LIBRARY OF CONGRESS
0 008 885 083 A

This eagerly-anticipated revision offers more than 50% new [illegible] the multitude of important recent discoveries and advances in device physics and integrated circuit processing.

The book offers a thorough introduction to physical principles of modern semiconductor devices and their fabrication technology. Readers are presented with theoretical and practical aspects of every step in device characterizations and fabrication, with an emphasis on integrated circuits.

The material is divided into three parts:

1. the basic properties of semiconductor materials, emphasizing silicon and gallium arsenide
2. the physics and characteristics of semiconductor devices bipolar, unipolar special microwave and photonic devices
3. the latest processing technologies, from crystal growth to lithographic pattern transfer

Each chapter is presented in a logical manner enabling readers to learn all important devices from a single source. Plus, the book covers historical developments of devices and technology in the last 100 years. Readers gain a sound perspective on the past and a foundation for projecting future trends.

ABOUT THE AUTHOR

S. M. Sze is UMC Chair Professor of the National Chiao Tung University and President of the National Nano Device Laboratories, Taiwan, R.O.C. For many years he was a member of the technical staff at Bell Laboratories. Professor Sze is the co-inventor of the nonvolatile semiconductor memory. He has written numerous texts on devices physics, including PHYSICS OF SEMICONDUCTOR DEVICES, considered a reference classic. In 1991, he received the IEEE J. J. Ebers award for his "fundamental and pioneering contributions..." He received his PhD in solid-state electronics from Stanford University in 1963.

ISBN 0-471-33372-7
90000>
9 780471 333722

JOHN WILEY & SONS, INC.
New York / Chichester
Weinheim / Brisbane
Singapore / Toronto
http://www.wiley.com/college/sze

Sci RR

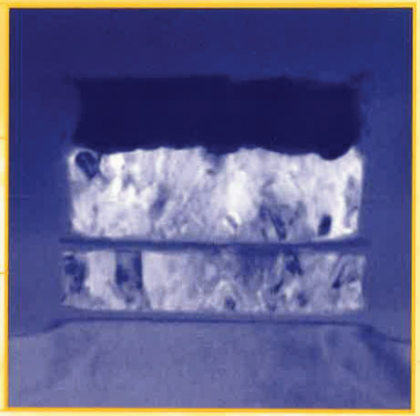
S.M. Sze

SEMICONDUCTOR DEVICES Physics and Technology

TK 7871
.85
.S9883
2001
COPY 1

SEMICONDUCTOR DEVICES

Physics and Technology



2nd Edition

S.M. Sze

TI 1022

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.

LIBRARY OF CONGRESS
SEP 19 2001
CIP

TK7871
.85
.S9883
2001
Copy 1
Sci&R

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 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 custome 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 2 1

In Memory of My Mentors

Dr. L. J. Chu — *Academia Sinica*

Dr. R. M. Ryder — *Bell Laboratories*

FOR SECTION 13.5 IMPLANT DAMAGE AND ANNEALING

16. If a 50 keV boron ion is implanted into the silicon substrate, calculate the damage density. Assume silicon atom density is 5.02×10^{22} atoms/cm^3, the silicon displacement energy is 15 eV, the range is 2.5 nm, and the spacing between silicon lattice plane is 0.25 nm.
17. Explain why high-temperature RTA is preferable to low-temperature RTA for defect-free shallow-junction formation.
18. Estimate the implant dose required to reduce a p-channel threshold voltage by 1 V if the gate oxide is 4 nm thick. Assume that the implant voltage is adjusted so that the peak of the distribution occurs at the oxide-silicon interface. Thus, half of the implant goes into the silicon. Further, assume that 90% of the implanted ions in the silicon are electrically activated by the annealing process. These assumptions allow 45% of the implanted ions to be used for threshold adjusting. Also assume that all of the charge in the silicon is effectively at the silicon-oxide interface.

FOR SECTION 13.6 IMPLANTATION-RELATED PROCESSES

19. We would like to form 0.1 μm deep, heavily doped junctions for the source and drain regions of a submicron MOSFET. Compare the options that are available to introduce and activate dopant for this application. Which option would you recommend and why?
20. When an arsenic implant at 100 keV is used and the photoresist thickness is 400 nm, find the effectiveness of the resist mask in preventing the transmission of ions (R_p = 0.6 μm, σ_p = 0.2 μm). If the resist thickness is changed to 1 μm, calculate the masking efficiency.
21. With reference to Ex. 4, what thickness of SiO_2 is required to mask 99.999% of the implanted ions?

CHAPTER 14

Integrated Devices

- 14.1 PASSIVE COMPONENTS
- 14.2 BIPOLAR TECHNOLOGY
- 14.3 MOSFET TECHNOLOGY
- 14.4 MESFET TECHNOLOGY
- 14.5 CHALLENGES FOR MICROELECTRONICS
- SUMMARY

Microwave, photonic, and power applications generally employ discrete devices. For example, an IMPATT diode is used as a microwave generator, an injection laser as an optical source, and a thyristor as a high-power switch. However, most electronic systems are built on the integrated circuit (IC), which is an ensemble of both active (e.g., transistor) and passive devices (e.g., resistor, capacitor, and inductor) formed on and within a single-crystal semiconductor substrate and interconnected by a metallization pattern.[1] ICs have enormous advantages over discrete devices connected by wire bondings. The advantages includes (a) reduction of the interconnection parasitics, because an IC with multilevel metallization can substantially reduce the overall wiring length, (b) full utilization of semiconductor wafer's "real estate," because devices can be closely packed within an IC chip, and (c) drastic reduction in processing cost, because wire bonding is a time-consuming and error-prone operation.

In this chapter we combine the basic processes described in previous chapters to fabricate active and passive components in an IC. Because the key element of an IC is the transistor, specific processing sequences are developed to optimize its performance. We consider three major IC technologies associated with the three transistor families: the bipolar transistor, the MOSFET, and the MESFET.

Specifically, we cover the following topics:

- The design and fabrication of IC resistor, capacitor, and inductor.
- The processing sequence for standard bipolar transistor and advanced bipolar devices.
- The processing sequence for MOSFET with special emphasis on CMOS and memory devices.
- The processing sequence for high-performance MESFET and monolithic microwave IC.
- The major challenges for future microelectronics, including ultrashallow junction, ultrathin oxide, new interconnection materials, low power dissipation, and isolation.

Figure 1 illustrates the interrelationship between the major process steps used for IC fabrication. Polished wafers with a specific resistivity and orientation are used as the starting material. The film formation steps include thermally grown oxide films, deposited polysilicon, dielectric, and metal films (Chapter 11). Film formation is often followed by lithography (Chapter 12) or impurity doping (Chapter 13). Lithography is generally followed by etching, which in turn is often followed by another impurity doping or film formation. The final IC is made by sequentially transferring the patterns from each mask, level by level, onto the surface of the semiconductor wafer.

After processing, each wafer contains hundreds of identical rectangular chips (or dice), typically between 1 and 20 mm on each side, as shown in Fig. 2*a*. The chips are separated by sawing or laser cutting; Figure 2*b* shows a separated chip. Schematic top views of a single MOSFET and a single bipolar transistor are shown in Fig. 2*c* to give some perspective of the relative size of a component in an IC chip. Prior to chip separation, each chip is electrically tested. Defective chips are usually marked with a dab of black ink. Good chips are selected and packaged to provide an appropriate thermal, electrical, and interconnection environment for electronic applications.[2]

IC chips may contain from a few components (transistors, diodes, resistors, capacitors, etc.) to as many as a billion or more. Since the invention of the monolithic IC in 1959, the number of components on a state-of-the-art IC chip has grown exponentially. We usually refer to the complexity of an IC as small-scale integration (SSI) for up to 100 components per chip, medium-scale integration (MSI) for up to 1000 components per chip, large-scale integration (LSI) for up to 100,000 components per chip, very-large-scale integrated (VLSI) for up to 10^7 components per chip, and ultra large-scale integration (ULSI) for larger numbers of components per chip. In Section 14.3, we show two ULSI chips, a 32-bit microprocessor chip, which contains over 42 million components, and a 1 Gbit dynamic random access memory (DRAM) chip, which contains over 2 billion components.

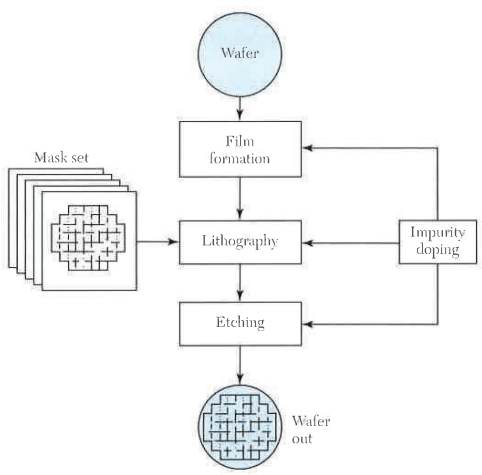


Fig. 1 Schematic flow diagram of integrated-circuit fabrication.

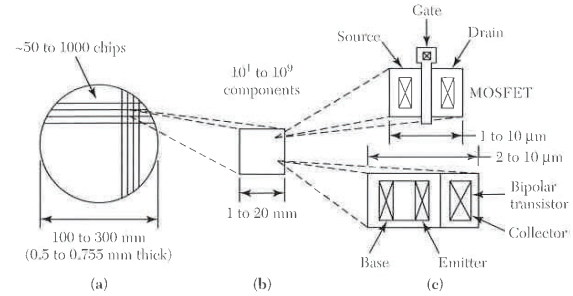


Fig. 2 Size comparison of a wafer to individual components. (*a*) Semiconductor wafer. (*b*) Chip. (*c*) MOSFET and bipolar transistor.

14.1 PASSIVE COMPONENTS

14.1.1 The Integrated-Circuit Resistor

To form an IC resistor, we can deposite a resistive layer on a silicon substrate, then pattern the layer by lithography and etching. We can also define a window in a silicon dioxide layer grown thermally on a silicon substrate and then implant (or diffuse) impurities of the opposite conductivity type into the wafer. Figure 3 shows the top and cross-sectional views of two resistors formed by the latter approach: one has a meander shape and the other has a bar shape.

Consider the bar-shaped resistor first. The differential conductance dG of a thin layer of the p-type material that is of thickness dx parallel to the surface and at a depth x (as shown by the B-B cross section) is

$$dG = q\mu_p p(x)\frac{W}{L}dx, \tag{1}$$

where W is the width of the bar, L is the length of the bar (we neglect the end contact areas for the time being), μ_p is mobility of hole, and $p(x)$ is the doping concentration. The total conductance of the entire implanted region of the bar is given by

$$\boxed{G = \int_0^{x_j} dG = q\frac{W}{L}\int_0^{x_j} \mu_p p(x)dx,} \tag{2}$$

where x_j is the junction depth. If the value of μ_p, which is a function of the hole concentration, and the distribution of $p(x)$ are known, the total conductance can be evaluated from Eq. 2. We can write

$$G \equiv g\frac{W}{L}, \tag{3}$$

where $g \equiv q\int_0^{x_j} \mu_p p(x)dx$ is the conductance of a square resistor pattern, that is, $G = g$ when $L = W$.

