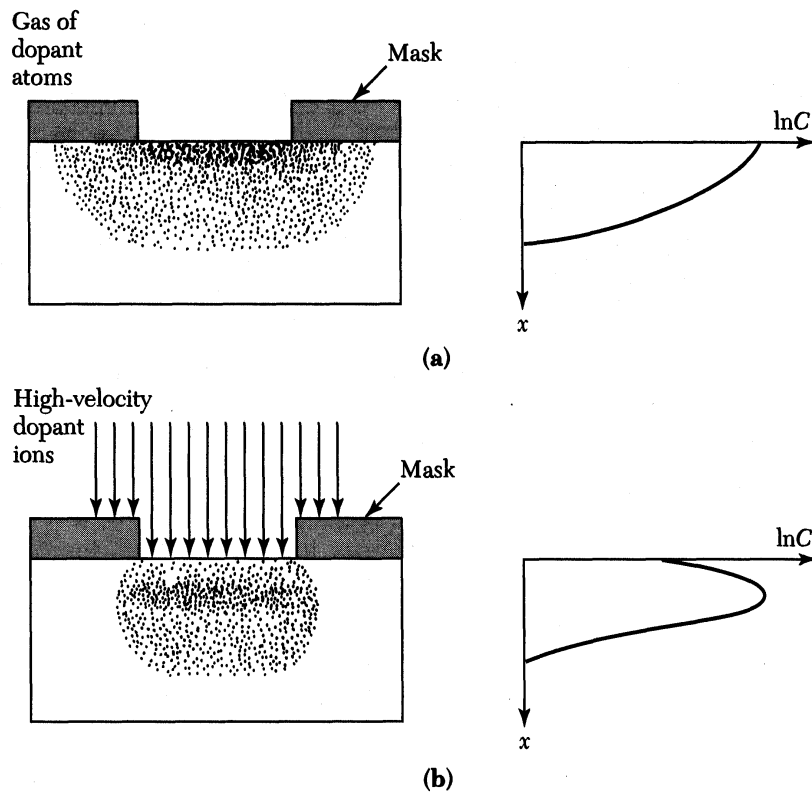


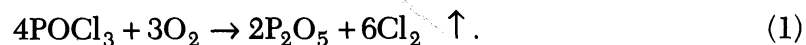
Fig. 1 Comparison of (a) diffusion and (b) ion-implantation techniques for the selective introduction of dopants into the semiconductor substrate.

- Implantation-related processes such as masking, high-energy implantation, and high-current implantation.

13.1 BASIC DIFFUSION PROCESS

Diffusion of impurities is typically done by placing semiconductor wafers in a carefully controlled high-temperature quartz-tube furnace and passing a gas mixture that contains the desired dopant through it. The temperature usually ranges between 800° and 1200°C for silicon and 600° and 1000°C for gallium arsenide. The number of dopant atoms that diffuse into the semiconductor is related to the partial pressure of the dopant impurity in the gas mixture.

For diffusion in silicon, boron is the most popular dopant for introducing a *p*-type impurity, whereas arsenic and phosphorus are used extensively as *n*-type dopants. These three elements are highly soluble in silicon, as they have solubilities above $5 \times 10^{20} \text{ cm}^{-3}$ in the diffusion temperature range. These dopants can be introduced in several ways, including solid sources (e.g., BN for boron, As_2O_3 for arsenic, and P_2O_5 for phosphorus), liquid sources (BBr_3 , AsCl_3 , and POCl_3), and gaseous sources (B_2H_6 , AsH_3 , and PH_3). However, liquid sources are most commonly used. A schematic diagram of the furnace and gas flow arrangement for a liquid source is shown in Fig. 2. This arrangement is similar to that used for thermal oxidation. An example of the chemical reaction for phosphorus diffusion using a liquid source is



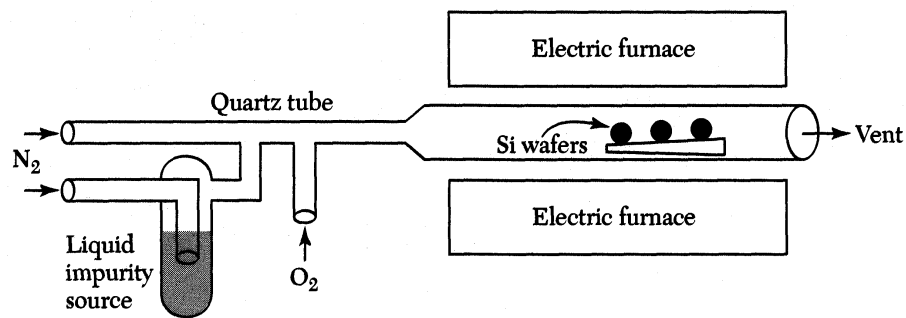
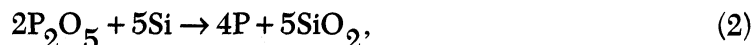


Fig. 2 The schematic diagram of a typical open-tube diffusion system.

The P_2O_5 forms a glass on silicon wafer and is then reduced to phosphorus by silicon,



and the phosphorus is released and diffuses into the silicon and Cl_2 is vented.

For diffusion in gallium arsenide, the high vapor pressure of arsenic requires special methods to prevent the loss of arsenic by decomposition or evaporation.² These methods include diffusion in sealed ampules with an overpressure of arsenic and diffusion in an open-tube furnace with a doped-oxide capping layer (e.g., silicon nitride). Most of the studies on *p*-type diffusion have been confined to the use of zinc in the forms of Zn-Ga-As alloys and $ZnAs_2$ for the sealed-ampule approach or $ZnO-SiO_2$ for the open-tube approach. The *n*-type dopants in gallium arsenide include selenium and tellurium.

13.1.1 Diffusion Equation

Diffusion in a semiconductor can be visualized as atomic movement of the diffusant (dopant atoms) in the crystal lattice by vacancies or interstitials. Figure 3 shows the two basic atomic diffusion models in a solid.^{1,3} The open circles represent the host atoms occupying the equilibrium lattice positions. The solid dots represent impurity atoms. At elevated temperatures, the lattice atoms vibrate around the equilibrium lattice sites. There is a finite probability that a host atom acquires sufficient energy to leave the lattice site and to become an interstitial atom, thereby creating a vacancy. When a neighboring impurity atom migrates

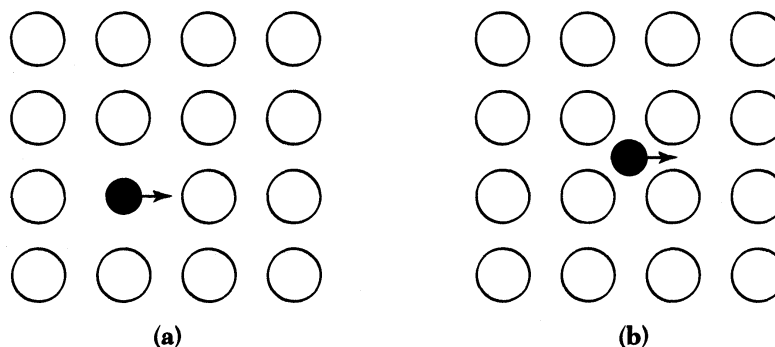


Fig. 3 Atomic diffusion mechanisms for a two-dimensional lattice.^{1,3} (a) Vacancy mechanism; (b) interstitial mechanism.

to the vacancy site, as illustrated in Fig. 3a, the mechanism is called *vacancy diffusion*. If an interstitial atom moves from one place to another without occupying a lattice site (Fig. 3b), the mechanism is *interstitial diffusion*. An atom smaller than the host atom often moves interstitially.

The basic diffusion process of impurity atoms is similar to that of charge carriers (electrons and holes), discussed in Chapter 3. Accordingly, we define a flux F as the number of dopant atoms passing through a unit area in a unit time and C as the dopant concentration per unit volume. From Eq. 27 in Chapter 3, we have

$$F = -D \frac{\partial C}{\partial x}, \quad (3)$$

where we have substituted C for the carrier concentration and the proportionality constant D is the diffusion coefficient or diffusivity. Note that the basic driving force of the diffusion process is the concentration gradient dC/dx . The flux is proportional to the concentration gradient, and the dopant atoms will move (diffuse) away from a high-concentration region toward a lower-concentration region.

If we substitute Eq. 3 into the one-dimensional continuity equation, Eq. 56, given in Chapter 3 under the condition that no materials are formed or consumed in the host semiconductor (i.e., $G_n = R_n = 0$), we obtain

$$\frac{\partial C}{\partial t} = -\frac{\partial F}{\partial x} = \frac{\partial}{\partial x} \left(D \frac{\partial C}{\partial x} \right). \quad (4)$$

When the concentration of the dopant atoms is low, the diffusion coefficient can be considered to be independent of doping concentration, and Eq. 4 becomes

$$\boxed{\frac{\partial C}{\partial t} = D \frac{\partial^2 C}{\partial x^2}}. \quad (5)$$

Equation 5 is often referred to as *Fick's diffusion equation*.

Figure 4 shows the measured diffusion coefficients for low concentrations of various dopant impurities in silicon and gallium arsenide.^{4,5} The logarithm of the diffusion coefficient plotted against the reciprocal of the absolute temperature gives a straight line in most of the cases. This implies that over the temperature range, the diffusion coefficients can be expressed as

$$\boxed{D = D_0 \exp\left(\frac{-E_a}{kT}\right)}, \quad (6)$$

where D_0 is the diffusion coefficient in cm^2/s extrapolated to infinite temperature and E_a is the activation energy in eV.

For the interstitial diffusion model, E_a is related to the energies required to move dopant atoms from one interstitial site to another. The values of E_a are found to be between 0.5 and 2 eV in both silicon and gallium arsenide. For the vacancy diffusion model, E_a is related to both the energies of motion and the energies of formation of vacancies. Thus,

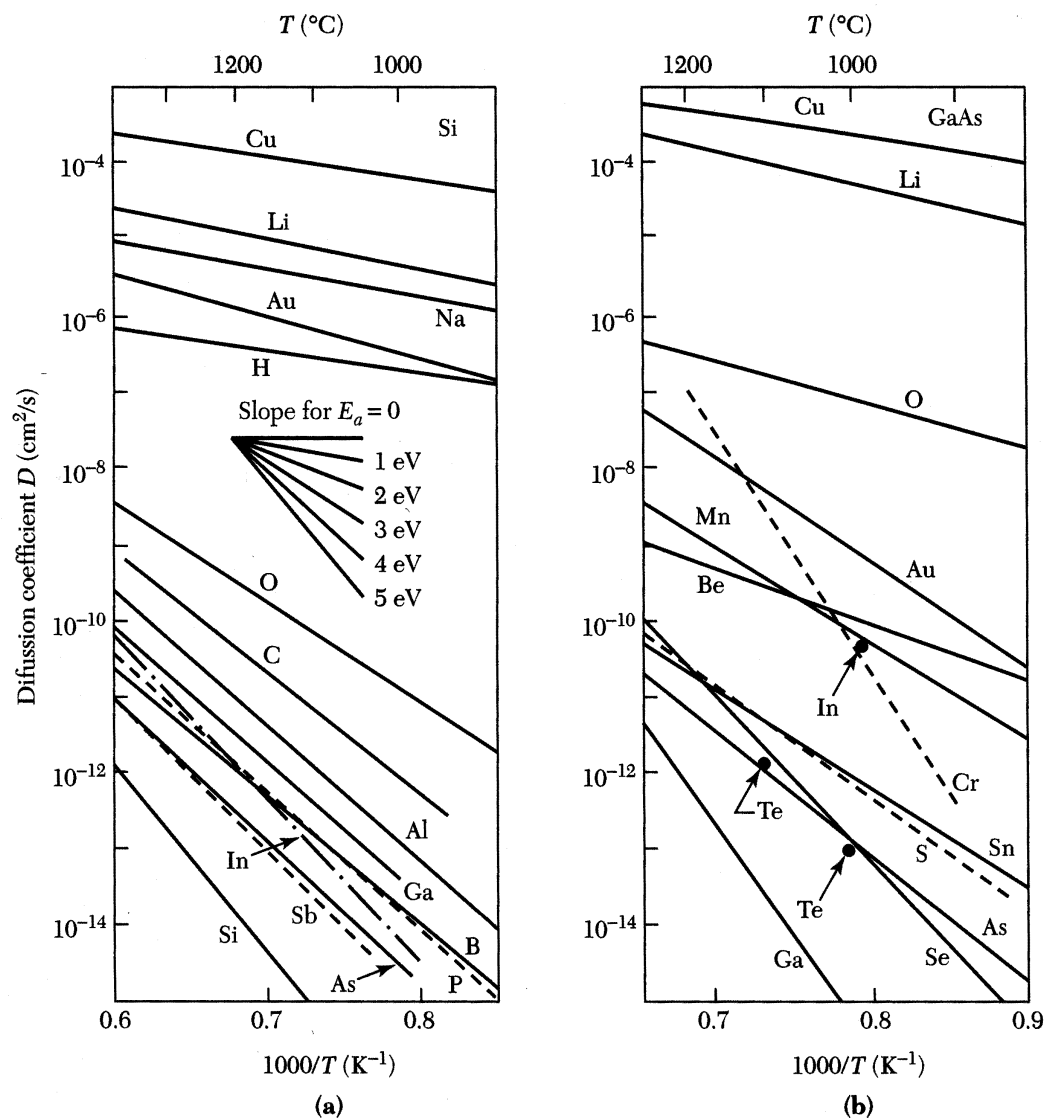


Fig. 4 Diffusion coefficient (also called diffusivity) as a function of the reciprocal of temperature for (a) silicon and (b) gallium arsenide.^{4,5}

E_a for vacancy diffusion is larger than that for interstitial diffusion, usually between 3 and 5 eV.

For fast diffusants, such as Cu in Si and GaAs, shown in the upper portion of Fig. 4a and 4b, the measured activation energies are less than 2 eV, and interstitial atomic movement is the dominant diffusion mechanism. For slow diffusants, such as As in Si and GaAs, shown in the lower portion of Fig. 4a and 4b, E_a is larger than 3 eV, and vacancy diffusion is the dominant mechanism.

13.1.2 Diffusion Profiles

The diffusion profile of the dopant atoms is dependent on the initial and boundary conditions. In this subsection we consider two important cases, namely, constant-surface-concentration diffusion and constant-total-dopant diffusion. In the first case, impurity atoms are transported from a vapor source onto the semiconductor surface and diffused into the semiconductor wafers. The vapor source maintains a constant level of surface

concentration during the entire diffusion period. In the second case, a fixed amount of dopant is deposited onto the semiconductor surface and is subsequently diffused into the wafers.

Constant-Surface-Concentration Diffusion

The initial condition at $t = 0$ is

$$C(x, 0) = 0, \quad (7)$$

which states that the dopant concentration in the host semiconductor is initially zero. The boundary conditions are

$$C(0, t) = C_s \quad (8a)$$

and

$$C(\infty, t) = 0 \quad (8b)$$

where C_s is the surface concentration (at $x = 0$), which is independent of time. The second boundary condition states that at large distances from the surface there are no impurity atoms.