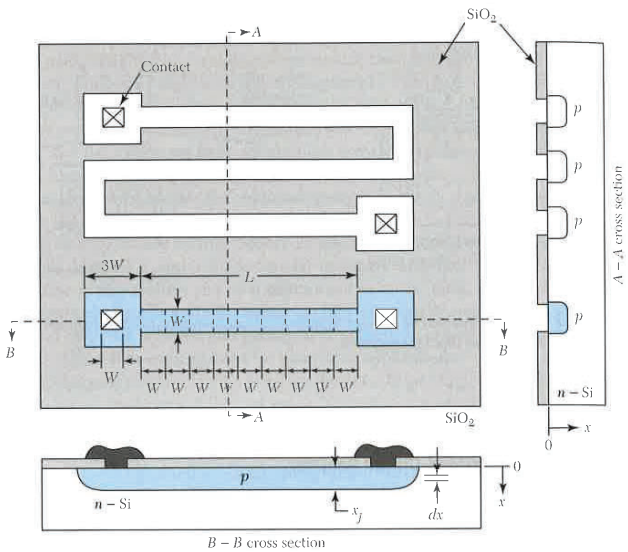


Fig. 3 Integrated-circuit resistors. All narrow lines in the large square area have the same width W, and all contacts are the same size.

The resistance is therefore given by

$$R \equiv \frac{1}{G} = \frac{L}{W}\left(\frac{1}{g}\right), \tag{4}$$

where $1/g$ usually is defined by the symbol $R_\square$ and is called the sheet resistance. The sheet resistance has units of ohms but is conventionally specified in units of ohms per square ($\Omega/\square$).

Many resistors in an integrated circuit are fabricated simultaneously by defining different geometric patterns in the mask such as those shown in Fig. 3. Since the same processing cycle is used for all these resistors, it is convenient to separate the resistance into two parts: the sheet resistance $R_\square$, determined by the implantation (or diffusion) process; and the ratio L/W, determined by the pattern dimensions. Once the value of $R_\square$ is known, the resistance is given by the ratio L/W, or the number of squares (each square has an area of $W \times W$) in the resistor pattern. The end contact areas will introduce additional resistance to the IC resistors. For the type shown in Fig. 3, each end contact corresponds to approximately 0.65 square. For the meander-shape resistor, the electric-field lines at the bends are not spaced uniformly across the width of the resistor but are crowded toward the inside corner. A square at the bend does not contribute exactly 1 square, but rather 0.65 square.

▶ **EXAMPLE 1**

Find the value of a resistor 90 μm long and 10 μm wide, such as the bar-shaped resistor in Fig. 3. The sheet resistance is 1 kΩ/□.

SOLUTION The resistor contains 9 squares. The two end contacts correspond to 1.3 □. The value of the resistor is (9 + 1.3) × 1 kΩ/□ = 10.3 kΩ. ◀

14.1.2 The Integrated-Circuit Capacitor

There are basically two types of capacitors used in integrated circuits: MOS capacitors and p–n junctions. The MOS (metal-oxide-semiconductor) capacitor can be fabricated by using a heavily doped region (such as an emitter region) as one plate, the top metal electrode as the other plate, and the intervening oxide layer as the dielectric. The top and cross-sectional views of a MOS capacitor are shown in Fig. 4*a*. To form a MOS capacitor, a thick oxide layer is thermally grown on a silicon substrate. Next, a window is lithographically defined and then etched in the oxide. Diffusion or ion implantation is used to form a p^+-region in the window area, whereas the surrounding thick oxide serves as a mask. A thin oxide layer is then thermally grown in the window area, followed by a metallization step. The capacitance per unit area is given by

$$C = \frac{\varepsilon_{ox}}{d} \text{ F/cm}^2, \tag{5}$$

where ε_{ox} is the dielectric permittivity of silicon dioxide (the dielectric constant $\varepsilon_{ox}/\varepsilon_0$ is 3.9) and d is the thin-oxide thickness. To increase the capacitance further, insulators with higher dielectric constants are being studied, such as Si_3N_4, and Ta_2O_5, with dielectric constants of 7 and 25, respectively. The MOS capacitance is essentially independent of the applied voltage, because the lower plate of the capacitor is made of heavily doped material. This also reduces the series resistance associated with it.

A p–n junction is sometimes used as a capacitor in an integrated circuit. The top and cross sectional views of an n^+-p junction capacitor are shown in Fig. 4*b*. The detailed

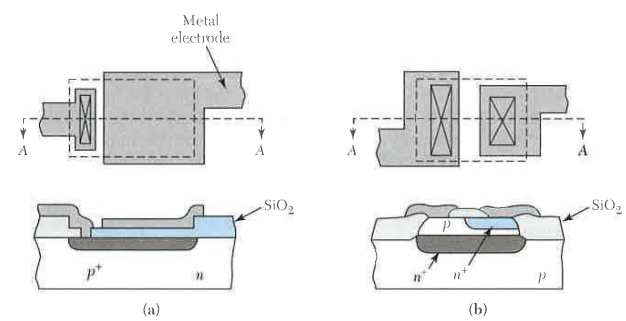


Fig. 4 (*a*) Integrated MOS capacitor. (*b*) Integrated p–n junction capacitor.

fabrication process is considered in Section 14.2, because this structure forms part of a bipolar transistor. As a capacitor, the device is usually reverse biased, that is, the p-region is reverse-biased with respect to the n^+-region. The capacitance is not a constant but varies as $(V_R + V_{bi})^{-1/2}$, where V_R is the applied reverse voltage and V_{bi} is the built-in potential. The series resistance is considerably higher than that of a MOS capacitor because the p-region has higher resistivity than does the p^+-region.

▸ **EXAMPLE 2**

What is the stored charge and the number of electrons on an MOS capacitor with an area of 4 μm^2, for (a) a dielectric of 10 nm thick SiO_2 and (b) a 5 nm thick Ta_2O_5. The applied voltage is 5 V for both cases.

SOLUTION

(a) $$Q = \varepsilon_{ox} \times A \times \frac{V_s}{d} = 3.9 \times 8.85 \times 10^{-14}\ \text{F/cm} \times 4 \times 10^{-8}\ \text{cm}^2 \times \frac{5\text{V}}{1 \times 10^{-6}\ \text{cm}} = 6.9 \times 10^{-14}\ \text{C}$$

or

$$Q_s = 6.9 \times 10^{-14}\ \text{C}/q = 4.3 \times 10^5 \text{ electrons.}$$

(b) Changing the dielectric constant from 3.9 to 25 and the thickness from 10 nm to 5 nm, we obtain $Q_s = 8.85 \times 10^{-13}$ C, and $Q_s = 8.85 \times 10^{-13}\ \text{C}/q = 5.53 \times 10^6$ electrons. ◂

14.1.3 The Integrated-Circuit Inductor

IC inductors have been widely used in III-V based monolithic microwave integrated circuits (MMIC)[3]. With the increased speed of silicon devices and advancement in multilevel interconnection technology, IC inductors have started to receive more and more attentions in silicon-based radio frequency (rf) and high-frequency applications. Many kinds of inductors can be fabricated using IC processes. The most popular method is the thin-film spiral inductor, Figure 5*a* and *b* shows the top-view and the cross section of a silicon-based, two-level–metal spiral inductor. To form a spiral inductor, a thick oxide is thermally grown or deposited on a silicon substrate. The first metal is then deposited and defined as one end of the inductor. Next, another dielectric is deposited onto the metal 1. A via hole is defined lithographically and etched in the oxide. Metal 2 is deposited and the via hole is filled. The spiral patterned can be defined and etched on the metal 2 as the second end of the inductor.

To evaluate the inductor, an important figure of merit is the quality factor, Q. The Q is defined as $Q = L\omega/R$, where L, R, and ω are the inductance, resistance, and frequency, respectively. The higher the Q values, the lower the loss from resistance, hence the better the performance of the circuits. Figure 5*c* shows the equivalent circuit model. R_1 is the inherent resistivity of the metal, C_{p1} and C_{p2} are the coupling capacitances between the metal lines and the substrate, and R_{sub1} and R_{sub2} are the resistances of the silicon substrate associated with the metal lines, respectively. The Q increases linearly with frequency initially and then drops at higher frequencies because of parasitic resistances and capacitances.

There are some approaches to improve the Q value. The first is to use low-dielectric-constant materials (<3.9) to reduce the C_p. The other is to use a thick film metal or low-resistivity metals (e.g., Cu, Au to replace Al) to reduce the R_1. The third approach

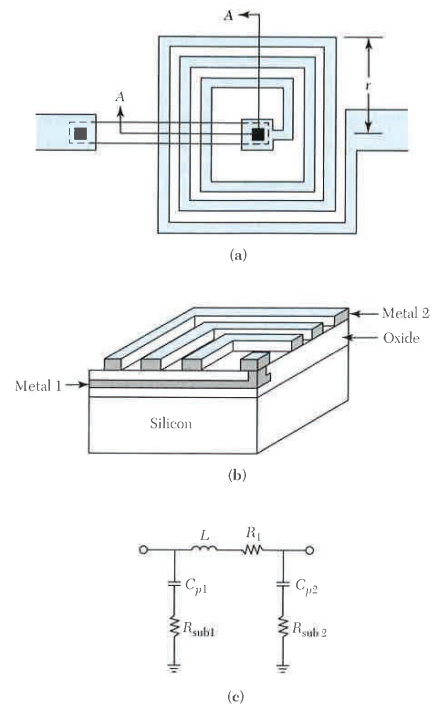


Fig. 5 (*a*) Schematic view of a spiral inductor on a silicon substrate. (*b*) Perspective view along *A-A′*. (*c*) An equivalent circuit model for an integrated inductor.

uses an insulating substrate (e.g., silicon-on-sapphire, silicon-on-glass, or quartz) to reduce R_{sub}.

To obtain the exact value of a thin-film inductor, complicated simulation tool, such as computer aided design, must be employed for both circuit simulation and inductor optimization. The model for thin-film inductor must take into account the resistance of the metal, the capacitance of the oxide, line-to-line capacitance, the resistance of the substrate, the capacitance to the substrate, and the inductance and mutual inductance of the metal lines. Hence, it is more difficult to calculate the integrated inductance compared with the integrated capacitors or resistors. However, a simple equation to estimate the square planar spiral inductor is given as[3]

$$\boxed{L \approx \mu_0 n^2 r \approx 1.2 \times 10^{-6} n^2 r,} \qquad (6)$$

where μ_0 is the permeability in vacuum ($4\pi \times 10^{-7}$ H/m), L is in henries, n is the number of turns, and r is the radius of the spiral in meters.

▶ EXAMPLE 3

For an integrated inductor with an inductance of 10 nH, what is the required radius if the number of turns is 20?

SOLUTION According to the Eq. 6,

$$r = \frac{10 \times 10^{-9}}{1.2 \times 10^{-6} \times 20^2} = 2.08 \times 10^{-5}\,(\text{m}) = 20.8\ \mu\text{m}.$$

▶ 14.2 BIPOLAR TECHNOLOGY

For IC applications, especially for VLSI and ULSI, the size of bipolar transistors must be reduced to meet the high-density requirement. Figure 6 illustrates the reduction in the size of the bipolar transistor in recent years.[4] The main differences in a bipolar transistor in an IC compared with a discrete transistor are that all electrode contacts are located on the *top surface* of the IC wafer, and each transistor must be electrically *isolated* to prevent interactions between devices. Prior to 1970, both the lateral and vertical isolations were provided by p–n junctions (Fig. 6*a*) and the lateral p-isolation region was always reverse biased with respect to the n-type collector. In 1971, thermal oxide was used for lateral isolation, resulting in a substantial reduction in device size (Fig. 6*b*), because the base and collector contacts abut the isolation region. In the mid-1970s, the emitter extended to the walls of the oxide, resulting in an additional reduction in area (Fig. 6*c*). At the present time, all the lateral and vertical dimensions have been scaled down and emitter stripe widths have dimensions in the submicron region (Fig. 6*d*).