The solution of the diffusion equation (Eq. 5) that satisfies the initial and boundary conditions is given by ⁶

$$C(x, t) = C_s \operatorname{erfc} \left(\frac{x}{2\sqrt{Dt}} \right), \quad (9)$$

where erfc is the complementary error function and $\sqrt{Dt}$ is the diffusion length. The definition of erfc and some properties of the function are summarized in Table 1. The diffusion profile for the constant-surface-concentration condition is shown in Fig. 5a, where we plot, on both linear (upper) and logarithmic (lower) scales, the normalized concentration as a function of depth for three values of the diffusion length $\sqrt{Dt}$ corresponding to three consecutive diffusion times and a fixed D for a given diffusion temperature. Note that as the time progresses, the dopant penetrates deeper into the semiconductor.

The total number of dopant atoms per unit area of the semiconductor is given by

$$Q(t) = \int_0^\infty C(x, t) dx. \quad (10)$$

Substituting Eq. 9 into Eq. 10 yields

$$Q(t) = \frac{2}{\sqrt{\pi}} C_s \sqrt{Dt} \cong 1.13 C_s \sqrt{Dt}. \quad (11)$$

This expression can be interpreted as follows. The quantity $Q(t)$ represents the area under one of the diffusion profiles of the linear plot in Fig. 5a. These profiles can be approximated by triangles with height C_s and base $2\sqrt{Dt}$. This leads to $Q(t) \cong C_s \sqrt{Dt}$, which is close to the exact result obtained from Eq. 11.

A related quantity is the gradient of the diffusion profile dC/dx . The gradient can be obtained by differentiating Eq. 9:

$$\left. \frac{dC}{dx} \right|_{x,t} = -\frac{C_s}{\sqrt{\pi Dt}} e^{-x^2/4Dt}. \quad (12)$$

TABLE 1 Error Function Algebra

$$\begin{aligned}
 \operatorname{erf}(x) &\equiv \frac{2}{\sqrt{\pi}} \int_0^x e^{-y^2} dy \\
 \operatorname{erfc}(x) &\equiv 1 - \operatorname{erf}(x) \\
 \operatorname{erf}(0) &= 0 \\
 \operatorname{erf}(\infty) &= 1 \\
 \operatorname{erf}(x) &\approx \frac{2}{\sqrt{\pi}} x \quad \text{for } x \ll 1 \\
 \operatorname{erfc}(x) &\approx \frac{1}{\sqrt{\pi}} \frac{e^{-x^2}}{x} \quad \text{for } x \gg 1. \\
 \frac{d}{dx} \operatorname{erf}(x) &= \frac{2}{\sqrt{\pi}} e^{-x^2} \\
 \frac{d^2}{dx^2} \operatorname{erf}(x) &= -\frac{4}{\sqrt{\pi}} x e^{-x^2} \\
 \int_0^x \operatorname{erfc}(y') dy' &= x \operatorname{erfc}(x) + \frac{1}{\sqrt{\pi}} (1 - e^{-x^2}) \\
 \int_0^\infty \operatorname{erfc}(x) dx &= \frac{1}{\sqrt{\pi}}
 \end{aligned}$$

► EXAMPLE 1

For a boron diffusion in silicon at 1000°C, the surface concentration is maintained at 10^{19} cm^{-3} and the diffusion time is 1 hour. Find $Q(t)$ and the gradient at $x = 0$ and at a location where the dopant concentration reaches 10^{15} cm^{-3} .

SOLUTION The diffusion coefficient of boron at 1000°C, as obtained from Fig. 4, is about $2 \times 10^{-14} \text{ cm}^2/\text{s}$, so that the diffusion length is

$$\sqrt{Dt} = \sqrt{2 \times 10^{-14} \times 3600} = 8.48 \times 10^{-6} \text{ cm},$$

$$Q(t) = 1.13 C_s \sqrt{Dt} = 1.13 \times 10^{19} \times 8.48 \times 10^{-6} = 9.5 \times 10^{13} \text{ atoms / cm}^2.$$

$$\left. \frac{dC}{dx} \right|_{x=0} = -\frac{C_s}{\sqrt{\pi Dt}} = \frac{-10^{19}}{\sqrt{\pi} \times 8.48 \times 10^{-6}} = -6.7 \times 10^{23} \text{ cm}^{-4}.$$

When $C = 10^{15} \text{ cm}^{-3}$, the corresponding distance x_j is given by Eq. 9, or

$$x_j = 2\sqrt{Dt} \operatorname{erfc}^{-1}\left(\frac{10^{15}}{10^{19}}\right) = 2\sqrt{Dt} (2.75) = 4.66 \times 10^{-5} \text{ cm} = 0.466 \text{ } \mu\text{m},$$

$$\left. \frac{dC}{dx} \right|_{x=0.466 \text{ } \mu\text{m}} = -\frac{C_s}{\sqrt{\pi Dt}} e^{-x_j^2/4Dt} = -3.5 \times 10^{20} \text{ cm}^{-4}.$$

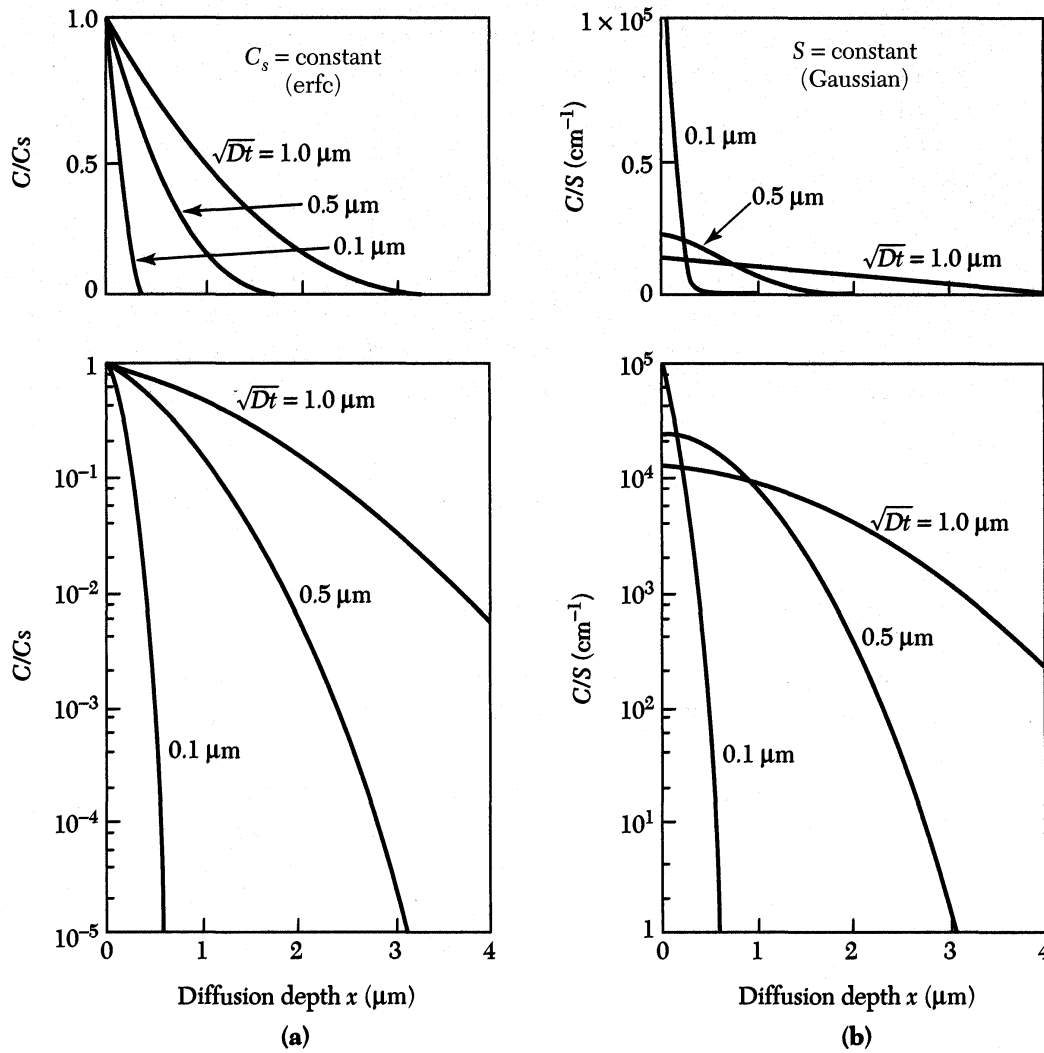


Fig. 5 Diffusion profiles. (a) Normalized complementary error function versus distance for successive diffusion times. (b) Normalized Gaussian function versus distance.

Constant-Total-Dopant Diffusion

For this case, a fixed (or constant) amount of dopant is deposited onto the semiconductor surface in a thin layer, and the dopant subsequently diffuses into the semiconductor. The initial condition is the same as in Eq. 7. The boundary conditions are

$$\int_0^{\infty} C(x, t) dx = S \quad (13a)$$

and

$$C(\infty, t) = 0 \quad (13b)$$

where S is the total amount of dopant per unit area.

The solution of the diffusion equation, Eq. 5, that satisfies the above conditions is

$$C(x, t) = \frac{S}{\sqrt{\pi Dt}} \exp\left(-\frac{x^2}{4Dt}\right). \quad (14)$$

This expression is the Gaussian distribution. Since the dopant will move into the semiconductor as time increases, to keep the total dopant S constant, the surface concentration must decrease. This is indeed the case, since the surface concentration is given by Eq. 14 with $x = 0$:

$$C_s(t) = \frac{S}{\sqrt{\pi Dt}}. \quad (15)$$

Figure 5b shows the dopant profile for a Gaussian distribution where we plot the normalized concentration (C/S) as a function of the distance for three increasing diffusion lengths. Note the reduction of the surface concentration as the diffusion time increases. The gradient of the diffusion profile is obtained by differentiating Eq. 14 and is

$$\left. \frac{dC}{dx} \right|_{x,t} = -\frac{xS}{2\sqrt{\pi(Dt)^{3/2}}} e^{-x^2/4Dt} = -\frac{x}{2Dt} C(x, t). \quad (16)$$

The gradient (or slope) is zero at $x = 0$ and at $x = \infty$ and the maximum gradient occurs at $x = \sqrt{2Dt}$.

In integrated-circuit processing, a two-step diffusion process is commonly used, in which a *predeposition* diffused layer is first formed under a constant-surface-concentration condition. This step is followed by a *drive-in* diffusion (also called *redistribution* diffusion) under a constant-total-dopant condition. For most practical cases, the diffusion length $\sqrt{Dt}$ for the predeposition diffusion is much smaller than the diffusion length for the drive-in diffusion. Therefore, we can consider the predeposition profile as a delta function at the surface, and we can regard the extent of the penetration of the predeposition profile to be negligibly small compared with that of the final profile that results from the drive-in step.

► EXAMPLE 2

Arsenic was predeposited by arsine gas and the resulting total amount of dopant per unit area is 1×10^{14} atoms/cm². How long would it take to drive the arsenic in to a junction depth of 1 μm ? Assume a background doping of 1×10^{15} atoms/cm³, and a drive-in temperature of 1200°C. For As diffusion, $D_0 = 24$ cm²/s, and $E_a = 4.08$ eV.

SOLUTION

$$D = D_0 \exp\left(\frac{-E_a}{kT}\right) = 24 \exp\left(\frac{-4.08}{8.614 \times 10^{-5} \times 1473}\right) = 2.602 \times 10^{-13} \text{ cm}^2/\text{s},$$

$$x_j^2 = 10^{-8} = 4Dt \ln\left(\frac{S}{C_B \sqrt{\pi Dt}}\right) = 1.04 \times 10^{-12} t \ln\left(\frac{1.106 \times 10^5}{\sqrt{t}}\right),$$

$$t \cdot \log t - 10.09t + 8350 = 0.$$

The solution to the above equation can be determined by the cross point of equation $y = t \cdot \log t$ and $y = 10.09t - 8350$.

Therefore, $t = 1190$ seconds $\cong 20$ minutes. ◀

13.1.3 Evaluation of Diffused Layers

The results of a diffusion process can be evaluated by three measurements—the junction depth, the sheet resistance, and the dopant profile of the diffused layer. The junction depth can be delineated by cutting a groove into the semiconductor and etching the surface with a solution (e.g., 100 cm³ HF and a few drops of HNO₃ for silicon) that stains the *p*-type region darker than the *n*-type region, as illustrated in Fig. 6a. If R_0 is the radius of the tool used to form the groove, then the junction depth x_j is given by

$$x_j = \sqrt{R_0^2 - b^2} - \sqrt{R_0^2 - a^2}, \quad (17)$$

where a and b are indicated in the figure. In addition, if R_0 is much larger than a and b , then

$$x_j \cong \frac{a^2 - b^2}{2R_0}. \quad (18)$$

The junction depth x_j as illustrated in Fig. 6b is the position where the dopant concentration equals the substrate concentration C_B , or

$$C(x_j) = C_B. \quad (19)$$

Thus, if the junction depth and C_B are known, the surface concentration C_s and the impurity distribution can be calculated, provided the diffusion profile follows one or the other simple equation derived in Section 13.1.2.

The resistance of a diffused layer can be measured by the four-point probe technique described in Chapter 3. The *sheet resistance* R is related to the junction depth x_j , the carrier mobility μ (which is a function of the total impurity concentration), and the impurity distribution $C(x)$ by the following expression:⁷

$$R = \frac{1}{q \int_0^{x_j} \mu C(x) dx}. \quad (20)$$

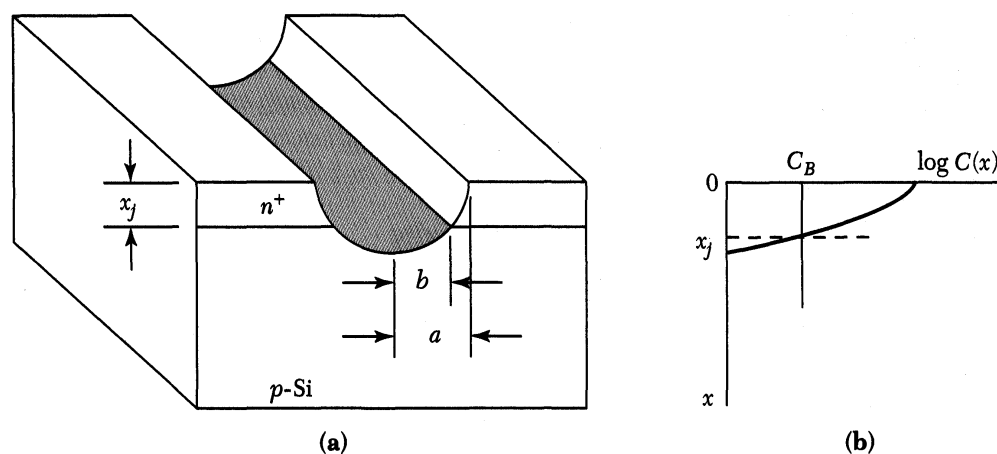


Fig. 6 Junction-depth measurement. (a) Grooving and staining; (b) position in which dopant and substrate concentration are equal.

For a given diffusion profile, the average resistivity $\bar{\rho} = R_x$ is uniquely related to the surface concentration C_s and the substrate-doping concentration for an assumed diffusion profile. Design curves relation C_s and $\bar{\rho}$ have been calculated for simple diffusion profiles, such as the erfc or Gaussian distribution.⁸ To use these curves correctly we must be sure that the diffusion profiles agree with the assumed profiles. For low concentration and deep diffusions, the diffusion profiles generally can be represented by the aforementioned simple functions. However, as we discuss in the next section, for high concentration and shallow diffusions, the diffusion profiles cannot be represented by these simple functions.

The diffusion profile can be measured using the capacitance-voltage technique described in Chapter 7. The majority carrier profile, which is equal to the impurity profile if impurities are fully ionized, can be determined by measuring the reverse-bias capacitance of a p - n junction or a Schottky barrier diode as a function of the applied voltage. A more elaborate method is the secondary-ion-mass spectroscope (SIMS) technique, which measures the total impurity profile. In the SIMS technique, an ion beam sputters material off the surface of a semiconductor, and the ion component is detected and mass analyzed. This technique has high sensitivity to many elements, such as boron and arsenic, and it is an ideal tool for providing the precision needed for profile measurements in high-concentration or shallow-junction diffusions.⁹

► 13.2 EXTRINSIC DIFFUSION

The diffusion profiles described in Section 13.1 are for constant diffusivities. These profiles occur when the doping concentration is lower than the intrinsic-carrier concentration $n_i(T)$ at the diffusion temperature. For example, at $T = 1000^\circ\text{C}$, n_i equals $5 \times 10^{18} \text{ cm}^{-3}$ for silicon and $5 \times 10^{17} \text{ cm}^{-3}$ for gallium arsenide. The diffusivity at low concentrations is often referred to as the intrinsic diffusivity. Doping profiles that have concentrations less than $n_i(T)$ are in the *intrinsic* diffusion region as indicated in the left side of Fig. 7. In this region, the resulting dopant profiles of sequential or simultaneous diffusions of n - and p -type impurities can be determined by superposition, that is, the diffusions can be treated independently. However, when the impurity concentration, including both the substrate and the dopant, is greater than $n_i(T)$, the semiconductor becomes extrinsic and the diffusivity is considered to be extrinsic. In the extrinsic diffusion region the diffusivity becomes concentration dependent.¹⁰ In the extrinsic diffusion region the diffusion profiles are more complicated, and there are interactions and cooperative effects among the sequential or simultaneous diffusions.

13.2.1 Concentration-Dependent Diffusivity

As mentioned previously, when a host atom acquires sufficient energy from the lattice vibration to leave its lattice site, a vacancy is created. Depending on the charges associated with a vacancy, we can have a neutral vacancy V^0 , an acceptor vacancy V^- , a double-charged acceptor vacancy V^{2-} , a donor vacancy V^+ , and so forth. We expect that the vacancy density of a given charge state (i.e., the number of vacancies per unit volume, C_V) has a temperature dependence similar to that of the carrier density (see to Eq. 28 in Chapter 2), that is ,

$$C_V = C_i \exp\left(\frac{E_F - E_i}{kT}\right), \quad (21)$$

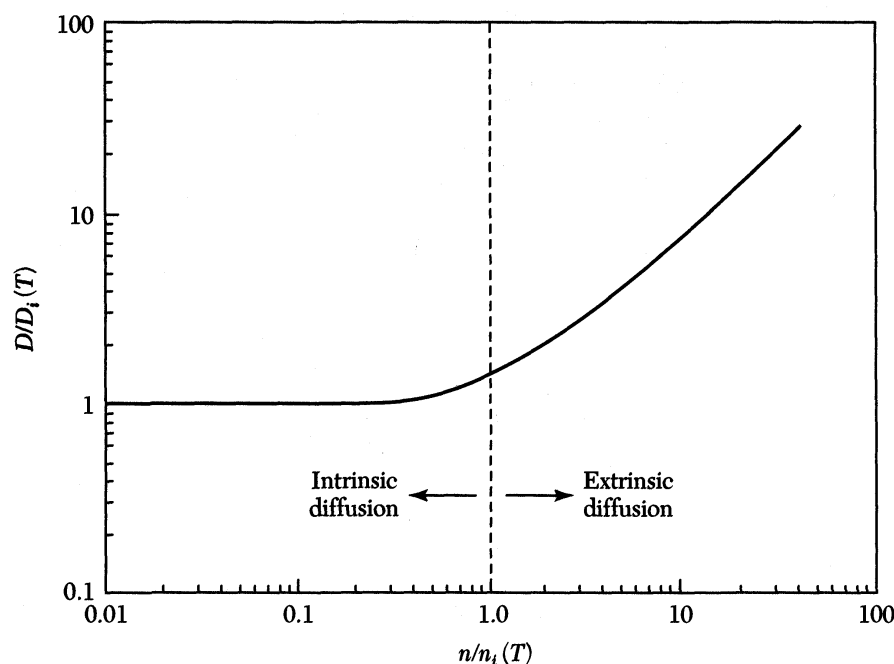


Fig. 7 Donor-impurity diffusivity versus electron concentration showing regions of intrinsic and extrinsic diffusion.¹⁰

where C_i is the intrinsic vacancy density, E_F is the Fermi level, and E_i is the intrinsic Fermi level.

If the dopant diffusion is dominated by the vacancy mechanism, the diffusion coefficient is expected to be proportional to the vacancy density. At low doping concentrations ($n < n_i$), the Fermi level coincides with the intrinsic Fermi level ($E_F = E_i$). The vacancy density is equal to C_i and is independent of doping concentration. The diffusion coefficient, which is proportional to C_v , also is independent of doping concentration. At high concentrations ($n > n_i$), the Fermi level will move toward the conduction band edge (for donor-type vacancies), and the term $[\exp (E_F - E_i)/kT]$ becomes larger than unity. This causes C_v to increase, which in turn causes the diffusion coefficient to increase, as shown in the right side of Fig. 7.

When the diffusion coefficient varies with dopant concentration, Eq. 4 should be used as the diffusion equation instead of Eq. 5, in which D is independent of C . We consider the case where the diffusion coefficient can be written as

$$D = D_s \left(\frac{C}{C_s} \right)^\gamma, \quad (22)$$

where C_s is the surface concentration, D_s is the diffusion coefficient at the surface, and γ is a parameter to describe the concentration dependence. For such a case, we can write the diffusion equation, Eq. 4, as an ordinary differential equation and solve it numerically.

Figure 8 shows the solutions¹¹ for a constant-surface-concentration diffusion with different values of γ . For $\gamma = 0$, we have the case of constant diffusivity and the profile is the same as that shown in Fig. 5a. For $\gamma > 0$ the diffusivity decreases as the concentration decreases, and increasingly steep and box-like concentration profiles result for increasing γ . Therefore, highly abrupt junctions are formed when diffusions are made into a background of an opposite impurity type. The abruptness of the doping profile

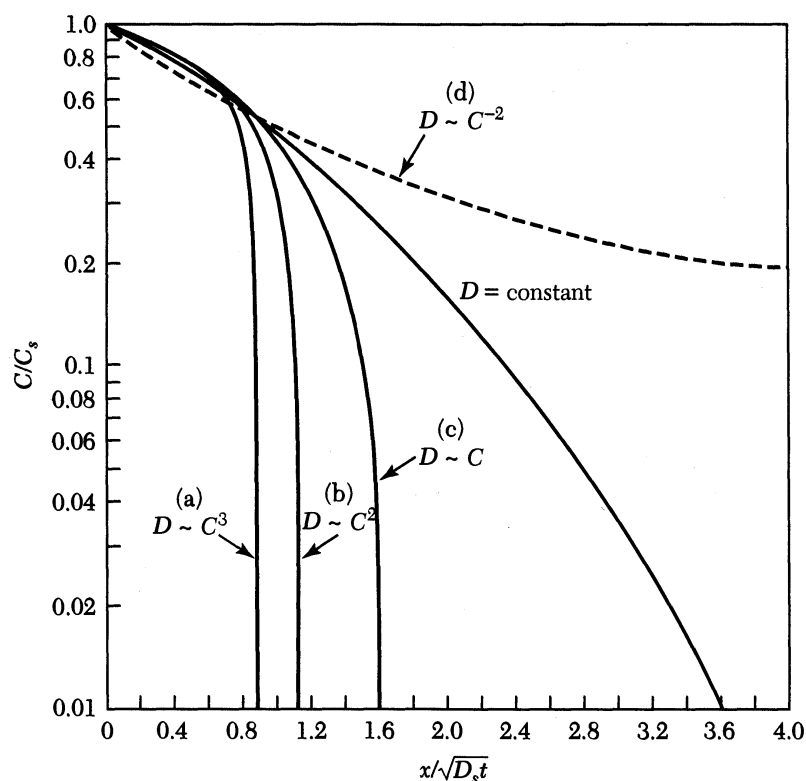


Fig. 8 Normalized diffusion profiles for extrinsic diffusion where the diffusion coefficient becomes concentration dependent.^{10,11}

results in a junction depth virtually independent of the background concentration. Note that the junction depth (see Fig. 8) is given by

$$\begin{aligned}
 x_j &= 1.6\sqrt{D_s t} \quad \text{for } D \sim C (\gamma = 1), \\
 x_j &= 1.1\sqrt{D_s t} \quad \text{for } D \sim C^2 (\gamma = 2), \\
 x_j &= 0.87\sqrt{D_s t} \quad \text{for } D \sim C^3 (\gamma = 3).
 \end{aligned}
 \tag{23}$$

In the case of $\gamma = -2$, the diffusivity increases with decreasing concentration, which leads to a concave profile, as opposed to the convex profiles for other cases.

13.2.2 Diffusion Profiles

Diffusion in Silicon

The measured diffusion coefficients of boron and arsenic in silicon have a concentration dependence with $\gamma \cong 1$. Their concentration profiles are abrupt, as depicted in curve *c* of Fig. 8. For gold and platinum diffusion in silicon, γ is close to -2 , and their concentration profiles have the concave shape shown in curve *d* of Fig. 8.

The diffusion of phosphorus in silicon is associated with the doubly charged acceptor vacancy V^{2-} , and the diffusion coefficient at high concentration varies as C^2 . We would expect that the diffusion profile of phosphorus resembles that shown in curve *b* of Fig. 8. However, because of a *dissociation effect*, the diffusion profile exhibits anomalous behavior.

Figure 9 shows phosphorus diffusion profiles for various surface concentrations after diffusion into silicon for 1 hour at 1000°C .¹² When the surface concentration is low, cor-

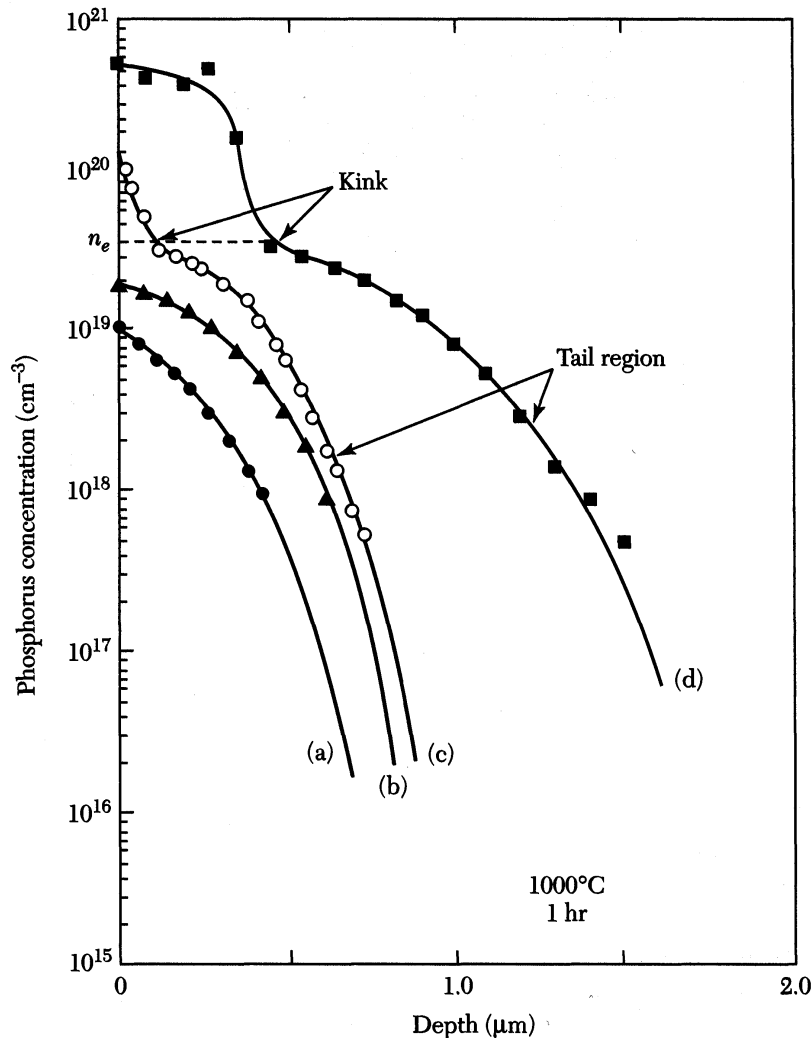


Fig. 9 Phosphorus diffusion profiles¹² for various surface concentrations after diffusion into silicon for 1 hour at 1000°C.

responding to the intrinsic diffusion region, the diffusion profile is given by an erfc (curve *a*). As the concentration increases, the profile begins to deviate from the simple expression (curves *b* and *c*). At very high concentration (curve *d*), the profile near the surface is indeed similar to that shown in curve *b* of Fig. 8. However, at concentration n_e , a kink occurs and is followed by a rapid diffusion in the tail region. The concentration n_e corresponds to a Fermi level 0.11 eV below the conduction band. At this energy level, the coupled impurity-vacancy pair (P^+V^{2-}) dissociates to P^+ , V^- , and an electron. Thus, the dissociation generates a large number of singly charged acceptor vacancies V^- , which in turn enhances the diffusion in the tail region of the profile. The diffusivity in the tail region is over 10^{-12} cm²/s, which is about two orders of magnitude larger than the intrinsic diffusivity at 1000°C. Because of its high diffusivity, phosphorus is commonly used to form deep junctions, such as the *n*-tubs in CMOS.