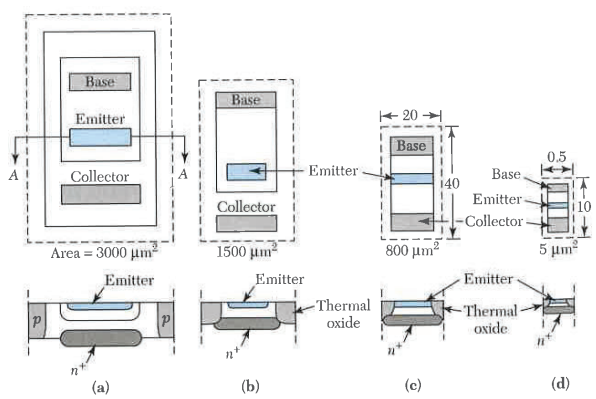


Fig. 6 Reduction of the horizontal and vertical dimensions of a bipolar transistor. (a) Junction isolation. (*b*) Oxide isolation. (*c* and *d*) Scaled oxide isolation.[4]

14.2.1 The Basic Fabrication Process

The majority of bipolar transistor used in ICs are of the n-p-n type because the higher mobility of minority carriers (electrons) in the base region results in higher-speed performance than can be obtained with p-n-p types. Figure 7 shows a perspective view of an n-p-n bipolar transistor, in which lateral isolation is provided by oxide walls and vertical isolation is provided by the n^+-p junction. The lateral oxide isolation approach reduces not only the device size but also the parasitic capacitance because of the smaller dielectric constant of silicon dioxide (3.9, compared with 11.9 for silicon). We consider the major process steps that are used to fabricate the device shown in Fig. 7.

For an n-p-n bipolar transistor, the starting material is a p-type lightly doped ($\sim 10^{15}$ cm^{-3}), ⟨111⟩- or ⟨100⟩-oriented, polished silicon wafer. Because the junctions are formed inside the semiconductor, the choice of crystal orientation is not as critical as for MOS devices. The first step is to form a buried layer. The main purpose of this layer is to minimize the series resistance of the collector. A thick oxide (0.5–1 μm) is thermally grown on the wafer, and a window is then opened in the oxide. A precisely controlled amount of low-energy arsenic ions (~30 keV, $\sim 10^{15}$ cm^{-2}) is implanted into the window region to serve as a predeposit (Fig. 8*a*). Next, a high temperature (~1100°C) drive-in step forms the n^+-buried layer, which has a typical sheet resistance of 20 Ω/□.

The second step is to deposit an n-type epitaxial layer. The oxide is removed and the wafer is placed in an epitaxial reactor for epitaxial growth. The thickness and the doping concentration of the epitaxial layer are determined by the ultimate use of the device. Analog circuits (with their higher voltages for amplification) require thicker layer (~10 μm) and lower dopings ($\sim 5 \times 10^{15}$ cm^{-3}), whereas digital circuits (with their lower voltages for switching) require thinner layers (~3 μm) and higher dopings ($\sim 2 \times 10^{16}$ cm^{-3}). Figure 8*b* shows a cross-sectional view of the device after the epitaxial process. Note that there is some outdiffusion from the buried layer into the epitaxial layer. To minimize the outdiffusion, a low-temperature epitaxial process should be employed, and low-diffusivity impurities should be used in the buried layer (e.g., As).

The third step is to form the lateral oxide isolation region. A thin-oxide pad (~50 nm) is thermally grown on the epitaxial layer, followed by a silicon-nitride deposition (~100 nm). If nitride is deposited directly onto the silicon without the thin-oxide pad, the nitride may cause damages to the silicon surface during the subsequent high-temperature steps. Next, the nitride-oxide layers and about half of the epitaxial layer are etched using a photoresist

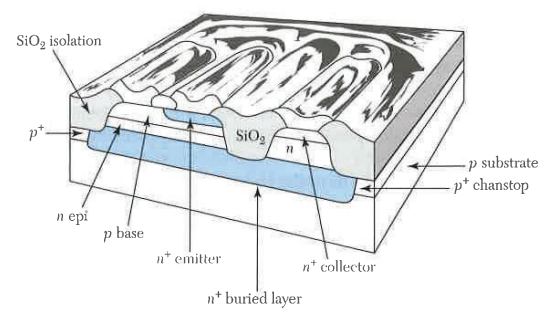


Fig. 7 Perspective view of an oxide-isolated bipolar transistor.

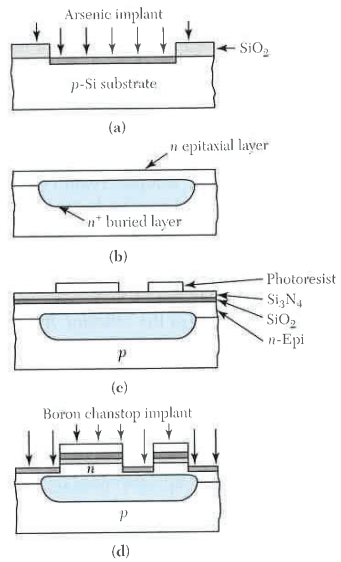


Fig. 8 Cross-sectional views of bipolar transistor fabrication. (*a*) Buried-layer implantation. (*b*) Epitaxial layer. (*c*) Photoresist mask. (*d*) Chanstop implant.

as mask (Fig. 8*c* and 8*d*). Boron ions are then implanted into the exposed silicon areas (Fig. 8*d*).

The photoresist is removed and the wafer is placed in an oxidation furnace. Since the nitride layer has a very low oxidation rate, thick oxides will be grown only in the areas not protected by the nitride layer. The isolation oxide is usually grown to a thickness such that the top of the oxide becomes coplanar with the original silicon surface to minimize the surface topography. This oxide isolation process is called local oxidation of silicon (LOCOS). Figure 9*a* shows the cross section of the isolation oxide after the removal of the nitride layer. Becaue of segregation effects, most of the implanted boron ions are pushed underneath the isolation oxide to form a p^+-layer. This is called p^+ channel stop (or chanstop), because the high concentration of p-type semiconductor will prevent surface inversion and eliminate possible high-conductivity paths (or channels) among neighboring buried layers.

The fourth step is to form the base region. A photoresist is used as a mask to protect the right half of the device; then, boron ions ($\sim 10^{12}$ cm^{-2}) are implanted to form the base regions, as shown in Fig. 9*b*. Another lithographic process removes all the thin-pad oxide except a small area near the center of the base region (Fig. 9*c*).

The fifth step is to form the emitter region. As shown in Fig. 9*d*, the base contact area is protected by a photoresist mask; then, a low-energy, high-arsenic–dose ($\sim 10^{16}$ cm^{-2}) implantation forms the n^+-emitter and the n^+-collector contact regions. The photoresist

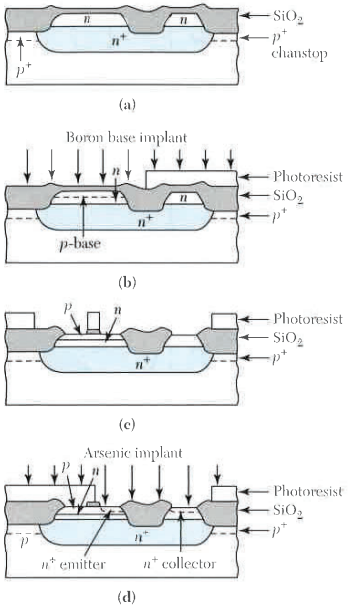


Fig. 9 Cross-section views of bipolar transistor fabrication. (*a*) Oxide isolation. (*b*) Base implant. (*c*) Removal of thin oxide. (*d*) Emitter and collector implant.

is removed; and a final metallization step forms the contacts to the base, emitter, and collector as shown in Fig. 7.

In this basic bipolar process, there are six film formation operations, six lithographic operations, four ion implantations, and four etching operations. Each operation must be precisely controlled and monitored. Failure of any one of the operations generally will render the wafer useless.

The doping profiles of the completed transistor along a coordinate perpendicular to the surface and passing through the emitter, base, and collector are shown in Fig. 10. The emitter profile is abrupt because of the concentration-dependent diffusivity of arsenic. The base doping profile beneath the emitter can be approximated by a Gaussian distribution for a limited-source diffusion. The collector doping is given by the epitaxial doping level ($\sim 2 \times 10^{16}$ cm^{-3}) for a representative switching transistor; however, at larger depths, the collector doping concentration increases because of outdiffusion from the buried layer.

14.2.2 Dielectric Isolation

In the isolation scheme described previously for the bipolar transistor, the device is isolated from other devices by the oxide layer around its periphery and is isolated from its common substrate by a n^+–p junction (buried layer). In high-voltage applications, a different

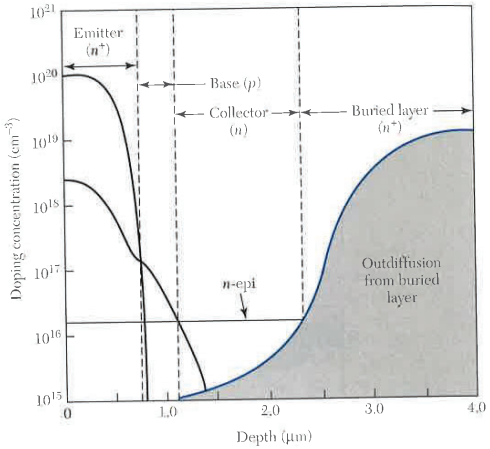


Fig. 10 n-p-n transistor doping profiles.

approach, called dielectric isolation, is used to form insulating tubs to isolate a number of pockets of single-crystal semiconductors. In this approach the device is isolated from both its common substrate and its surrounding neighbors by a dielectric layer.

A process sequence for the dielectric isolation is shown in Fig. 11. An oxide layer is formed inside a <100>-oriented n-type silicon substrate using high-energy oxygen ion implantation (Fig. 11a). Next, the wafer undergoes a high-temperature annealing process so that the implanted oxygen will react with silicon to form the oxide layer. The damage resulting from implantation is also annealed out in this process (Fig. 11b). After this, we can obtain an n-silicon layer that is fully isolated on an oxide [namely, silicon-on-insulator, (SOI)]. This process is called SIMOX (separation by implanted oxygen). Since the top silicon is so thin, the isolation region is easily formed by the LOCOS process illustrated in Fig. 8c or by etching a trench (Fig. 11c) and refilling it with oxide (Fig. 11d). The other processes are almost the same as those from Fig. 8c through Fig. 9 to form the p-type base, n^+-emitter, and collector.

The main advantage of this technique is its high breakdown voltage between the emitter and the collector, which can be in excess of several hundreds volts. This technique is also compatible with modern CMOS integration. This CMOS-compatible process is very useful for mixed high-voltage and high-density IC.

14.2.3 Self-Aligned Double-Polysilicon Bipolar Structure

The process shown in Fig. 9c needs another lithographic process to define an oxide region to separate the base and emitter contact regions. This gives rise to a large inactive device area within the isolated boundary, which increases not only the parasitic capacitances but also the resistance that degrades the transistor performance. The most effective way to reduce these effects is by using the self-aligned structure.

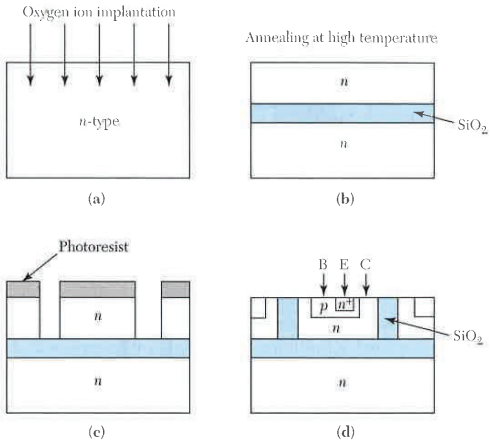


Fig. 11 Process sequence for dielectric isolation bipolar device using silicon-on-insulator for high-voltage application. (a) Oxygen ion implantation. (b) Annealing at high temperature to form the isolation dielectric. (c) Trench isolation formed by a dry-etching process. (d) Base, emitter, and collector formation.