Zinc Diffusion in Gallium Arsenide

We expect diffusion in gallium arsenide to be more complicated than that in silicon because the diffusion of impurities may involve atomic movements on both the gallium and arsenic sublattices. Vacancies play a dominant role in diffusion processes in gallium arsenide

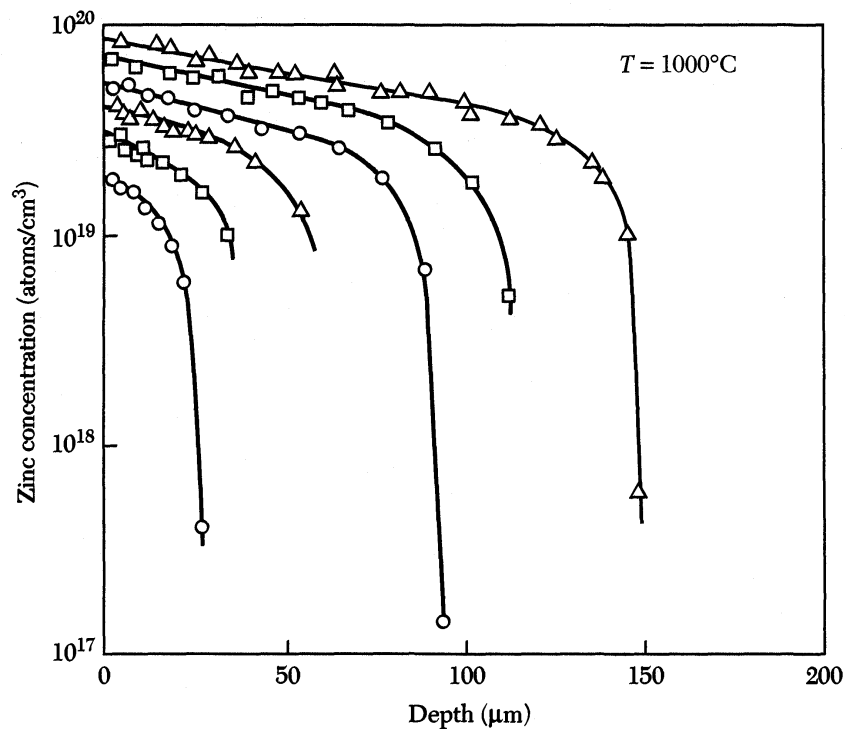


Fig. 10 Diffusion profiles¹³ of zinc in GaAs after annealing at 1000°C for 2.7 hours. The different surface concentrations are obtained by maintaining the Zn source at temperatures in the range 600°–800°C.

because both *p*- and *n*-type impurities must ultimately reside in lattice sites. However, the charge states of the vacancies have not been established.

Zinc is the most extensively studied diffusant in gallium arsenide. Its diffusion coefficient is found to vary as C^2 . Therefore, the diffusion profiles are steep, as shown¹³ in Fig. 10, and resemble curve *b* of Fig. 8. Note that even for the case of the lowest surface concentration, the diffusion is in the extrinsic-diffusion region, because n_i for GaAs at 1000°C is less than 10^{18} cm^{-3} . As can be seen in Fig. 10, the surface concentration has a profound effect on the junction depth. The diffusivity varies linearly with the partial pressure of the zinc vapor, and the surface concentration is proportional to the square root of the partial pressure. Therefore, from Eq. 23, the junction depth is linearly proportional to the surface concentration.

► 13.3 DIFFUSION-RELATED PROCESSES

In this section we consider two processes in which diffusion plays an important role and the impact of these processes on device performances.

13.3.1 Lateral Diffusion

The one-dimensional diffusion equation discussed previously can describe satisfactorily the diffusion process, except at the edge of the mask window. Here the impurities will diffuse downward and sideways (i.e., laterally). In this case, we must consider a two-dimensional diffusion equation and use a numerical technique to obtain the diffusion profiles under different initial and boundary conditions.

Figure 11 shows the contours of constant doping concentration for a constant-surface-concentration diffusion condition assuming that the diffusivity is independent of concentration.¹⁴ At the far right of the figure, the variation of the dopant concentration from $0.5 C_s$ to $10^{-4} C_s$ (where C_s is the surface concentration) corresponds to the erfc distribution given by Eq. 9. The contours are in effect a map of the location of the junctions created by diffusing into various background concentrations. For example, at $C/C_s = 10^{-4}$ (i.e., the background doping is 10^4 times lower than the surface concentration), we see from this constant-concentration curve that the vertical penetration is about $2.8 \mu\text{m}$, whereas the lateral penetration is about $2.3 \mu\text{m}$ (i.e., the penetration along the diffusion mask-semiconductor interface). Therefore, the lateral penetration is about 80% of the penetration in the vertical direction for concentrations three or more orders of magnitude below the surface concentration. Similar results are obtained for a constant-total-dopant diffusion condition. The ratio of lateral to vertical penetration is about 75%. For concentration-dependent diffusivities, the ratio is found to be reduced slightly, to about 65%–70%.

Because of the lateral-diffusion effect, the junction consists of a central plane (or flat) region with approximately cylindrical edges with a radius of curvature r_j , as shown in Fig. 11. In addition, if the diffusion mask contains sharp corners, the shape of the junction near the corner will be roughly spherical because of lateral diffusion. Since the electric-field intensities are higher for cylindrical and spherical junction regions, the avalanche breakdown voltages of such regions can be substantially lower than that of a plane junction having the same background doping. This junction “curvature effect” was discussed in Chapter 4.

13.3.2 Impurity Redistribution During Oxidation

Dopant impurities near the silicon surface will be redistributed during thermal oxidation. The redistribution depends on several factors. When two solid phases are brought together, an impurity in one solid will redistribute between the two solids until it reaches equilibrium. This is similar to our previous discussion on impurity redistribution in crystal growth

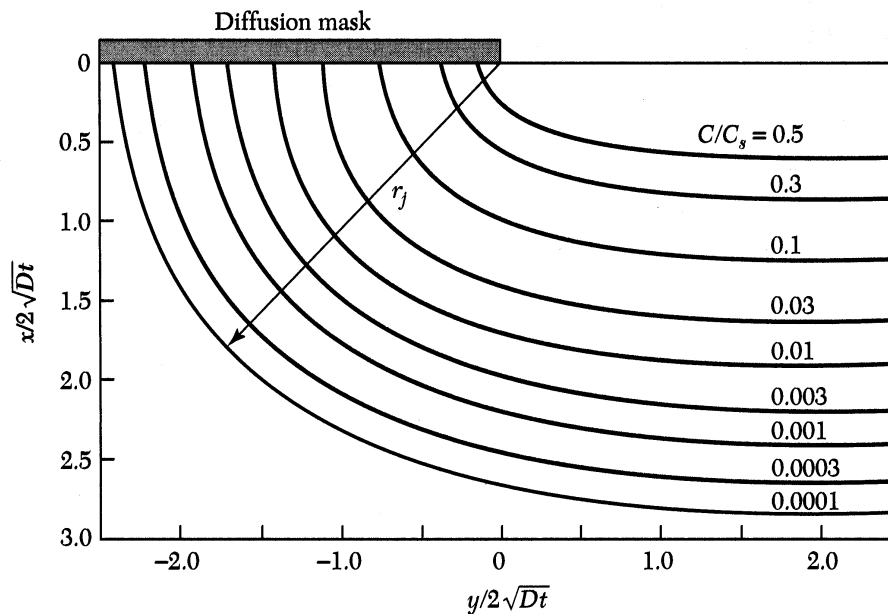


Fig. 11 Diffusion contours at the edge of an oxide window, where r_j is the radius of curvature.¹⁴

from the melt. The ratio of the equilibrium concentration of the impurity in the silicon to that in the silicon dioxide is called the *segregation coefficient* and is defined as

$$k = \frac{\text{equilibrium concentration of impurity in silicon}}{\text{equilibrium concentration of impurity in SiO}_2} \quad (24)$$

A second factor that influences impurity distribution is that the impurity may diffuse rapidly through the silicon dioxide and escape to the gaseous ambient. If the diffusivity of the impurity in silicon dioxide is large, this factor will be important. A third factor in the redistribution process is that the oxide is growing, and thus, the boundary between the silicon and the oxide is advancing into the silicon as a function of time. The relative rate of this advance compared with the diffusion rate of the impurity through the oxide is important in determining the extent of the redistribution. Note that even if the segregation coefficient of an impurity k equals unity, some redistribution of the impurity in the silicon will still take place. As indicated in Fig. 3 of Chapter 11, the oxide layer will be about twice as thick as the silicon layer it replaced. Therefore, the same amount of impurity will now be distributed in a larger volume, resulting in a depletion of the impurity from the silicon.

Four possible redistribution processes are illustrated⁶ in Fig. 12. These processes can be classified into two groups. In one group the oxide takes up the impurity (Fig. 12a and

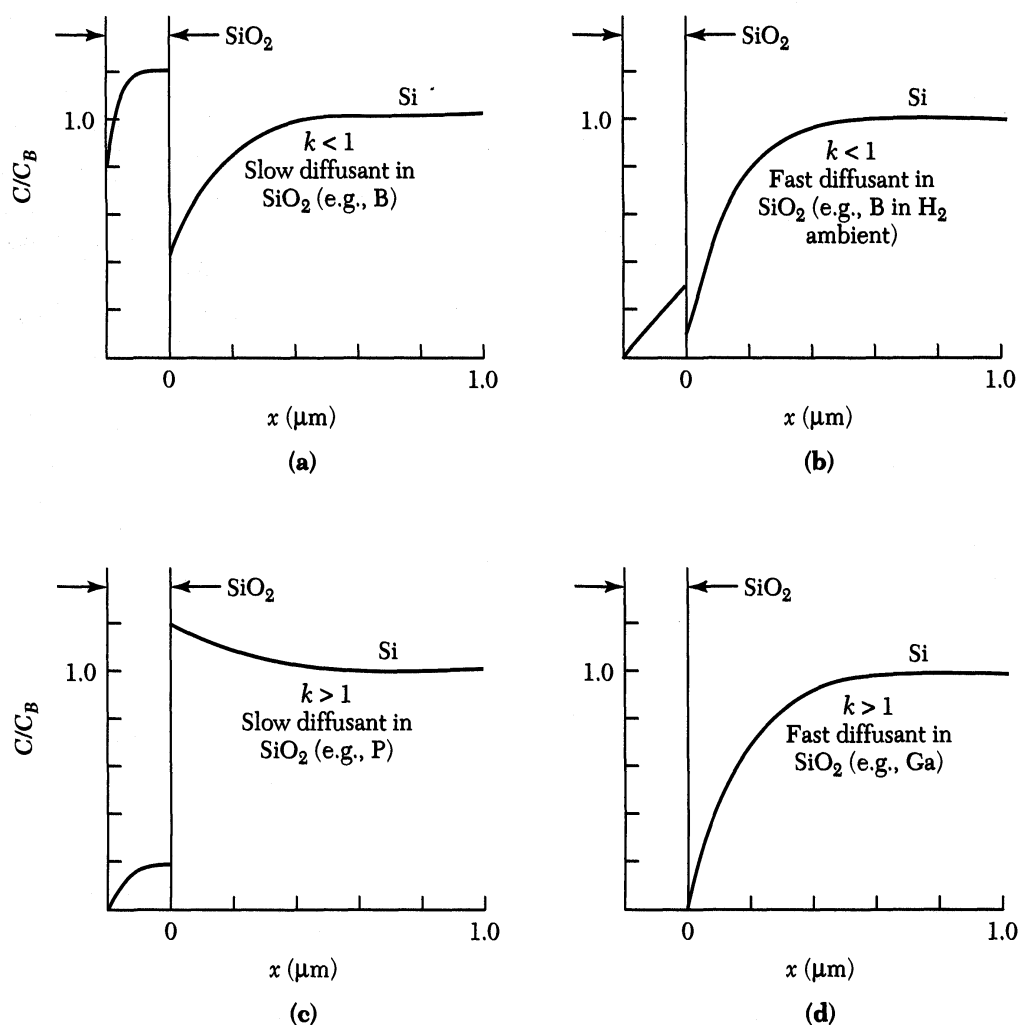


Fig. 12 Four different cases of impurity redistribution in silicon due to thermal oxidation.⁶

b for $k < 1$), and in the other the oxide rejects the impurity (Fig. 12c and d for $k > 1$). In each case, what happens depends on how rapidly the impurity can diffuse through the oxide. In group 1, the silicon surface is depleted of impurities; an example is boron with k approximately equal to 0.3. Rapid diffusion of the impurity through the silicon dioxide increases the amount of depletion; an example is boron-doped silicon heated in a hydrogen ambient because hydrogen in silicon dioxide enhances the diffusivity of boron. In group 2, k is greater than unity, so that the oxide rejects the impurity. If diffusion of the impurity through the silicon dioxide is relatively slow, the impurity piles up near the silicon surface; an example is phosphorus, with k approximately equal to 10. When diffusion through the silicon dioxide is rapid, so much impurity may escape from the solid to the gaseous ambient that the overall effect will be a depletion of the impurity; an example is gallium, with k approximately equal to 20.

The redistributed dopant impurities in silicon dioxide are seldom electrically active. However, redistribution in silicon has an important effect on processing and device performance. For example, nonuniform dopant distribution will modify the interpretation of the measurements of interface-trap properties (see Chapter 6), and the change of the surface concentration will modify the threshold voltage and device contact resistance (see Chapter 7).

13.4 RANGE OF IMPLANTED IONS

Ion implantation is the introduction of energetic, charged particles into a substrate such as silicon. Implantation energies are between 1 keV and 1 MeV, resulting in ion distributions with average depths ranging from 10 nm to 10 μm . Ion doses vary from 10^{12} ions/cm² for threshold voltage adjustment to 10^{18} ions/cm² for the formation of buried insulating layer. Note that the dose is expressed as the number of ions implanted into 1 cm² of the semiconductor surface area. The main advantages of ion implantation are its more precise control and reproducibility of impurity dopings and its lower processing temperature compared with those of the diffusion process.