The most widely used self-aligned structure is the double-polysilicon structure with the advanced isolation provided by a trench refilled with polysilicon,[5] shown in Fig. 12. Figure 13 shows the detail sequence of the steps for the self-aligned double-polysilicon (n-p-n) bipolar structure.[6] The transistor is built on an n-type epitaxial layer. A trench of 5.0 μm in depth is etched by reactive ion etching through the n^+-subcollector region into the p^-substrate region. A thin layer of thermal oxide is then grown and serves as the screen oxide for the channel stop implant of boron at the bottom of the trench. The trench is then filled with undoped polysilicon and capped by a thick planar field oxide.

The first polysilicon layer is deposited and heavily doped with boron. The p^+-polysilicon (called poly 1) will be used as a solid-phase diffusion source to form the extrinsic base region and the base electrode. This layer is covered with a chemical-vapor deposition

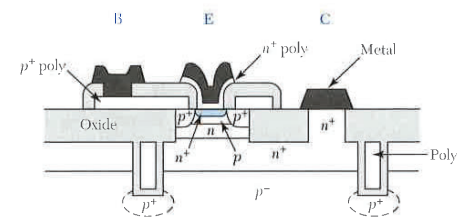


Fig. 12 Cross-section of a self-aligned, double-polysilicon bipolar transistor with advanced trench isolation.[5]

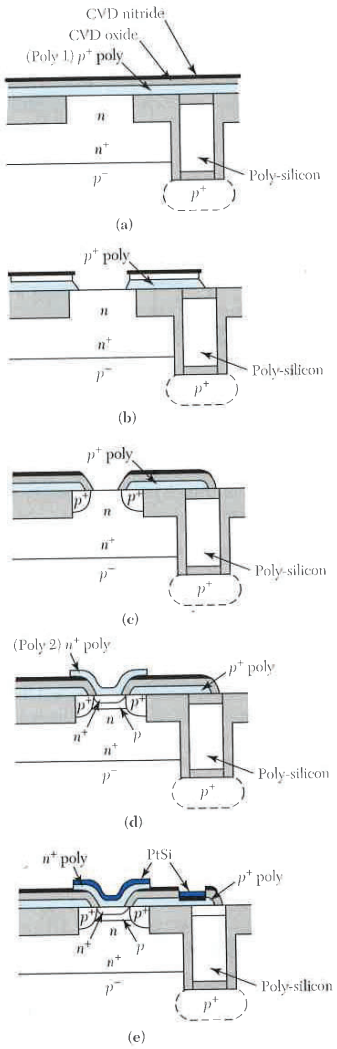


Fig. 13 Process sequence for fabricating double-polysilicon self-aligned n-p-n transistors.[6]

(CVD) oxide and nitride (Fig. 13*a*). The emitter mask is used to pattern the emitter-area regions, and a dry-etch process is used to produce an opening in the CVD oxide and poly 1 (Fig. 13*b*). A thermal oxide is then grown over the etched structure, and a relatively thick oxide (approximately 0.1–0.4 μm) is grown on the vertical sidewalls of the heavily doped poly. The thickness of this oxide determines the spacing between the edges of the base and emitter contacts. The extrinsic p^+ base regions are also formed during the thermal-oxide growth step as a result of the outdiffusion of boron from the poly 1 into the substrate (Fig. 13*c*). Because boron diffuses laterally as well as vertically, the extrinsic base region will be able to make contact with the intrinsic base region that is formed next, under the emitter contact.

Following the oxide-grown step, the intrinsic base region is formed using ion implantation of boron (Fig. 13*d*). This serves to self-align the intrinsic and extrinsic base regions. After the contact is cleaned to remove any oxide layer, the second polysilicon layer is deposited and implanted with As or P. The n^+–polysilicon (called poly 2) is used as a solid-phase diffusion source to form the emitter region and the emitter electrode. A shallow emitter region is then formed through dopant outdiffusion from poly 2. A rapid thermal anneal for the base and emitter outdiffusion steps facilitates the formation of shallow emitter-base and collector-base junctions. Finally, Pt film is deposited and sintered to form PtSi over the n^+-polysilicon emitter and the p^+-polysilicon base contact (Fig. 13*e*).

This self-aligned structure allows the fabrication of emitter regions smaller than the minimum lithographic dimension. When the sidewall-spacer oxide is grown, it fills the contact hole to some degree because the thermal oxide occupies a larger volume than the original volume of polysilicon. Thus, an opening 0.8 μm wide will shrink to about 0.4 μm if sidewall oxide a 0.2 μm thick is grown on each side.

14.3 MOSFET TECHNOLOGY

At present, the MOSFET is the dominant device used in ULSI circuits because it can be scaled to smaller dimensions than other types of devices. The dominant technology for MOSFET is the CMOS (complementary MOSFET) technology, in which both n-channel and p-channel MOSFETs (called NMOS and PMOS, respectively) are provided on the same chip. CMOS technology is particular attractive for ULSI circuits because it has the lowest power consumption of all IC technology.

Figure 14 shows the reduction in the size of the MOSFET in recent years. In the early 1970s, the gate length was 7.5 μm and the corresponding device area was about 6000 μm^2. As the device is scaled down, there is a drastic reduction in the device area. For a MOSFET with a gate length of 0.5 μm, the device area shrinks to less than 1% of the early MOSFET. We expect that device miniaturization will continue. The gate length will be less than 0.10 μm in the early twenty-first century. We consider the future trends of the devices in Section 14.5.

14.3.1 The Basic Fabrication Process

Figure 15 shows a perspective view of an n-channel MOSFET prior to its final metallization.[7] The top layer is a phosphorus-doped silicon dioxide (P-glass) that is used as an insulator between the polysilicon gate and the gate metallization and also as a gettering layer for mobile ions. Compare Fig. 15 with Fig. 7 for the bipolar transistor and note that a MOSFET is considerably simpler in its basic structure. Although both devices use

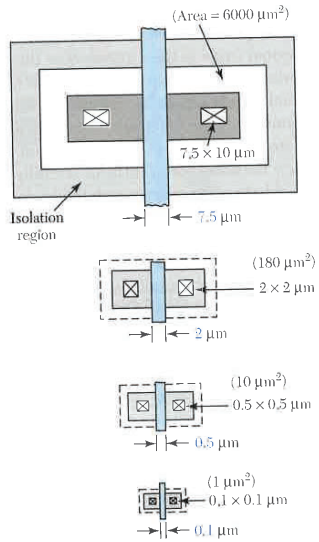


Fig. 14 Reduction in the area of the MOSFET as the gate length (minimum feature length) is reduced.

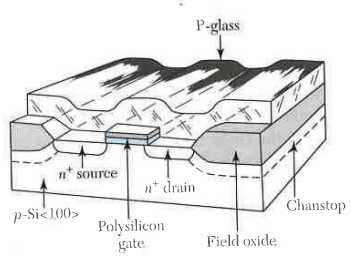


Fig. 15 Perspective view of an n-channel MOSFET.[7]

lateral oxide isolation, there is no need for vertical isolation in the MOSFET, whereas a buried-layer n^+-p junction is required in the bipolar transistor. The doping profile in a MOSFET is not as complicated as that in a bipolar transistor and the control of the dopant distribution is also less critical. We consider the major process steps that are used to fabricate the device shown in Fig. 15.

To process an n-channel MOSFET (NMOS), the starting material is a p-type, lightly doped ($\sim 10^{15}$ cm^{-3}), ⟨100⟩-oriented, polished silicon wafer. The ⟨100⟩-orientation is preferred over ⟨111⟩ because it has an interface-trap density that is about one-tenth that of ⟨111⟩. The first step is to form the oxide isolation region using LOCOS technology. The

process sequence for this step is similar to that for the bipolar transistor. A thin-pad oxide (~35 nm) is thermally grown, followed by a silicon nitride (~150 nm) deposition (Fig. 16*a*).[7] The active device area is defined by a photoresist mask and a boron chanstop layer is then implanted through the composite nitride-oxide layer (Fig. 16*b*). The nitride layer not covered by the photoresist mask is subsequently removed by etching. After stripping the photoresist, the wafer is placed in an oxidation furnace to grow an oxide (called the field oxide), where the nitride layer is removed, and to drive in the boron implant. The thickness of the field oxide is typically 0.5–1 μm.

The second step is to grow the gate oxide and to adjust the threshold voltage (see Section 6.2.3). The composite nitride-oxide layer over the active device area is removed, and a thin-gate oxide layer (less than 10 nm) is grown. For an enhancement-mode n-channel device, boron ions are implanted in the channel region, as shown in Fig. 16*c*, to increase the threshold voltage to a predetermined value (e.g., + 0.5V). For a depletion-mode n-channel device, arsenic ions are implanted in the channel region to decrease the threshold voltage (e.g., –0.5V).

The third step is to form the gate. A polysilicon is deposited and is heavily doped by diffusion or implantation of phosphorus to a typical sheet resistance of 20–30 Ω/□. This resistance is adequate for MOSFETs with gate lengths larger than 3 μm. For smaller

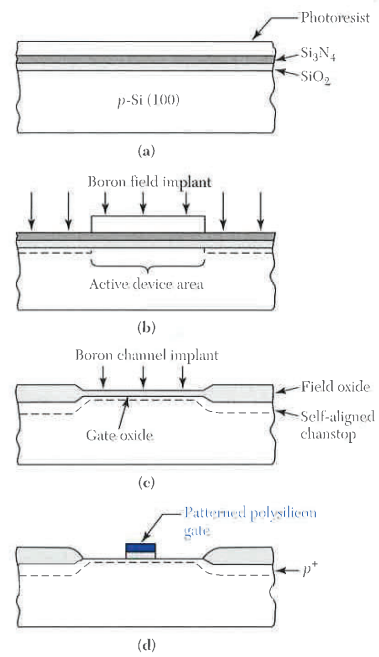


Fig. 16 Cross-sectional view of NMOS fabrication sequence.[7] (*a*) Formation of SiO_2, Si_3N_4, and photoresist layer. (*b*) Boron implant. (*c*) Field oxide. (*d*) Gate.

devices, polycide, a composite layer of metal silicide and polysilicon such as W-polycide, can be used as the gate materials to reduce the sheet resistance to about 1 Ω/□.

The fourth step is to form the source and drain. After the gate is patterned (Fig. 16*d*), it serves as a mask for the arsenic implantation (~30 keV, ~5×10^{15} cm^{-2}) to form the source and drain (Fig. 17*a*), which are self-aligned with respect to the gate[7]. At this stage, the only overlapping of the gate is due to lateral straggling of the implanted ions (for 30 keV As, $\sigma_\perp$ is only 5 nm). If low-temperature processes are used for subsequent steps to minimize lateral diffusion, the parasitic gate-drain and gate-source coupling capacitances can be much smaller than the gate-channel capacitance.

The last step is the metallization. A phosphorus-doped oxide (P-glass) is deposited over the entire wafer and is flowed by heating the wafer to give a smooth surface topography (Fig. 17*b*). Contact windows are defined and etched in the P-glass. A metal layer, such as aluminum, is then deposited and patterned. A cross-section view of the completed MOSFET is shown in Fig. 17*c*, and the corresponding top view is shown in Fig. 17*d*. The gate contact is usually made outside the active device area to avoid possible damage to the thin-gate oxide.

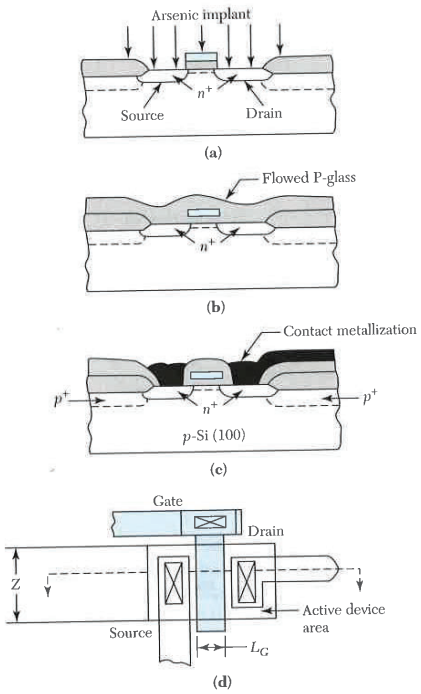


Fig. 17 NMOS fabrication sequence.[7] (*a*) Source and drain. (*b*) P-glass deposition. (*c*) Cross section of the MOSFET. (*d*) Top view of the MOSFET.