Figure 13 shows schematically a medium-energy ion implanter.¹⁵ The ion source has a heated filament to break up source gas such as BF_3 or AsH_3 into charged ions (B^+ or As^+). An extraction voltage, around 40 kV, causes the charged ions to move out of the ion-source chamber into a mass analyzer. The magnetic field of the analyzer is chosen such that only ions with the desired mass-to-charge ratio can travel through it without being filtered. The selected ions then enter the acceleration tube, where they are accelerated to the implantation energy as they move from high voltage to ground. Apertures

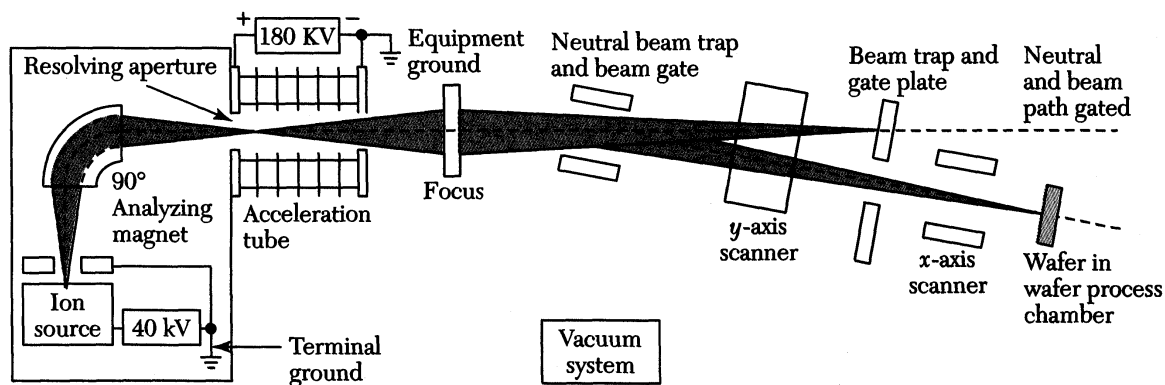


Fig. 13 Schematic of a medium-current ion implanter.

ensure that the ion beam is well collimated. The pressure in the implantor is kept below 10^{-4} Pa to minimize ion scattering by gas molecules. The ion beam is then scanned over the wafer surface using electrostatic deflection plates and is implanted into the semiconductor substrate.

The energetic ions lose their energies through collision with electrons and nuclei in the substrate and finally come to rest at some depth within the lattice. The average depth can be controlled by adjusting the acceleration energy. The dopant dose can be controlled by monitoring the ion current during implantation. The principle side effect is the disruption or damage of the semiconductor lattice due to ion collisions. Therefore, a subsequent annealing treatment is needed to remove these damages.

13.4.1 Ion Distribution

The total distance that an ion travels in coming to rest is called its *range* R and is illustrated¹⁶ in Fig. 14a. The projection of this distance along the axis of incidence is called the *projected range* R_p . Because the number of collisions per unit distance and the energy lost per collision are random variables, there will be a spatial distribution of ions having

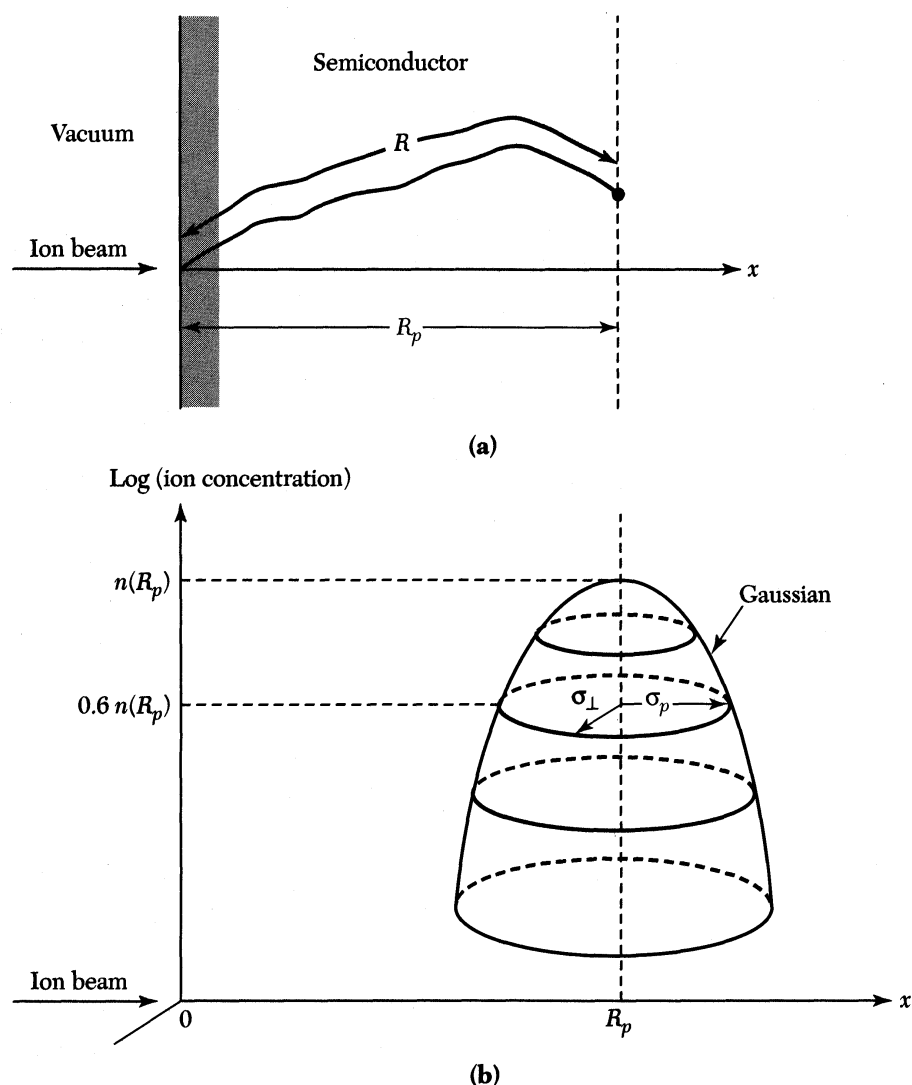


Fig. 14 (a) Schematic of the ion range R and projected range R_p . (b) Two-dimensional distribution of the implanted ions.¹⁶

the same mass and the same initial energy. The statistical fluctuations in the projected range are called the *projected straggle* σ_p . There is also a statistical fluctuation along an axis perpendicular to the axis of incidence, called the *lateral straggle* $\sigma_\perp$.

Figure 14b shows the ion distribution. Along the axis of incidence, the implanted impurity profile can be approximated by a Gaussian distribution function:

$$n(x) = \frac{S}{\sqrt{2\pi}\sigma_p} \exp\left[-\frac{(x - R_p)^2}{2\sigma_p^2}\right], \quad (25)$$

where S is the ion dose per unit area. This equation is similar to Eq. 14 for constant-total-dopant diffusion, except that the quantity $4Dt$ is replaced by $2\sigma_p^2$ and the distribution is shifted along the x -axis by R_p . Thus, for diffusion, the maximum concentration is at $x = 0$, whereas for ion implantation the maximum concentration is at the projected range R_p . The ion concentration is reduced by 40% from its peak value at $(x - R_p) = \pm\sigma_p$, by one decade at $\pm 2\sigma_p$, by two decades at $\pm 3\sigma_p$, and by five decades at $\pm 4.8\sigma_p$.

Along the axis perpendicular to the axis of incidence, the distribution is also a Gaussian function of the form $\exp(-y^2/2\sigma_\perp^2)$. Because of this distribution, there will be some lateral implantation.¹⁷ However, the lateral penetration from the mask edge (on the order of $\sigma_\perp$) is considerably smaller than that from the thermal diffusion process discussed in Section 13.3.

13.4.2 Ion Stopping

There are two stopping mechanisms by which an energetic ion, on entering a semiconductor substrate (also called the target), can be brought to rest. The first is by transferring its energy to the target nuclei. This causes deflection of the incident ion and also dislodges many target nuclei from their original lattice sites. If E is the energy of the ion at any point x along its path, we can define a nuclear stopping power $S_n(E) \equiv (dE/dx)_n$ to characterize this process. The second stopping mechanism is by the interaction of the incident ion with the cloud of electrons surrounding the target's atoms. The ion loses energy in collisions with electrons through Coulombic interaction. The electrons can be excited to higher energy levels (excitation), or they can be ejected from the atom (ionization). We can define an electronic stopping power $S_e(E) \equiv (dE/dx)_e$ to characterize this process.

The average rate of energy loss with distance is given by a superposition of the above two stopping mechanisms:

$$\frac{dE}{dx} = S_n(E) + S_e(E). \quad (26)$$

If the total distance traveled by the ion before coming to rest is R , then

$$R = \int_0^R dx = \int_0^{E_0} \frac{dE}{S_n(E) + S_e(E)}, \quad (27)$$

where E_0 is the initial ion energy. The quantity R has been defined previously as the range.

We can visualize the nuclear stopping process by considering the elastic collision between an incoming hard sphere (energy E_0 and mass M_1) and a target hard sphere (initial energy zero and mass M_2), as illustrated in Fig. 15. When the spheres collide, momentum is transferred along the centers of the spheres. The deflection angle θ and the velocities v_1 and v_2 can be obtained from the requirements for conservation of momentum and

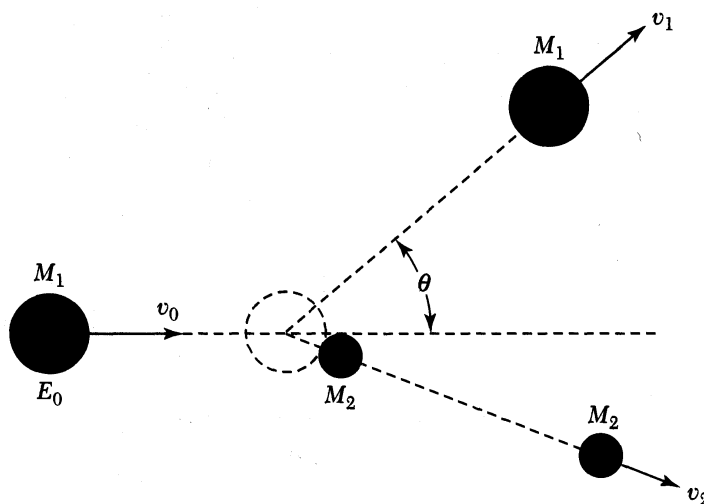


Fig. 15 Collision of hard spheres.

energy. The maximum energy loss is in a head-on collision. For this case, the energy loss by the incident particle M_1 or the energy transferred to M_2 is

$$\frac{1}{2} M_2 v_2^2 = \left[\frac{4 M_1 M_2}{(M_1 + M_2)^2} \right] E_0. \quad (28)$$

Since M_2 is usually of the same order of magnitude as M_1 , a large amount of energy can be transferred in nuclear stopping process.

Detailed calculations show that the nuclear stopping power increases linearly with energy at low energies (similar to Eq. 28), and $S_n(E)$ reaches a maximum at some intermediate energy. At high energies, $S_n(E)$ becomes smaller because fast particles may not have sufficient interaction time with the target atoms to achieve effective energy transfer. The calculated values of $S_n(E)$ for arsenic, phosphorus, and boron in silicon at various energies are shown in Fig. 16 (solid lines, where the superscript indicates the atomic weight).¹⁸ Note that heavier atoms, such as arsenic, have larger nuclear stopping power, that is, larger energy loss per unit distance.

The electronic stopping power is found to be proportional to the velocity of the incident ion, or

$$S_e(E) = k_e \sqrt{E} \quad (29)$$

where the coefficient k_e is a relatively weak function of atomic mass and atomic number. The value of k_e is approximately 10^7 (eV)^{1/2}/cm for silicon and 3×10^7 (eV)^{1/2}/cm for gallium arsenide. The electronic stopping power in silicon is plotted in Fig. 16 (dotted line). Also shown in the figure are the crossover energies, at which $S_e(E)$ equals $S_n(E)$. For boron, which has a relatively low ion mass compared with the target silicon atom, the crossover energy is only 10 keV. This means that over most of the implantation energy range of 1 keV to 1 MeV, the main energy loss mechanism is due to electronic stopping. On the other hand, for arsenic with relatively high ion mass, the crossover energy is 700 keV. Thus, nuclear stopping dominates over most of the energy range. For phosphorus, the crossover energy is 130 keV. For an E_0 less than 130 keV, nuclear stopping will dominate; for higher energies, electronic stopping will take over.

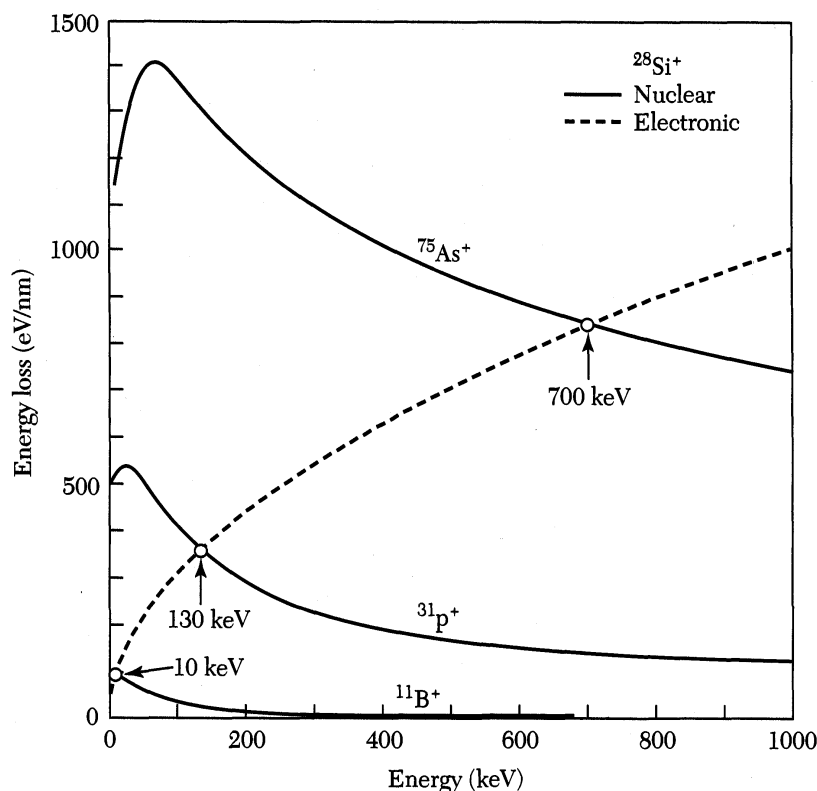


Fig. 16 Nuclear stopping power $S_n(E)$, and electronic stopping power $S_e(E)$ for As, P, and B in Si. The points of intersection of the curves correspond to the energy at which nuclear and electronic stopping are equal.¹⁸

Once $S_n(E)$ and $S_e(E)$ are known, we can calculate the range from Eq. 27. This in turn can give us the projected range and projected straggle with the help of the following approximate equations¹⁵:

$$R_p \cong \frac{R}{1 + (M_2 / 3M_1)}, \quad (30)$$

$$\sigma_p \cong \frac{2}{3} \left[\frac{\sqrt{M_1 M_2}}{M_1 + M_2} \right] R_p. \quad (31)$$

Figure 17a shows the projected range (R_p), the projected straggle (σ_p), and the lateral straggle ($\sigma_\perp$) for arsenic, boron, and phosphorus in silicon.¹⁹ As expected, the larger the energy loss, the smaller the range. Also, the projected range and straggles increase with ion energy. For a given element at a specific incident energy, σ_p and $\sigma_\perp$ are comparable and usually within $\pm 20\%$. Figure 17b shows the corresponding values for hydrogen, zinc, and tellurium in gallium arsenide.¹⁷ If we compare Fig. 17a with Fig. 17b, we see that most of the popular dopants (except hydrogen) have larger projected ranges in silicon than they have in gallium arsenide.

► EXAMPLE 3

Assume 100 keV boron implants on a 200 mm silicon wafer at a dose of 5×10^{14} ions/cm². Calculate the peak concentration and the required ion beam current for 1 minute of implantation.

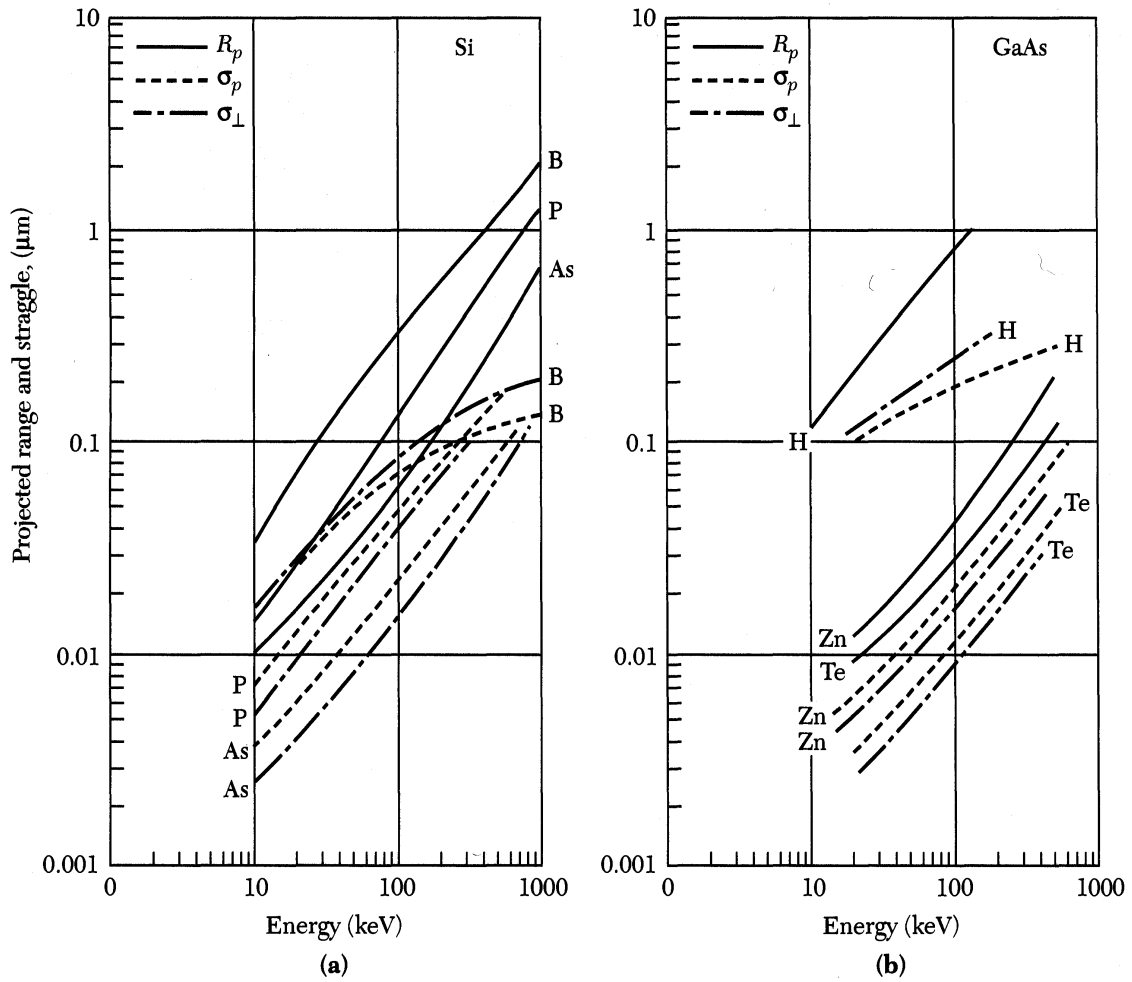


Fig. 17 Projected range, projected straggle, and lateral straggle for (a) B, P, and As in Si, and (b) H, Zn, and Te in GaAs.^{17,19}

SOLUTION From Fig. 17a, we obtain 0.31 and 0.07 μm for the projected range and project straggle, respectively.

$$\text{From Eq. 25, } n(x) = \frac{S}{\sqrt{2\pi}\sigma_p} \exp\left[\frac{-(x - R_p)^2}{2\sigma_p^2}\right],$$

$$\frac{dn}{dx} = -\frac{S}{\sqrt{2\pi}\sigma_p} \frac{2(x - R_p)}{2\sigma_p^2} \exp\left[\frac{-(x - R_p)^2}{2\sigma_p^2}\right] = 0.$$

The peak concentration is at $x = R_p$, and $n(x) = 2.85 \times 10^{19}$ ions/cm³.

The total number of implanted ions = $Q = 5 \times 10^{14} \times \pi \times \left(\frac{20}{2}\right)^2 = 1.57 \times 10^{17}$ ions.

$$\begin{aligned} \text{The required ion current} = I &= \frac{qQ}{t} = \frac{1.6 \times 10^{-19} \times 1.57 \times 10^{17}}{60} = 4.19 \times 10^{-4} \text{ A.} \\ &= 0.42 \text{ mA.} \end{aligned}$$

13.4.3 Ion Channeling

The projected range and straggle of the Gaussian distribution discussed previously give a good description of the implanted ions in amorphous or fine-grain polycrystalline substrates. Both silicon and gallium arsenide behave as if they were amorphous semiconductors, provided the ion beam is misoriented from the low-index crystallographic direction (e.g., $\langle 111 \rangle$). In this situation, the doping profile described by Eq. 25 is followed closely near the peak and extended to one or two decades below the peak value. This is illustrated¹⁶ in Fig. 18. However, even for a misorientation of 7° from the $\langle 111 \rangle$ -axis, there still is a tail that varies exponentially with distance as $\exp(-x/\lambda)$, where λ is typically on the order of $0.1 \mu\text{m}$.

The exponential tail is related to the ion-channeling effect. Channeling occurs when incident ions align with a major crystallographic direction and are guided between rows of atoms in a crystal. Figure 19 illustrates a diamond lattice viewed along a $\langle 110 \rangle$ -direction.²⁰ Ions implanted in the $\langle 110 \rangle$ -direction will follow trajectories that will not bring them close enough to a target atom to lose significant amounts of energy in nuclear collisions. Thus, for channeled ions, the only energy loss mechanism is electronic stopping, and the range of channeled ions can be significantly larger than it would be in an amorphous target. Ion channeling is particularly critical for low-energy implant and heavy ions.

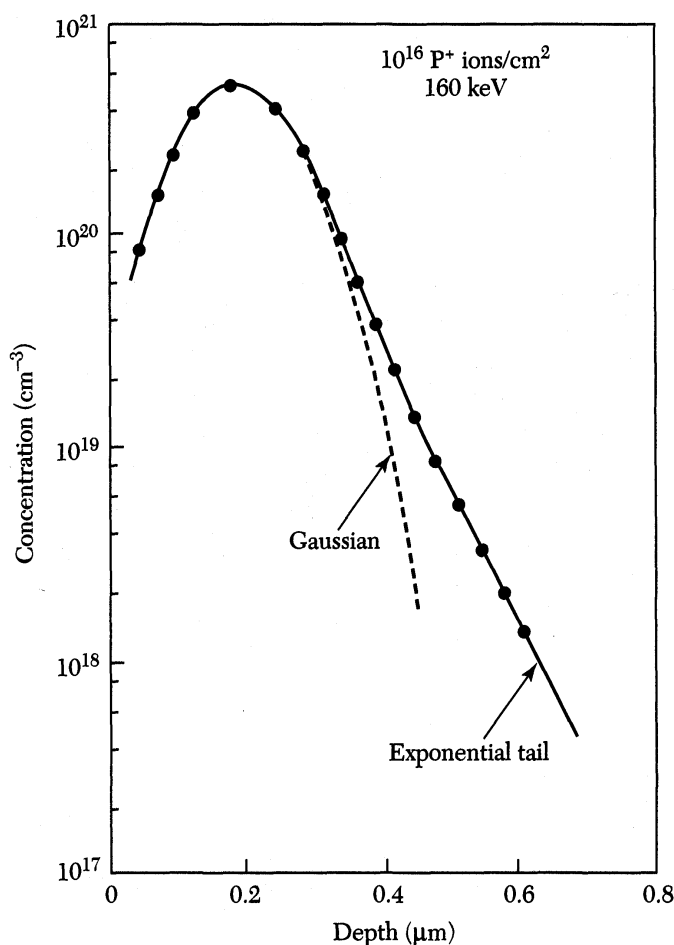


Fig. 18 Impurity profile obtained in a purposely misoriental target. Ion beam is incident 7° from the $\langle 111 \rangle$ -axis.¹⁶

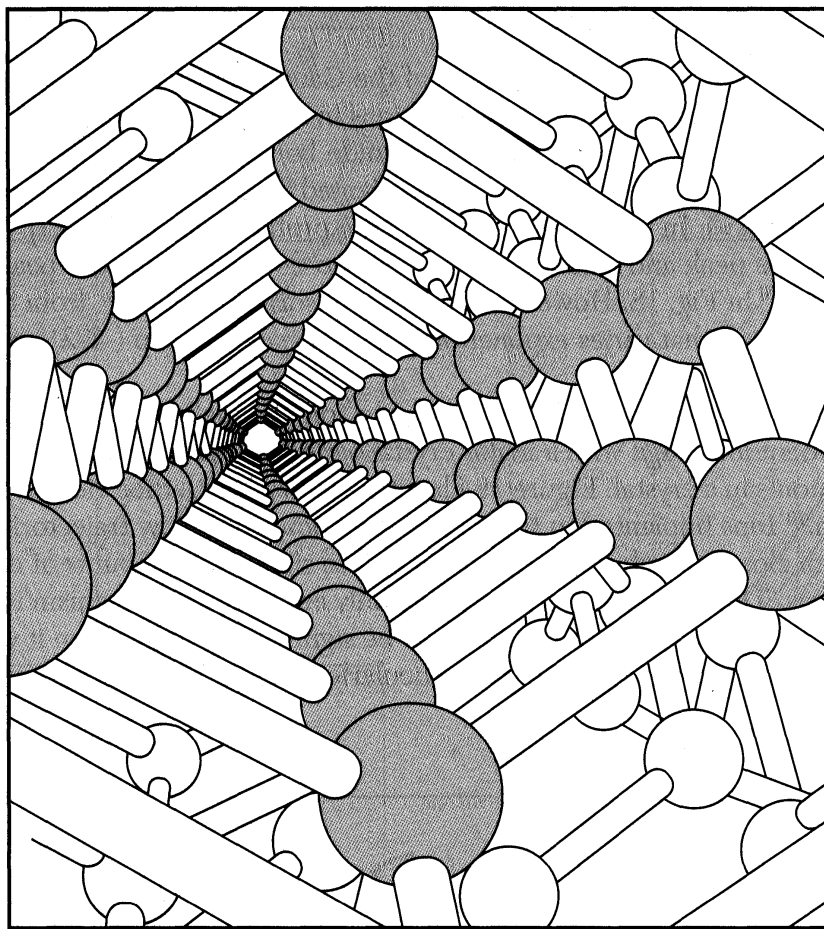


Fig. 19 Model for a diamond structure, viewed along a $\langle 110 \rangle$ -axis.²⁰

Channeling can be minimized by several techniques: a blocking amorphous surface layer, misorientation of the wafer, and creating a damage layer in the wafer surface. The usual blocking amorphous layer is simply a thin layer of grown silicon dioxide (Fig. 20a). The layer randomizes the direction of the ion beam so that the ions enter the wafer at different angles and not directly down the crystal channels. Misorientation of the wafers 5° – 10° off the major plane also has the effect of preventing the ions from entering the channels (Fig. 20b). With this method, most implantation machines tilt the wafer by 7° and then apply a 22° twist from the flat to prevent channeling. Predamaging the wafer surface with a heavy silicon or germanium implant creates a randomizing layer in the wafer surface (Fig. 20c). This method, however, increases the use of the expensive ion implanter.

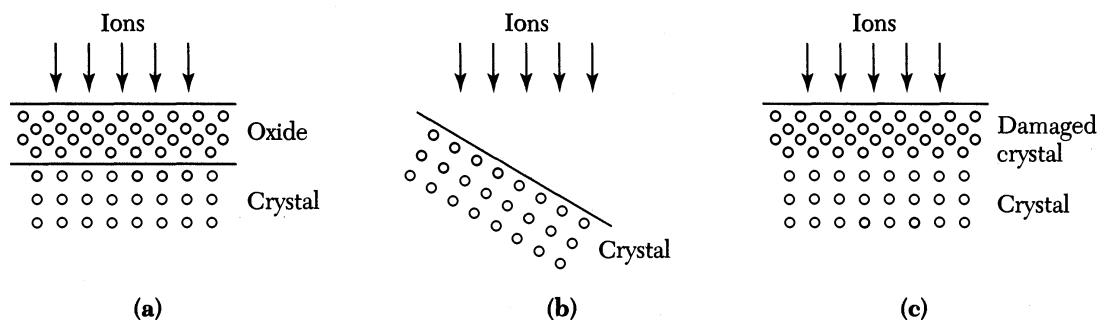


Fig. 20 (a) Implant through an amorphous oxide layer, (b) misorient the beam direction to all crystal axes, and (c) predamage on the crystal surface.

▶ 13.5 IMPLANT DAMAGE AND ANNEALING

13.5.1 Implant Damage

When energetic ions enter a semiconductor substrate, they lose their energy in a series of nuclear and electronic collisions and finally come to rest. The electronic-energy loss can be accounted for in terms of electronic excitations to higher energy levels or in the generation of electron-hole pairs. However, electronic collisions do not displace semiconductor atoms from their lattice positions. Only nuclear collisions can transfer sufficient energy to the lattice so that host atoms are displaced resulting in implant damage (also called lattice disorder).²¹ These displaced atoms may possess large fractions of the incident energy, and they in turn cause cascades of secondary displacements of nearby atoms to form a *tree of disorder* along the ion path. When the displaced atoms per unit volume approach the atomic density of the semiconductor, the material becomes amorphous.

The tree of disorder for light ions is quite different from that for heavy ions. Much of the energy loss for light ions (e.g., $^{11}\text{B}^+$ in silicon) is due to electronic collisions (see Fig. 16), which do not cause lattice damage. The ions lose their energies as they penetrate deeper into the substrate. Eventually, the ion energy is reduced below the crossover energy (10 keV for boron) where nuclear stopping becomes dominant. Therefore, most of the lattice disorder occurs near the final ion position. This is illustrated in Fig. 21a.