▶ EXAMPLE 4

What is the maximum gate-to-source voltage that a MOSFET with a 5 nm gate oxide can withstand. Assume that the oxide breaks down at 8 MV/ cm and the substrate voltage is zero.

SOLUTION

$V = \mathscr{E} \times d = 8 \times 10^6 \times 5 \times 10^{-7} = 4$ V. ◀

14.3.2 Memory Devices

Memories are devices that can store digital information (or data) in terms of *bits* (*bi*nary digi*ts*). Various memory chips have been designed and fabricated using NMOS technology. For most large memories, the random access memory (RAM) organization is preferred. In a RAM, memory cells are organized in a matrix structure and data can be accessed (i.e., stored, retrieved, or erased) in random order, independent of their physical locations. A static random access memory (SRAM) can retain stored data indefinitely as long as the power supply is on. The SRAM is basically a flip-flop circuit that can store one bit of information. A SRAM cell has four enhancement-mode MOSFETs and two depletion-mode MOSFETs. The depletion-mode MOSFETs can be replaced by resistors formed in undoped polysilicon to minimize power consumption.[8]

To reduce the cell area and power consumption, the dynamic random access memory (DRAM) has been developed. Figure 18*a* shows the circuit diagram of the one-transistor DRAM cell in which the transistor serves as a switch and one bit of information can be stored in the storage capacitor. The voltage level on the capacitor determines the state of the cell. For example, +1.5 V may be defined as logic 1 and 0 V defined as logic 0. The stored charge will be removed typically in a few milliseconds mainly because of the leakage current of the capacitors; thus, dynamic memories require periodic "refreshing" of the stored charge.

Figure 18*b* shows the layout of a DRAM cell, and Fig. 18*c* shows the corresponding cross section through AA′. The storage capacitor uses the channel region as one plate, the polysilicon gate as the other plate, and the gate oxide as the dielectric. The row line is a metal track to minimize the delay due to parasitic resistance (R) and parasitic capacitance (C), the RC delay. The column line is formed by n^+-diffusion. The internal drain region of the MOSFET serves as a conductive link between the inversion layers under the storage gate and the transfer gate. The drain region can be eliminated by using the double-level polysilicon approach shown in Fig. 18*d*. The second polysilicon electrode is separated from the first polysilicon capacitor plate by an oxide layer that is thermally grown on the first-level polysilicon before the second electrode has been defined. The charge from the column line can therefore be transmitted directly to the area under the storage gate by the continuity of inversion layers under the transfer and storage gates.

To meet the requirements of high-density DRAM, the DRAM structure has been extended to the third dimension with stacked or trench capacitors. Figure 19*a* shows a simple trench cell structure.[9] The advantage of the trench type is that the capacitance of the cell could be increased by increasing the depth of the trench without increasing the surface area of silicon occupied by the cell. The main difficulties of making trench-type cells are the etching of the deep trench, which needs a rounded bottom corner and the growth of a uniform thin dielectric film on trench walls. Figure 19*b* shows a stacked cell structure. The storage capacitance increases as a result of stacking the storage capacitor on top of the access transistor. The dielectric is formed using the thermal oxidation

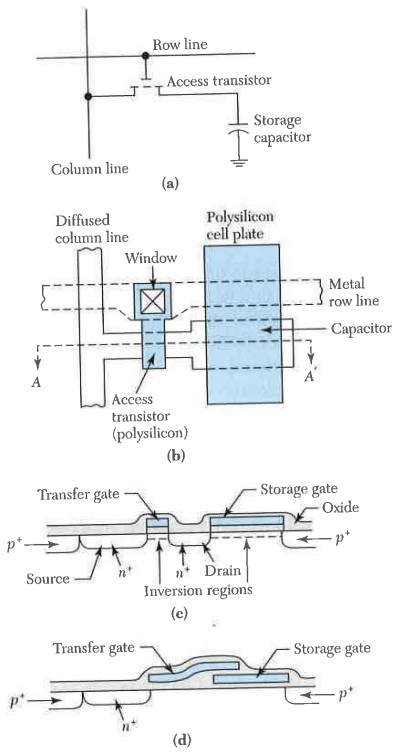


Fig. 18 Single-transistor dynamic random access memory (DRAM) cell with a storage capacitor.[8] (*a*) Circuit diagram. (*b*) Cell layout. (*c*) Cross section through *A-A′*. (*d*) Double-level polysilicon.

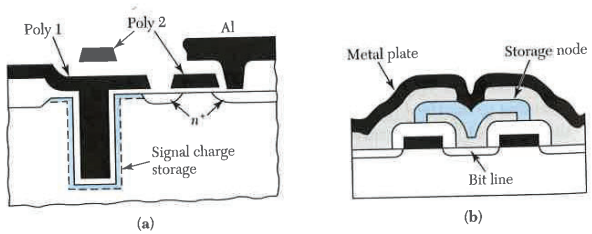


Fig. 19 (*a*) DRAM with a trench cell structure.[9] (*b*) DRAM with a single-layer stacked-capacitor cell.

or CVD nitride methods between the two-polysilicon plates. Hence, the stacked cell process is easier than the trench type process.

Figure 20 shows a 1 Gb DRAM chip. This memory chip uses 0.18 μm design rules. Trench capacitors and its peripheral circuits are in CMOS, which are considered in Section 14.3.3. The memory chip has an area of 390 mm^2 (14.3 mm × 27.3 mm) that contains over 2 billion components and operates at 2.5 V. This 1 Gb DRAM is mounted in an 88-pin ceramic package, which can provide adequate heat dissipation.

Both SRAM and DRAM are volatile memories, that is, they lose their stored data when power is switched off. Nonvolatile memories, on the other hand, can retain their data. Figure 21a shows a floating-gate nonvolatile memory, which is basically a conventional MOSFET that has a modified gate electrode. The composite gate has a regular (control gate) and a floating gate which is surrounded by insulators. When a large positive voltage is applied to the control gate, charge will be injected from the channel region through the gate oxide into the floating gate. When the applied voltage is removed, the injected charge can be stored in the floating gate for a long time. To remove this charge, a large negative voltage must be applied to the control gate, so that the charge will be injected back into the channel region.

Another version of the nonvolatile memory is the metal-nitride-oxide-semiconductor (MNOS) type shown inn Fig. 21*b*. When a positive gate voltage is applied, electrons can tunnel through the thin oxide layer (~2 nm) and be captured by the traps at the oxide-nitride interface, and thus become stored charges there. The equivalent circuit for both types of nonvolatile memories can be represented by two capacitors in series for the gate structure, as illustrated in Fig. 21*c*. The charge stored in the capacitor C_1 causes a shift in the threshold voltage, and the device remains at the higher threshold voltage-state (logic 1). For a well-designed memory device, the charge retention time can be over 100 years.

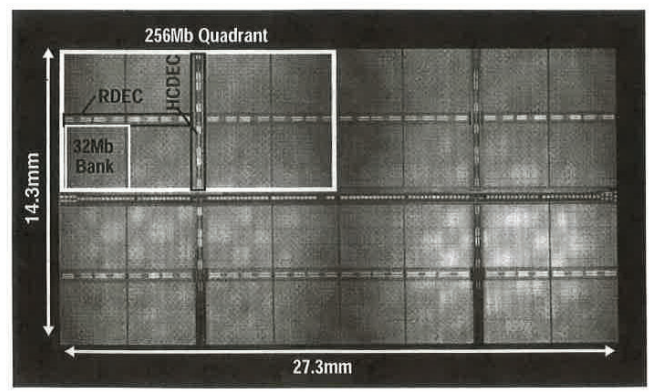


Fig. 20 A 1 Gb DRAM that contains over 2 billion components. (Photography courtesy of IBM/Siemens, 1999 IEEE Int. Solid State Circuit Conference.)

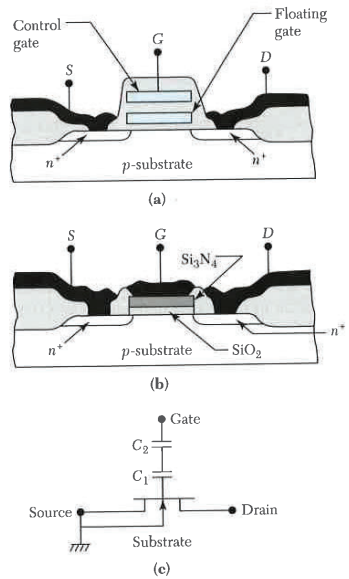


Fig. 21 Nonvolatile memory devices. (*a*) Floating-gate, nonvolatile memory. (*b*) MNOS nonvolatile memory. (*c*) Equivalent circuit of either type of nonvolatile memory.

To erase the memory (e.g., the store charge) and return the device to a lower threshold voltage state (logic 0), a gate voltage or other means (such as ultraviolet light) can be used.

The nonvolatile semiconductor memory (NVSM) has been extensively used in portable electronics systems, such as cellular phones and the digital cameras. Another interesting application is the chip card, also called IC card.

The top photo in Fig. 22 shows an IC card. The diagram at the bottom of Fig. 22 illustrates the nonvolatile memory device that stores the data that can be read and written through the bus to a central processing unit (CPU). In contrast to the limited volume (1 kbytes) inside a conventional magnetic tape card, the size of the nonvolatile memory can be increased to 16 kbytes, 64 kbytes, or even larger depending on the applications (e.g., you can store personal photos or finger prints). Through the IC card read/write machines, the data can be used in numerous applications, such as telecommunications (card telephone, mobile radio), payment transactions (electronic purse, credit card), pay television, transport (electronic ticket, public transport), health care (patient-data card), and access control. The IC card will play a central role in the global information and service society of the future.[10]

14.3.3 CMOS Technology

Figure 23*a* shows a CMOS inverter. The gate of the upper PMOS device is connected to the gate of the lower NMOS device. Both devices are enhancement-mode MOSFETs

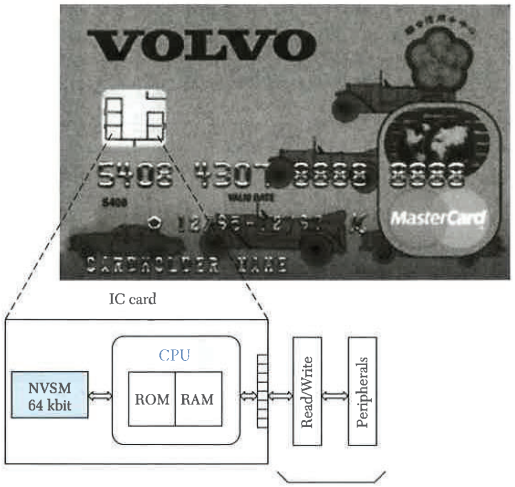


Fig. 22 An integrated-circuit (IC) card. The data stored in the NVSM can be accessed through the bus of the central processing unit (CPU). There are several metal pads connecting to the read/write machine. (Photography courtesy of Retone Information System Co., LTD.)

with the threshold voltage V_{Tp} less than zero for the PMOS device and V_{Tn} greater than zero for the NMOS device (typically the threshold voltage is about 1/4 V_{DD}). When the input voltage V_i is at ground or at small positive values, the PMOS device is turned on (the gate-to-ground potential of PMOS is $-V_{DD}$, which is more negative than V_{Tp}), and the NMOS device is off. Hence, the output voltage V_o is very close to V_{DD} (logic 1). When the input is at V_{DD}, the PMOS (with $V_{GS} = 0$) is turned off, and the NMOS is turned on ($V_i = V_{DD} > V_{Tn}$). Therefore, the output voltage V_o equals zero (logic 0). The CMOS inverter has a unique feature: in either logic state, one device in the series path from V_{DD} to ground is nonconductive. The current that flows in either steady state is a small leakage current, and only when both devices are on during switching does a significant current flow through the CMOS inverter. Thus, the average power dissipation is small, on the order of nanowatts. As the number of components per chip increases, the power dissipation becomes a major limiting factor. The low power consumption is the most attractive feature of the CMOS circuit.