We can estimate the damage by considering a 100 keV boron ion. Its projected range is $0.31\text{ }\mu\text{m}$, (Fig. 17a), and its initial nuclear energy loss is only $3\text{ eV/\AA}$ (Fig. 16). Since the spacing between lattice planes in silicon is about $2.5\text{ }\text{\AA}$, this means that the boron ion will lose 7.5 eV for each lattice plane because of nuclear stopping. The energy required to displace a silicon atom from its lattice position is about 15 eV . Therefore, the incident boron ion does not release enough energy from nuclear stopping to displace a silicon atom when it first enters the silicon substrate. When the ion energy is reduced to about

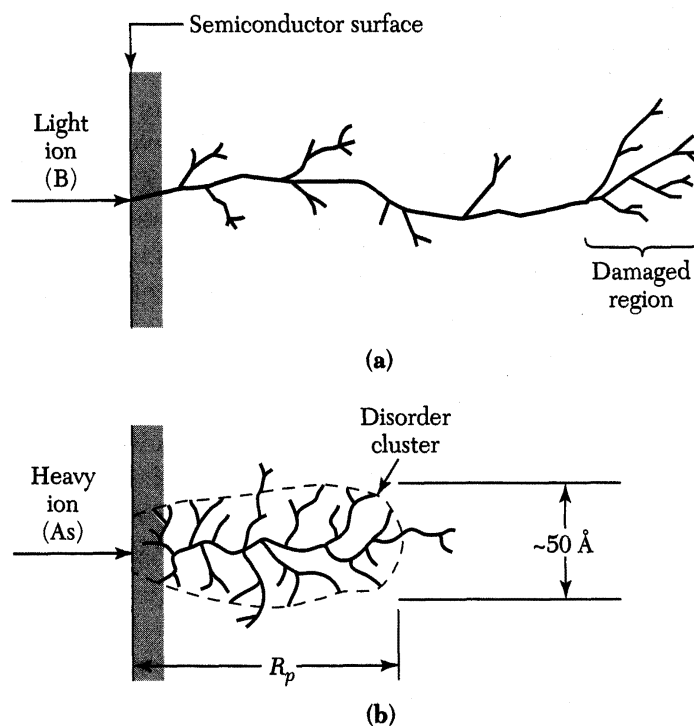


Fig. 21 Implantation disorder caused by (a) light ions and (b) heavy ions.^{2,15}

50 keV (at a depth of 1500 Å), the energy loss due to nuclear stopping increases to 15 eV for each lattice plane (i.e., 6 eV/Å), sufficient to create a lattice disorder. Assuming that one atom is displaced per lattice plane for the remaining ion range, we have 600 lattice atoms displaced (i.e., 1500 Å/2.5 Å). If each displaced atom moves roughly 25 Å from its original position, the damage volume is given by $V_D \cong \pi(25\text{Å})^2(1500\text{Å}) = 3 \times 10^{-18} \text{cm}^3$. The damage density is $600/V_D \cong 2 \times 10^{20} \text{cm}^{-3}$, which is only 0.4% of the atoms. Thus, very high doses of light ions are needed to create an amorphous layer.

For heavy ions, the energy loss is primarily due to nuclear collisions; therefore, we expect substantial damage. Consider a 100 keV arsenic ion with a projected range of 0.06 µm or 60 nm. The average nuclear energy loss over the entire energy range is about 1320 eV/nm (Fig. 16). This means that the arsenic ion loses about 330 eV for each lattice plane on the average. Most of the energy is given to one primary silicon atom. Each primary atom will subsequently cause 22 displaced target atoms (i.e., 330 eV/15 eV). The total number of displaced atom is 5280. Assuming a range of 2.5 nm for the displaced atoms, the damage volume is $V_D \cong \pi(2.5 \text{ nm})^2 (60 \text{ nm}) = 10^{-18} \text{cm}^3$. The damage density is then $5280/V_D \cong 5 \times 10^{21} \text{cm}^{-3}$, or about 10% of the total number of atoms in V_D . As a result of the heavy-ion implantation, the material has become essentially amorphous. Figure 21b illustrates the situation where the damage forms a disordered cluster over the entire projected range.

To estimate the dose required to convert a crystalline material to an amorphous form, we can use the criterion that the energy density is of the same order of magnitude as that needed for melting the material (i.e., 10^{21}keV/cm^3). For 100 keV arsenic ions, the dose required to make amorphous silicon is then

$$S = \frac{(10^{21} \text{ keV / cm}^3) R_p}{E_0} = 6 \times 10^{13} \text{ ions / cm}^2. \quad (32)$$

For 100 keV boron ions, the dose required is $3 \times 10^{14} \text{ ions/cm}^2$ because R_p for boron is five times larger than for arsenic. However, in practice, higher doses ($>10^{16} \text{ ions/cm}^2$) are required for boron implant into a target at room temperature because of the nonuniform distribution of the damage along the ion path.

13.5.2 Annealing

Because of the damaged region and the disorder cluster that result from ion implantation, semiconductor parameters such as mobility and lifetime are severely degraded. In addition, most of the ions as implanted are not located in substitutional sites. To activate the implanted ion and to restore mobility and other material parameters, we must anneal the semiconductor at an appropriate combination of time and temperature.

In conventional annealing, we use an open-tube batch-furnace system similar to that for thermal oxidation. This process requires long time and high temperature to remove the implant damages. However, conventional annealing may cause substantial dopant diffusion and cannot meet the requirement for shallow junctions and narrow doping profiles. Rapid thermal annealing (RTA) is an annealing process that employs a variety of energy sources with a wide range of times, from 100 seconds down to nanoseconds—all short compared with the conventional annealing. RTA can activate the dopant fully with minimal redistribution.

Conventional Annealing of B and P

Annealing characteristics depend on the dopant type and the dose involved. Figure 22 shows the annealing behaviors of boron and phosphorus implantation into silicon substrates.¹⁹ The substrate is held at room temperature (T_s) during implantation. At a given

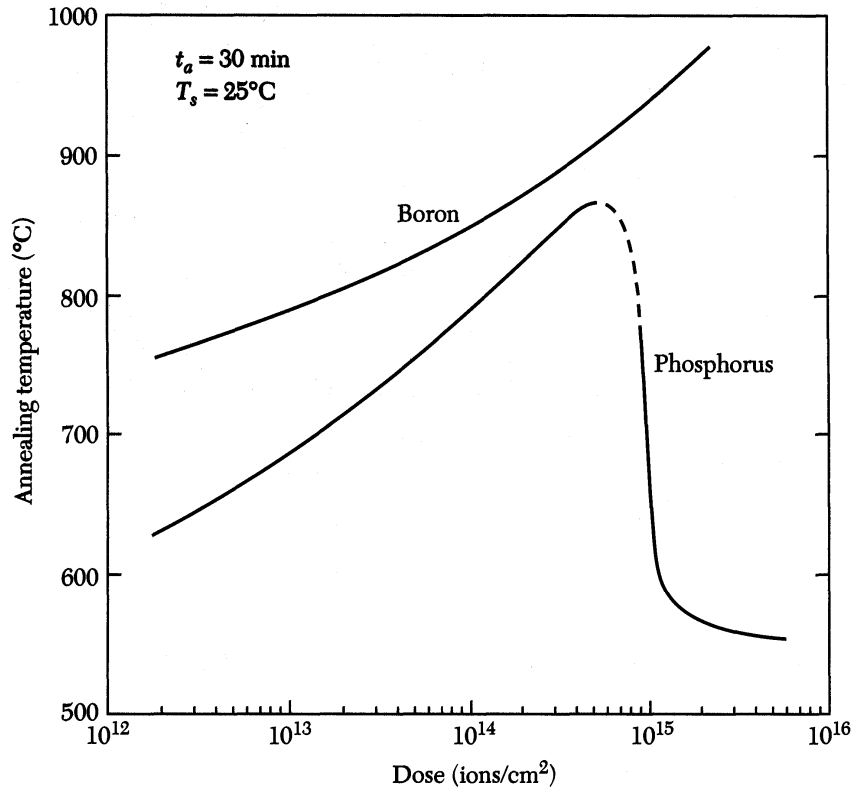


Fig. 22 Annealing temperature versus dose for 90% activation of boron and phosphorus.^{1,19}

ion dose, the annealing temperature is defined as the temperature at which 90% of the implanted ions are activated by a 30 minute annealing in a conventional annealing furnace. For boron implantation, higher annealing temperatures are needed for higher doses. For phosphorus at lower doses, the annealing behavior is similar to that for boron. However, when the dose is greater than 10^{15} cm^{-2} , the annealing temperature drops to about 600°C . This phenomenon is related to the *solid-phase epitaxy* process. At phosphorus doses greater than $6 \times 10^{14} \text{ cm}^{-2}$, the silicon surface layer becomes amorphous. The single-crystal semiconductor underneath the amorphous layer serves as a seeding area for recrystallization of the amorphous layer. The epitaxial-growth rate along the $\langle 100 \rangle$ direction is 10 nm/min at 550°C and 50 nm/min at 600°C , with an activation energy at 2.4 eV. Therefore, a 100–500 nm amorphous layer can be recrystallized in a few minutes. During the solid-phase epitaxial process, the impurity dopant atoms are incorporated into the lattice sites along with the host atoms; thus, full activation can be obtained at relatively low temperatures.

Rapid Thermal Annealing

The machine for RTA with a transient lamp heating is shown in Fig. 23. The temperature measured²² from the heated wafer is usually from 600°C – 1100°C . A wafer is heated quickly under atmospheric conditions or at low pressure under isothermal conditions. Typical lamps in a RTA system are tungsten filaments or arc lamps. The processing chamber is made of either quartz, silicon carbide, stainless steel, or aluminum and has quartz windows through which the optical radiation passes to illuminate the wafer. The wafer holder is often made of quartz and contacts the wafer in a minimum number of places. A measurement system is placed in a control loop to set wafer temperature. The RTA system interfaces with a gas-handling system and a computer that controls system operation. Typically, wafer temperature in a RTA system is measured with a noncontact optical pyrometer that determines temperature from radiated infrared energy.

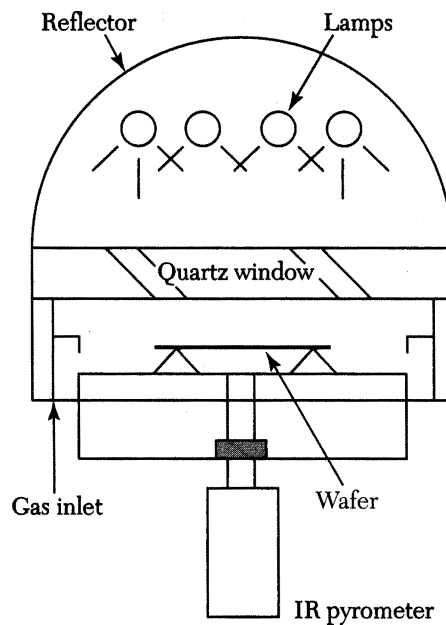


Fig. 23 Rapid thermal annealing system that is optically heated.

A comparison between conventional furnace and RTA technologies is shown in Table 2. To achieve short processing times using RTA, trade-offs must be made in temperature and process uniformity, temperature measurement and control, and wafer stress and throughput. In addition, there are concerns over the introduction of electrically active wafer defects during the very fast (100° – 300°C/s) thermal transients. Rapid heating with temperature gradients in the wafers can cause wafer damage in the form of slip dislocations induced by thermal stress. On the other hand, conventional furnace processing brings with it significant problems, such as particle generation from the hot walls, limited ambient control in an open system, and a large thermal mass that restricts controlled heating times to tens of minutes. In fact, requirements on contamination, process control, and cost of manufacturing floor space have resulted in the paradigm shift to the RTA process.

TABLE 2 Technology Comparison

Determinant	Conventional furnace	Rapid thermal annealing
Process	Batch	Single-wafer
Furnace	Hot-wall	Cold-wall
Heating rate	Low	High
Cycle time	High	Low
Temperature monitor	Furnace	Wafer
Thermal budget	High	Low
Particle problem	Yes	Minimal
Uniformity and repeatability	High	Low
Throughput	High	Low

13.6 IMPLANTATION-RELATED PROCESSES

In this section we consider a few implantation-related processes, such as multiple implantation, masking, tilt-angle implantation, high-energy implantation, and high-current implantation.

13.6.1 Multiple Implantation and Masking

In many applications doping profiles other than the simple Gaussian distribution are required. One such case is the preimplantation of silicon with an inert ion to make the silicon surface region amorphous. This technique allows close control of the doping profile and permits nearly 100% dopant activation at low temperatures, as discussed previously. In such a case, a deep amorphous region may be required. To obtain this type of region, we must make a series of implants at varying ion energies and doses.

Multiple implantation can also be used to form a flat doping profile, as shown in Fig. 24. Here, four boron implants into silicon are used to provide a composite doping profile.²³ The measured carrier concentration and that predicted using range theory are shown in the figure. Other doping profiles, unavailable from diffusion techniques, can be obtained by using various combinations of impurity dose and implantation energy. Multiple implants have been used to preserve stoichiometry during the implantation and annealing of GaAs. This approach, whereby equal amount of gallium and *n*-type dopant (or arsenic and *p*-type dopant) are implanted prior to annealing, has resulted in higher carrier activation.