Figure 23*b* shows a layout of the CMOS inverter, and Fig. 23*c* shows the device cross section along the *A-A′* line. In the processing, a *p*-tub (also called a *p*-well) is first implanted and subsequently driven into the *n*-substrate. The *p*-type dopant concentration must be high enough to overcompensate the background doping of the *n*-substrate. The subsequent processes for the *n*-channel MOSFET in the *p*-tub are identical to those described previously. For the *p*-channel MOSFET, $^{11}B^+$ or $^{49}(BF_2)^+$ ions are implanted into the *n*-substrate to form the source and drain regions. A channel implant of $^{75}As^+$ ions may be used to adjust the threshold voltage and a n^+-chanstop is formed underneath the field

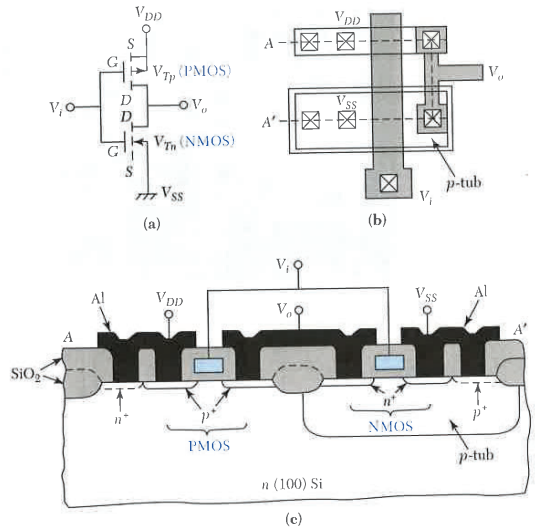


Fig. 23 Complementary MOS (CMOS) inverter. (*a*) Circuit diagram. (*b*) Circuit layout. (*c*) Cross section along dotted *A-A′* line of (*b*).

oxide around the *p*-channel device. Because of the *p*-tub and the additional steps needed to make the *p*-channel MOSFET, the number of steps to make a CMOS circuit is essentially double that to make an NMOS circuit. Thus, we have a trade-off between the complexity of processing and a reduction in power consumption.

Instead of the *p*-tub described above, an alternate approach is to use an *n*-tub formed in *p*-type substrate, as shown in Fig. 24*a*. In this case, the *n*-type dopant concentration must be high enough to overcompensate for the background doping of the p-substrate (i.e., $N_D > N_A$). In both the *p*-tub and the *n*-tub approach, the channel mobility will be degraded because mobility is determined by the total dopant concentration ($N_A + N_D$). A recent approach using two separated tubs implanted into a lightly doped substrate is shown in Fig. 24*b*. This structure is called a *twin* tub.[1] Because no overcompensation is needed in either of the twin tubs, higher channel mobility can be obtained.

All CMOS circuits have the potential for a troublesome problem called latchup that is associated with parasitic bipolar transistors (to see how this problem can occur, see Chapter 5). An effective processing technique to eliminate latchup problem is to use the deep-trench isolation, as shown[11] in Fig. 24*c*. In this technique, a trench with a depth deeper than the well is formed in the silicon by anisotropic reactive sputter etching. An oxide layer is thermally grown on the bottom and walls of the trench, which is then refilled by deposited polysilicon or silicon dioxide. This technique can eliminate latchup because the *n*-channel and *p*-channel devices are physically isolated by the refilled trench. The detailed steps for trench isolation and some related CMOS processes are now considered.

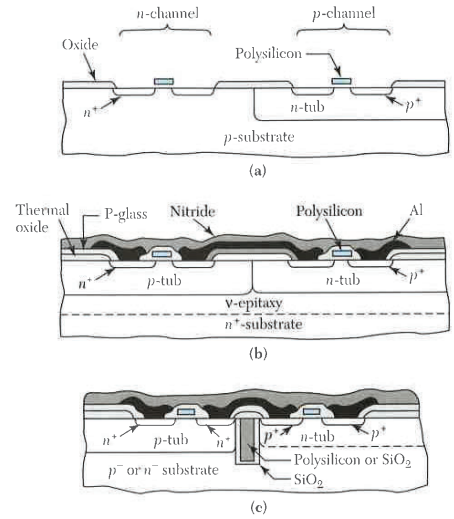


Fig. 24 Various CMOS structures. (*a*) *n*-tub. (*b*) Twin tub[1]. (*c*) Refilled trench.[11]

Well-Formation Technology

The well of a CMOS can be a single well, a twin well, or a retrograde well. The twin-well process exhibits some disadvantages, e.g., it needs high temperature processing (above 1050°C) and a long diffusion time (longer than 8 hours) to achieve the required depth of 2–3 μm. In this process, the doping concentration is highest at the surface and decreases monotonically with depth. To reduce the process temperature and time, high-energy implantation is used, i.e., implanting the ion to the desired depth instead of diffusion from the surface. Since the depth is determined by the implantion energy, we can design the well depth with different implantation energy. The profile of the well in this case can have a peak at a certain depth in the silicon substrate. This is called a retrograde well. Figure 25 shows a comparison of the impurity profiles in the retrograde well and the conventional thermal diffused well.[12] The energy for the *n*- and *p*-type retrograde wells is around 700 keV and 400 keV, respectively. As mentioned above, the advantage of the high-energy implantation is that it can form the well under low-temperature and short-time conditions, hence, it can reduce the lateral diffusion and increase the device density. The retrograde well can offer some additional advantages over the conventional well: (a) because of high doping near the bottom, the well resistivity is lower than that of the conventional well and the latchup problem can be minimized, (b) the chanstop can be formed at the same time as the retrograde well implantation, reducing processing steps and time, (c) higher well doping in the bottom can reduce the chance of punchthrough from the drain to the source.

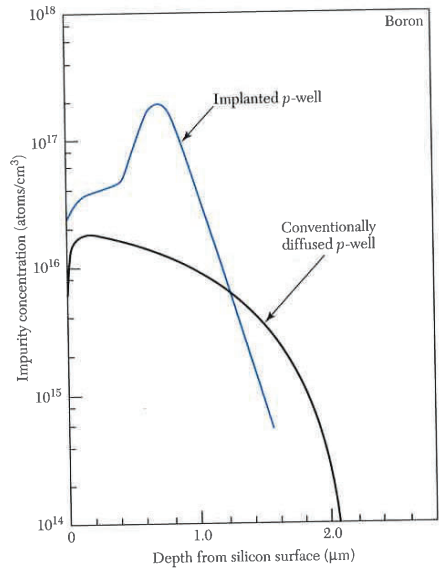


Fig. 25 Retrograded p-well implanted impurity concentration profile. Also shown is a conventionally diffused well.[12]

Advanced Isolation Technology

The conventional isolation process (Section 14.3.1) has some disadvantages that make it unsuitable for deep-submicron (0.25 μm and smaller) fabrications. The high-temperature oxidation of silicon and long oxidation time result in the encroachment of the chanstop implantation (usually boron for n-MOSFET) to the active region and cause V_T shift. The area of the active region is reduced because of the lateral oxidation. In addition, the field-oxide thickness in submicron-isolation spacings is significantly less than the thickness of field oxide grown in wider spacings. The trench isolation technology can avoid these problems and has become the mainstream technology for isolation. Figure 26 shows the process sequence for forming a deep (larger than 3 μm) but narrow (less than 2 μm) trench-isolation structure. There are four steps: patterning the area, trench etching and oxide growth, refilling with dielectric materials such as oxide or undoped polysilicon, and planarization. This deep trench isolation can be used in both advanced CMOS and bipolar devices and for the trench-type DRAM. Since the isolation material is deposited by CVD, it does not need a long-time or a high-temperature process, and it eliminates the lateral oxidation and boron encroachment problems.

Another example is the shallow trench (depth is less than 1 μm) isolation for CMOS, shown in Fig. 27. After patterning (Fig. 27a), the trench area is etched (Fig. 27b) and then re-filled with oxide (Fig. 27c). Before refilling, a chanstop implantation can be performed. Since the oxide has over filled the trench, the oxide on the nitride should be removed. Chemical-mechanical polishing (CMP) is used to remove the oxide on the nitride

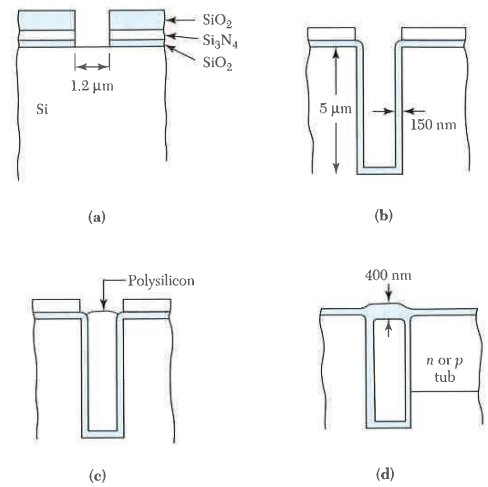


Fig. 26 Process sequence for forming a deep, narrow-trench, isolation structure. (a) Trench mask patterning. (b) Trench etching and oxide growth. (c) Polysilicon deposition to fill the trench. (d) planarization.

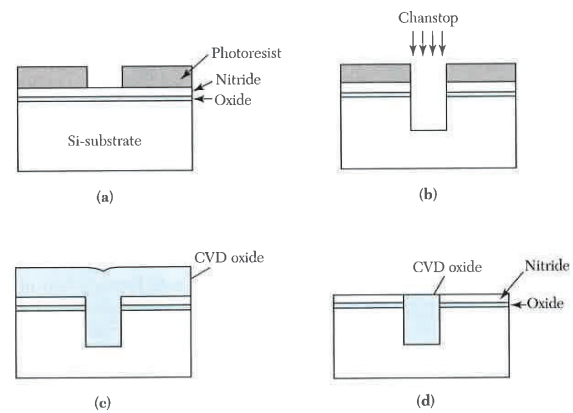


Fig. 27 A shallow trench isolation for CMOS. (a) Patterning with photoresist on nitride/oxide films. (b) Dry etching and chanstop implantation. (c) Chemical-vapor deposition (CVD) oxide to refill. (d) Planar surface after chemical-mechanical polishing (CMP).

and to get a flat surface (Fig. 27*d*). Due to its high resistance to polishing, the nitride acts as a stop-layer for the CMP process. After the polishing, the nitride layer and the oxide layer can be removed by H_3PO_4 and HF, respectively. This initial planarization step at the beginning is helpful for the subsequent polysilicon patterning and planarizations of the multilevel interconnection processes.