To form *p-n* junctions in selected areas of the semiconductor substrate, an appropriate mask should be used for the implantation. Because implantation is a low-temperature process, a large variety of masking materials can be used. The minimum thickness of masking material required to stop a given percentage of incident ions can be estimated from the range parameters for ions. The inset of Fig. 25 shows a profile of an implant in

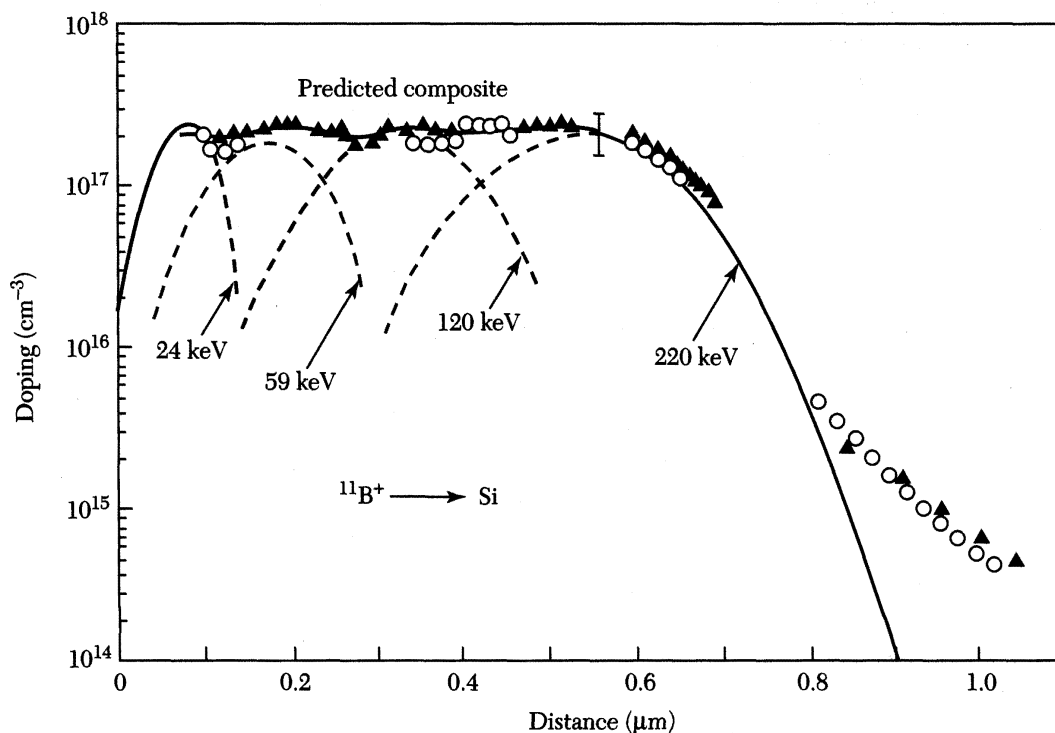


Fig. 24 Composite doping profile using multiple implants.²³

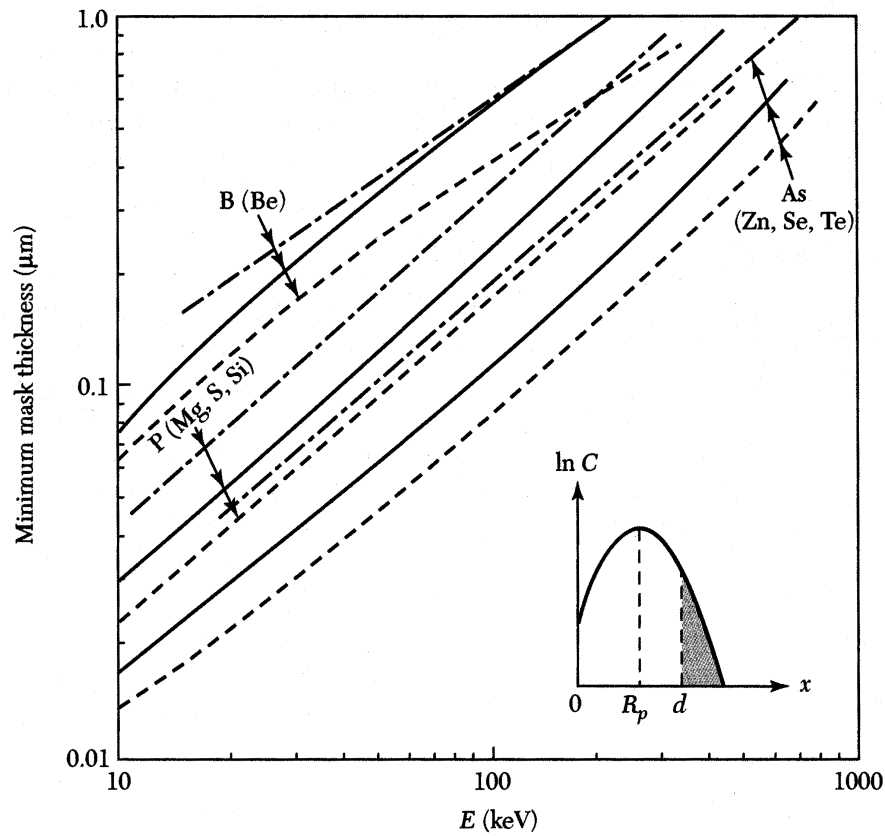


Fig. 25 Minimum thickness of SiO_2 (—), Si_3N_4 (---), and photoresist (- · - · -) to produce a masking effectiveness of 99.99%.

a masking material. The dose implanted in the region beyond a depth d (shown shaded) is given by integration of Eq. 25 as

$$S_d = \frac{S}{\sqrt{2\pi} \sigma_p} \int_d^\infty \exp \left[-\left(\frac{x - R_p}{\sqrt{2}\sigma_p} \right)^2 \right] dx. \quad (33)$$

From Table 1 we can derive the expression

$$\int_x^\infty e^{-y^2} dy = \frac{\sqrt{\pi}}{2} \text{erfc}(x). \quad (34)$$

Therefore, the fraction of the dose that has “transmitted” beyond a depth d is given by the transmission coefficient T :

$$T \equiv \frac{S_d}{S} = \frac{1}{2} \text{erfc} \left(\frac{d - R_p}{\sqrt{2} \sigma_p} \right). \quad (35)$$

Once T is given, we can obtain the mask thickness d from Eq. 35 for any given R_p and σ_p .

The values of d to stop 99.99% of the incident ions ($T = 10^{-4}$) are shown in Fig. 25 for SiO_2 , Si_3N_4 , and a photoresist as masking materials.^{19,24} Mask thicknesses given in this

figure are for boron, phosphorus, and arsenic implanted into silicon. These mask thicknesses also can be used as guidelines for impurity masking in gallium arsenide. The dopants are shown in the parentheses. Since both R_p and σ_p vary approximately linearly with energy, the minimum thickness of the masking material also increases linearly with energy. In certain applications, instead of totally stopping the beam, the masks can be used as attenuators, which can provide an amorphous surface layer to the incident ion beam to minimize the channeling effect.

► EXAMPLE 4

When boron ions are implanted at 200 keV, what thickness of SiO_2 will be required to mask 99.996% of the implanted ions ($R_p = 0.53 \mu\text{m}$, $\sigma_p = 0.093 \mu\text{m}$)?

SOLUTION The complementary error function in Eq. 35 can be approximated if the argument is large (see Table 1):

$$T \cong \frac{1}{2\sqrt{\pi}} \frac{e^{-u^2}}{u},$$

where the parameter u is given by $(d - R_p)/\sqrt{2}\sigma$. For $T = 10^{-4}$, we can solve the above equation to give $u = 2.8$. Thus,

$$d = R_p + 3.96\sigma_p = 0.53 + 3.96 \times 0.093 = 0.898 \mu\text{m}$$

13.6.2 Tilt-Angle Ion Implantation

In scaling devices to submicron dimensions, it is important also to scale dopant profiles vertically. We need to produce junction depths less than 100 nm, including diffusion during dopant activation and subsequent processing steps. Modern device structures, such as the lightly doped drain (LDD), require precise control of dopant distributions vertically and laterally.

It is the ion velocity perpendicular to the surface that determines the projected range of an implanted ion distribution. If the wafer is tilted at a large angle to the ion beam, then the effective ion energy is greatly reduced. Figure 26 illustrates this for 60 keV arsenic ions as a function of the tilt angle, showing that it is possible to achieve extremely shallow distributions using a high tilt angle (86°). In tilt-angle ion implantation, we should consider the shadow effect (inset in Fig. 26) for the patterned wafer. A lower tilt angle leads to a small shadow area. For example, if the height of patterned mask is $0.5 \mu\text{m}$, with vertical sidewall, a 7° incident ion beam will induce a 61 nm shadow. This shadow effect may introduce an unexpected series resistance in the device.

13.6.3 High-Energy and High-Current Implantation

High-energy implantors, capable of energies as high as 1.5–5 MeV, are available and have been used for a number of novel applications. The majority of these depend on the ability to dope the semiconductor to many micrometers in depth, without the need for long diffusion times at high temperatures. High-energy implantors can also be used to produce low-resistivity buried layers. For example, a buried layer $1.5\text{--}3 \mu\text{m}$ below the surface for a CMOS device can be achieved by high-energy implantation.

High-current implantors (10–20 mA), operating in the 25–30 keV range, are routinely used for the predeposition step in diffusion technology because the amount of total

dopant can be controlled precisely. After the predeposition, the dopant impurities can be driven in by a high-temperature diffusion step at the same time that implant damage at the surface region is annealed out. Another application is the threshold voltage adjustment in MOS devices. A precisely controlled amount of dopant (e.g., boron) is implanted through the gate oxide to the channel region²⁵ (Fig. 27a). Because the projected range of boron in silicon and silicon oxide are comparable, if we choose a suitable incident energy, the ions will penetrate just the thin gate oxide, not the thicker-field oxide. The threshold voltage will vary approximately linearly with the implanted dose. After boron implantation, polysilicon can be deposited and patterned to form the gate electrode of the MOSFET. The thin oxide surrounding the gate electrode is removed, and the source and drain regions are formed as shown in Fig. 27b by another high-dose arsenic implantation.

High-current implantors with energies in the 150–200 keV range are now available. A major use for these machines is to form high-quality silicon films, which are insulated

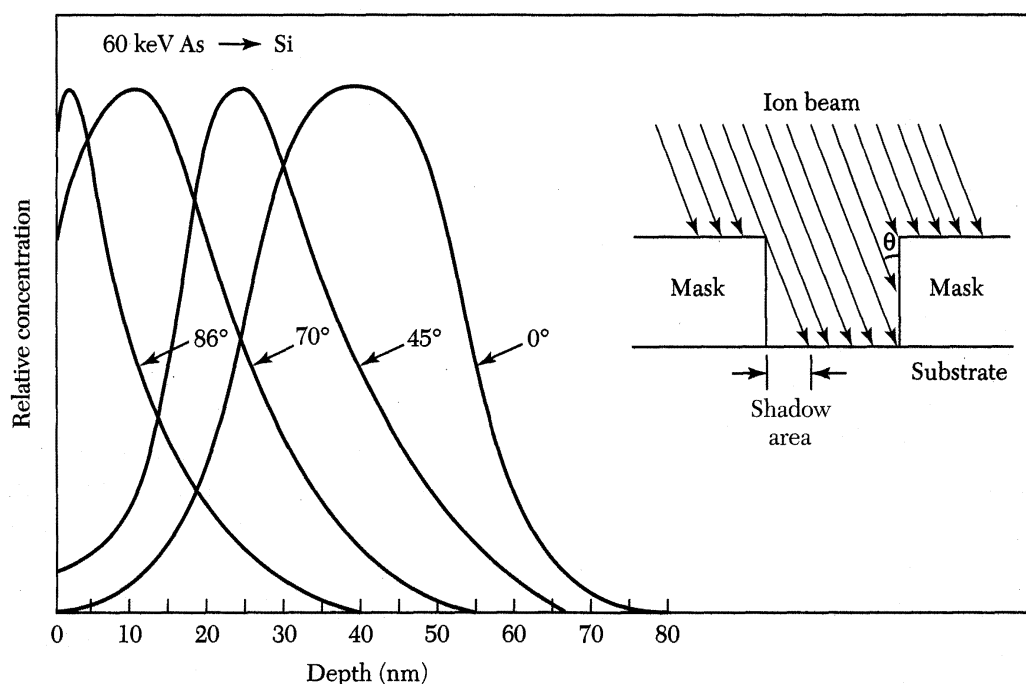


Fig. 26 60 keV arsenic implanted into silicon, as a function of beam tilt angle. Inset shows the shadow area for tilt-angle ion implantation.

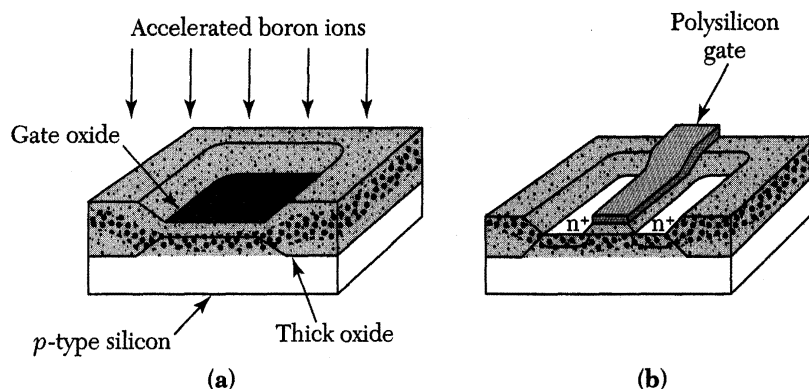


Fig. 27 Threshold voltage adjustment using boron ion implantation.²⁵

from the substrate by implanting oxygen, creating an intervening layer of silicon dioxide. This separation by implantation of oxygen (SIMOX) is a key SOI (silicon-on-insulator) technology.

The SIMOX process uses a high-energy O^+ beam, typically in the 150 to 200 keV range, so that the oxygen ions have projected ranges of 100–200 nm. Additionally, a heavy dose, $1\text{--}2 \times 10^{18}$ ions/cm², is used to produce an insulating layer of SiO_2 that is 100–500 nm thick. The use of SIMOX material leads to a significant reduction of source/drain capacitances in MOS devices. Moreover, it reduces coupling between devices and, thus, allows tighter packing without the problem of latchup. As a result, it is widely proposed as the material of choice for advanced, high-speed CMOS circuits.

► SUMMARY

Diffusion and ion implantation are the two key methods for impurity doping. We first considered the basic diffusion equation for constant diffusivity. We obtain the complementary error function (erfc) and the Gaussian function for the constant-surface-concentration case and constant-total-dopant case, respectively. The results of a diffusion process can be evaluated by measurements of the junction depth, the sheet resistance, and the dopant profile.

When the doping concentration is higher than the intrinsic carrier concentration n_i at the diffusion temperature, the diffusivity becomes concentration dependent. This dependence has profound effect on the resulting doping profile. For example, arsenic and boron diffusivity in silicon vary linearly with the impurity concentration. Their doping profiles are much more abrupt than the erfc profile. Phosphorus diffusivity in silicon varies as the square of concentration. This dependence and a dissociation effect give rise to a phosphorus diffusivity that is 100 times larger than its intrinsic diffusivity.

Lateral diffusion at the edge of a mask and impurity redistribution during oxidation are two processes in which diffusion can have an important impact on device performance. The former can substantially reduce the breakdown voltage, and the latter will influence the threshold voltage as well as the contact resistance.

The key parameters for ion implantation are the projected range R_p and its standard deviation σ_p , also called the projected straggle. The implantation profile can be approximated by a Gaussian distribution with its peak located at R_p from the surface of the semiconductor substrate. The advantages of ion implantation process are more precise control of the amount of dopant, a more reproducible doping profile, and lower processing temperature compared with the diffusion process.

We considered R_p and σ_p for various elements in silicon and gallium arsenide and discussed the channeling effect and ways to minimize this effect. However, implantation may cause severe damage to the crystal lattice. To remove the implant damage and to restore mobility and other device parameters, we must anneal the semiconductor at an appropriate combination of time and temperature. Currently, the rapid thermal annealing (RTA) is preferred to conventional furnace annealing because RTA can remove implant damage without thermal broadening of the doping profile.

Ion implantation has wide applications for advanced semiconductor devices. These include (a) multiple implantation to form novel distributions, (b) selection of masking materials and thickness to stop a given percentage of incident ions from reaching the substrate, (c) tilt-angle implantation to form ultrashallow junctions, (d) high-energy implantation to form buried layers, and (e) high-current implantation for predeposition and threshold voltage adjustment and to form an insulating layer for SOI applications.