Gate-Engineering Technology

If we use n^+-polysilicon for both PMOS and NMOS gates, the threshold voltage for PMOS ($V_{TP} \cong -0.5$ to -1.0 V) has to be adjusted by boron implantation. This makes the channel of the PMOS a buried type, shown in Fig. 28*a*. The buried-type PMOS suffers serious short-channel effects as the device size shrinks to 0.25 μm and less. The most noticeable phenomena for short-channel effects are the V_T roll-off, drain-induced barrier lowering (DIBL), and the large leakage current at the off state so that even with the gate voltage at zero, leakage current flows through source and drain. To alleviate this problem, one can change n^+-polysilicon to p^+-polysilicon for PMOS. Due to the work function difference (there is a 1.0 eV difference from n^+- to p^+-polysilicon), one can obtain a surface p-type channel device without the boron V_T adjustment implantation. Hence, as the technology shrinks to 0.25 μm and less, dual-gate structures are required, i.e., p^+-polysilicon gate for PMOS, and n^+-polysilicon for NMOS (Fig. 28*b*). A comparison of V_T for the surface channel and the buried channel is shown in Fig. 29. We note that the V_T of surface channel rolls off slowly in the deep-submicron regime compared with the buried-channel device. This makes the surface-channel device with the p^+-polysilicon suitable for deep-submicron device operation.

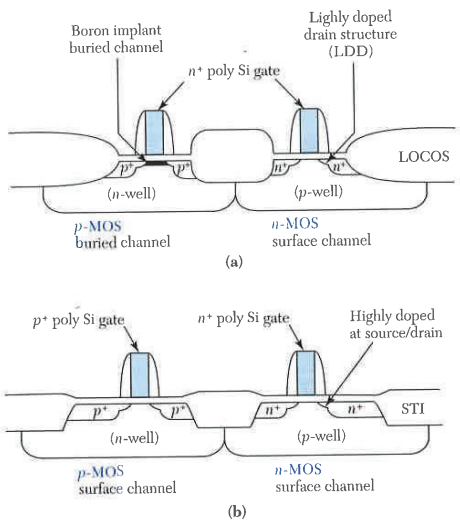


Fig. 28 (*a*) A conventional long-channel CMOS structure with a single-polysilicon gate (n^+). (*b*) Advanced CMOS structures with dual-polysilicon gates.

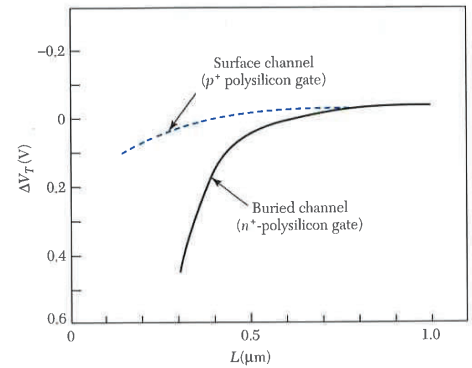


Fig. 29 The V_T roll-off for a buried type channel and for a surface type channel. The V_T drops very quickly as the channel length becomes less than 0.5 μm.

To form the p^+-polysilicon gate, ion implantation of BF_2^+ is commonly used. However, boron penetrates easily from the polysilicon through the oxide into the silicon substrate at high temperatures, resulting in a V_T shift. This penetration is enhanced in the presence of a F-atom. There are methods to reduce this effect: use of rapid thermal annealing to reduce the time at high temperatures and, consequently, the diffusion of boron; use of nitrided oxide to suppress the boron penetration, since boron can easily combine with nitrogen and becomes less mobile; and the making of a multilayer of polysilicon to trap the boron atoms at the interface of the two layers.

Figure 30 shows a microprocessor chip (Pentium 4) that has an area of about 200 mm^2 and contains 42 million components. This ULSI chip is fabricated using 0.18 μm CMOS technology with a six-level aluminum metallization.

14.3.4 BiCMOS Technology

BiCMOS is a technology that combines both CMOS and bipolar device structures in a single IC. The reason to combine these two different technologies is to create an IC chip that has the advantages of both CMOS and bipolar devices. As we know that CMOS exhibits advantages in power dissipation, noise margin, and packing density, whereas bipolar shows advantages in switching speed, current drive capability, and analog capability. As a result, for a given design rule, BiCMOS can have a higher speed than CMOS, better performance in analog circuits than CMOS, a lower power-dissipation than bipolar, and a higher component density than bipolar. Figure 31 shows the comparison of a BiCMOS and a CMOS logic gates. For a CMOS inverter, the current to drive (or to charging) the next loading, C_L, is the drain current I_{DS}. For a BiCMOS inverter, the current is $h_{fe}I_{DS}$, where h_{fe} is the current-gain of the bipolar transistor and I_{DS} is the base current of the bipolar transistor and is equal to the drain current of M_2 in the CMOS. Since h_{fe} is much larger than 1, the speed can be substantially enhanced.

BiCMOS has been widely used in many applications. In the early days, it was used in SRAM. At the present time, BiCMOS technology has been successfully developed for transceiver, amplifier, and oscillator applications in wireless-communication equipment.

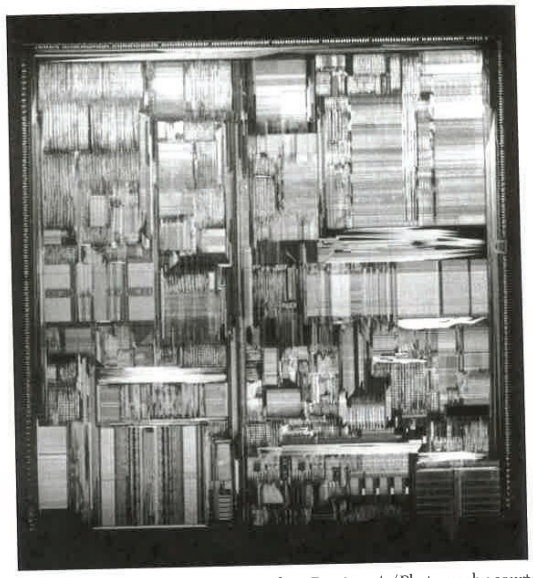


Fig. 30 Micrograph of a 32-bit microprocessor chip, Pentium 4. (Photography courtesy of Intel Corporation.)

Most of the BiCMOS processes are based on the CMOS process, with some modifications, such as adding masks for bipolar transistor fabrication. The following example is for a high-performance BiCMOS process based on the twin-well CMOS process, shown[13] in Fig. 32.

The initial material is a p-type silicon substrate, and then an n^+-buried layer is formed to reduce the collector's resistance. The buried p-layer is formed through ion implantation to increase the doping level to prevent punchthrough. A lightly doped n-epi layer is grown on the wafer and a twin-well process for the CMOS is performed. To achieve high performance of the bipolar transistor, four additional masks are needed. They are the

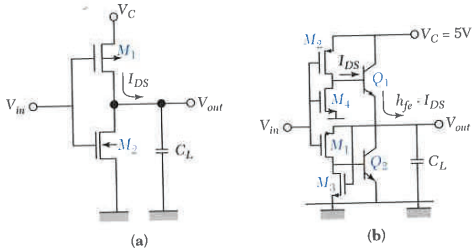


Fig. 31 (*a*) CMOS logic gate. (*b*) Bipolar CMOS (BiCMOS) logic gate.

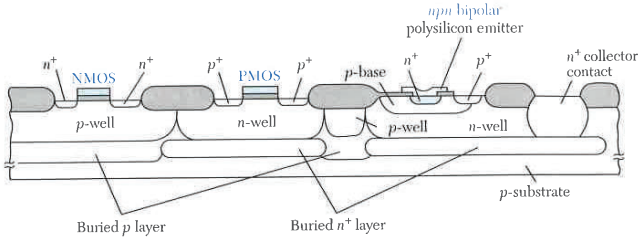


Fig. 32 Optimized BiCMOS device structure. Key features include self-aligned p and n^+ buried layers for improved packing density, separately optimized n- and p-well (twin-well CMOS) formed in an epitaxial layer with intrinsic background doping, and a polysilicon emitter for improved bipolar performance.[13]

buried n^+-mask, the collector deep-n^+-mask, the base p-mask, and the poly-emitter mask. In other processing steps, the p^+-region for base contact can be formed with the p^+-implant in the source/drain implantation of the PMOS, and the n^+-emitter can be formed with the source/drain implantation of the NMOS. The additional masks and longer processing time compared with a standard CMOS are the main drawbacks of BiCMOS. The additional cost should be justified by the enhanced performances of BiCMOS.

14.4 MESFET TECHNOLOGY

Recent advances in gallium arsenide processing techniques in conjunction with new fabrication and circuit approaches have made possible the development of "silicon-like" gallium arsenide IC technology. There are three inherent advantages of gallium arsenide compared with silicon: higher electron mobility, which results in lower series resistance for a given device geometry; higher drift velocity at a given electric field, which improves device speed; and the ability to be made semiinsulating, which can provide a lattice-matched dielectric-insulated substrate. However, gallium arsenide also has three disadvantages: a very short minority-carrier lifetime; lack of a stable, passivating native oxide; and crystal defects that are many orders of magnitude higher than in silicon. The short minority-carrier lifetime and the lack of high-quality insulating films have prevented the development of bipolar devices and delayed MOS technology using gallium arsenide. Thus, the emphasis of gallium arsenide IC technology is in the MESFET area, in which our main concerns are the majority carriers transport and the metal-semiconductor contact.

A typical fabrication sequence[14] for a high-performance MESFET is shown in Fig. 33. A layer of GaAs is epitaxially grown on a semiinsulating GaAs substrate, followed by an n^+-contact layer (Fig. 33*a*). A mesa etch step is performed for isolation (Fig. 33*b*), and a metal layer is evaporated for the source and drain ohmic contacts (Fig. 33*c*). A channel recess etch is followed by a gate recess etch and gate evaporation (Fig. 33*d* and *e*). After a liftoff process that removes the photoresist, shown in Fig. 33*e*, the MESFET is completed (Fig. 33*f*).

The n^+-contact layer reduces the source and drain ohmic contact resistances. Note that the gate is offset toward the source to minimize the source resistance. The epitaxial layer is thick enough to minimize the effect of surface depletion on the source and drain resistance. The gate electrode has maximal cross-sectional area with a minimal foot

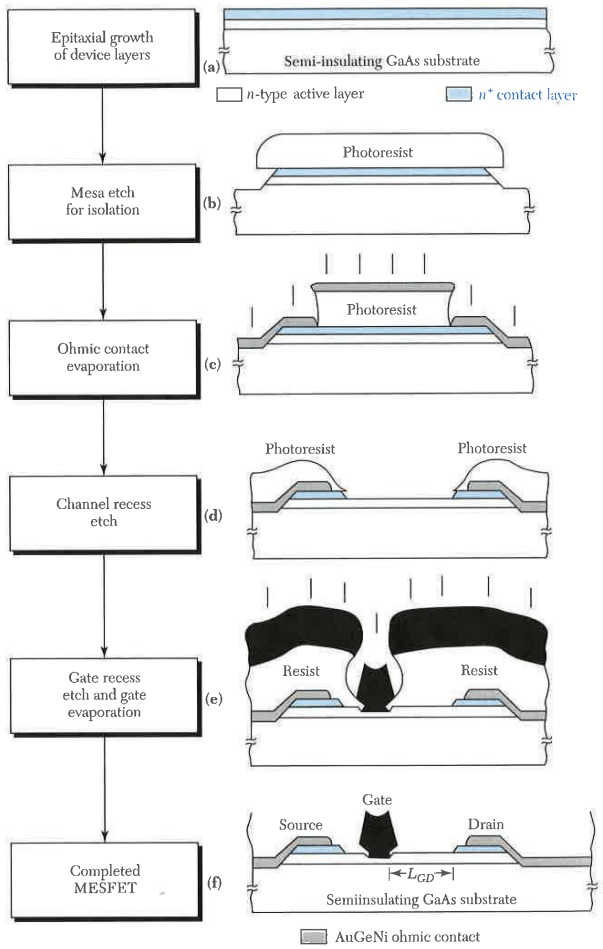


Fig. 33 Fabrication sequence of a GaAs MESFET.[14]

print, which provides low gate resistance and minimal gate length. In addition, the length L_{GD} is designed to be greater than the depletion width at gate-drain breakdown.

A representative fabrication sequence for a MESFET integrated circuit is shown[15] in Fig. 34. In this process, n^+-source and drain regions are self-aligned to the gate of each MESFET. A relatively light channel implant is used for the enhancement-mode switching

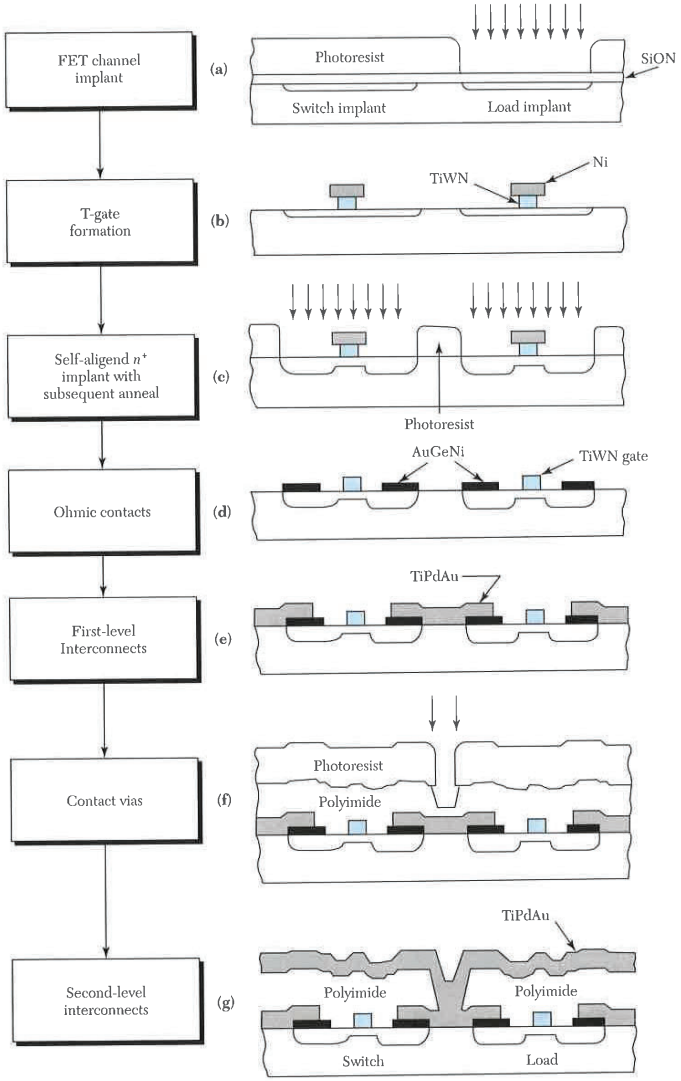


Fig. 34 Fabrication process for MESFET direct-coupled FET logic (DCFL) with active loads. Note that the n^+-source and drain regions are self-aligned to the gate.[15]

device and a heavier implant is used for the depletion-mode load device. A gate recess is usually not used for such digital IC fabrication because the uniformity of each depth has been difficult to control, leading to an unacceptable variation of the threshold voltage. This process sequence can also be used for a monolithic microwave integrated circuit (MMIC). Note that the gallium arsenide MESFET processing technology is similar to the silicon-based MOSFET processing technology.

Gallium arsenide ICs with complexities up to the large-scale integration level (~10,000 components per chip) have been fabricated. Because of the higher drift velocity (~20% higher than silicon), gallium arsenide ICs will have a 20% higher speed than silicon ICs that use the same design rules. However, substantial improvements in crystal quality and processing technology are needed before gallium arsenide can seriously challenge the preeminent position of silicon in ULSI applications.

14.5 CHALLENGES FOR MICROELECTRONICS

Since the beginning of the integrated-circuit era in 1959, the minimum device dimension, also called the minimum feature length, has been reduced at an annual rate of about 13% (i.e., a reduction of 30% every 3 years). According to the prediction by the International Technology Roadmap for Semiconductors[16], the minimum feature length will shrink from 130 nm (0.13 μm) in the year 2002 to 35 nm (0.035 μm) around 2014, as shown in Table 1. Also shown in Table 1 is the DRAM size. The DRAM has increased its memory cell capacity four times every 3 years and 64 Gbit DRAM is expected to be available in year 2011 using 50 nm design rules. The table also shows that the wafer size will increase to 450 mm (18 in. diameter) in 2014. In addition to the feature size reduction, challenges come from the device level, material level, and system level, discussed in the following subsections.

14.5.1 Challenges for Integration

Figure 35 shows the trends of power supply voltage V_{DD}, threshold voltage V_T, and gate oxide thickness d versus channel length for CMOS logic technology.[17] From the figure, one can find that the gate oxide thickness will soon approach the tunneling-current limit of 2 nm. V_{DD} scaling will slow down because of nonscalable V_T (i.e., to a minimum V_T of about 0.3 V due to subthreshold leakage and circuit noise immunity). Some challenges of the 180 nm technology and beyond are shown[18] in Fig. 36. The most stringent requirements are as follows.

TABLE 1 The Technology Generation[16] from 1997 to 2014

Year of the first product shipment	1997	1999	2002	2005	2008	2011	2014
Feature size (nm)	250	180	130	100	70	50	35
DRAM[a] size (bit)	256M	1G	—	8G	—	64G	—
Wafer size (mm)	200	300	300	300	300	300	450
Gate oxide (nm)	3–4	1.9–2.5	1.3–1.7	0.9–1.1	<1.0	—	—
Junction depth (nm)	50–100	42–70	25–43	20–33	15–30	—	—

[a]DRAM, dynamic random access memory.

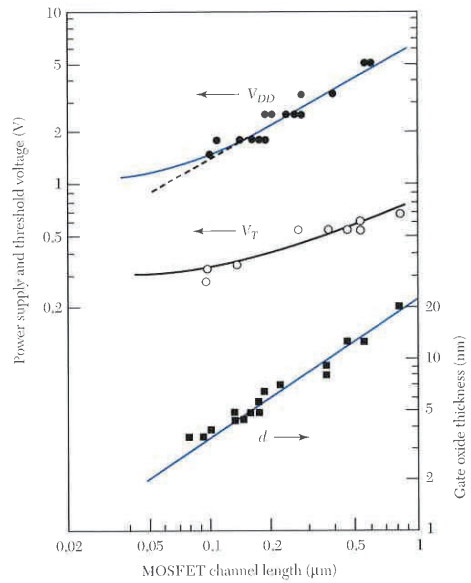


Fig. 35 Trends of power supply voltage V_{DD}, threshold voltage V_T, and gate oxide thickness d versus channel length for CMOS logic technologies. Points are collected from data published over recent years.[17]

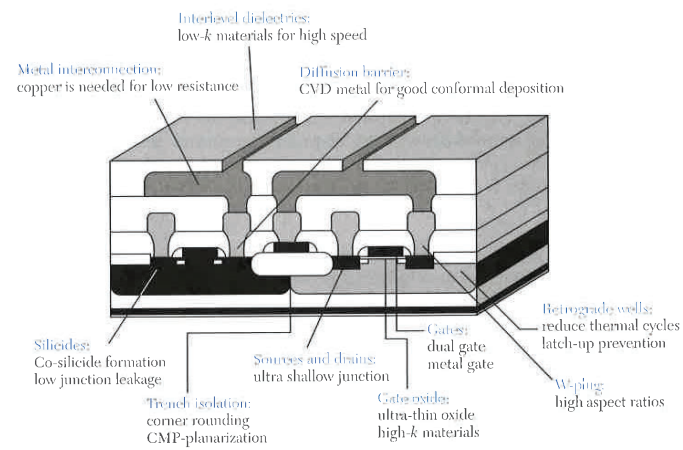


Fig. 36 Challenges for 180 nm and smaller MOSFET.[18]

Ultrashallow Junction Formation
As mentioned in Chapter 6, the short-channel effect happens as the channel length is reduced. This problem becomes critical as the device dimension is scaled down to 100 nm. To achieve an ultrashallow junction with low sheet resistance, low-energy (less than 1 keV) implantation technology with high dosage must be employed to reduce the short-channel effect. Table 1 shows the required junction depth versus the technology generation. The requirements of the junction for 100 nm are depths around 20–33 nm with a doping concentration of 1×10^{20} /cm^3.

Ultrathin Oxide
As the gate length shrinks below 130 nm, the oxide equivalent thickness of gate dielectric must be reduced around 2 nm to maintain the performance. However, if only SiO_2 (with a dielectric constant of 3.9) is used, the leakage through the gate becomes very high because of direct tunneling. For this reason, thicker high-k dielectric materials that have lower leakage current are suggested to replace oxide. Candidates for the short term are silicon nitride (with a dielectric constant of 7), Ta_2O_5 (25), and TiO_2 (60–100).

Silicide Formation
Silicide-related technology has become an integral part of submicron devices for reducing the parasitic resistance to improve device and circuit performance. The conventional Ti-silicide process has been widely use in 350–250 nm technology. However, the sheet resistance of a $TiSi_2$ line increases with decreasing line width, which limits the use of $TiSi_2$ in 100 nm CMOS applications and beyond. $CoSi_2$ or NiSi processes will replace $TiSi_2$ in the technology beyond 100 nm.

New Materials for Interconnection
To achieve high-speed operation, the RC time delay of the interconnection must be reduced. In Fig. 14 of Chapter 11 we have shown the delay as a function of feature size.[19] It is obvious that the gate delay decreases as the channel length decreases, meanwhile the delay resulting from interconnect increases significantly as the size decreases. This causes the total delay time to increase as the dimension of the device size scales down to 250 nm. Consequently, both high-conductivity metals, such as Cu, and low-dielectric-constant (low-k) insulators, such as organic (polyimide) or inorganic (F-doped oxide) materials offer major performance gains. Cu exhibits superior performance because of its high conductivity (1.7 μΩ-cm compared with 2.7 μΩ-cm of Al) and is 10–100 times more resistant to electromigration. The delay using the Cu and low-k material shows a significant decrease compared with that of the conventional Al and oxide. Hence, Cu with the low-k material is essential in multilevel interconnection for future deep-submicron technology.

Power Limitations
The power required merely to charge and discharge circuit nodes in an IC is proportional to the number of gates and the frequency at which they are switched (clock frequency). The power can be expressed as $P \equiv 1/2CV^2 nf$, when C is the capacitance per device, V is the applied voltage, n is the number of devices per chip, and f is the clock frequency. The temperature rise caused by this power dissipation in an IC package is limited by the thermal conductivity of the package material, unless auxiliary liquid or gas cooling is used. The maximum allowable temperature rise is limited by the bandgap of the semiconductor (~100°C for Si with a bandgap of 1.1 eV). For such a temperature rise, the maximum power dissipation of a typical high-performance package is about 10 W. As a result, we must limit either the maximum clock rate or the number of gates on a chip. As an example, in an IC containing 100 nm MOS devices with $C = 5 \times 10^{-2}$ fF, running at a 20 GHz clock rate, the maximum number of gates we can have is about 10^7 if we assume a 10% duty cycle. This is a design constraint fixed by basic material parameters.

SOI Integration
Mentioned in Section 14.2.2 was the isolation of the SOI wafer. Recently SOI technology has received more attention. The advantages of the SOI integration become significant as the minimum feature length approaches 100 nm. From the process point of view, SOI does not need the complex well structure and isolation processes. In addition, shallow junctions are directly obtained through the SOI film thickness. There is no risk of nonuniform interdiffusion of silicon and Al in the contact regions because of oxide isolation at the bottom of the junction. Hence, the contact barrier is not necessary. From the device point of view, the modern bulk silicon device needs high doping at the drain and substrate to eliminate short-channel effects and punch-through. This high doping results in high capacitance when the junction is reversed bias. On the other hand, in SOI, the maximum capacitance between the junction and substrate is the capacitance of the buried insulator whose dielectric constant is three times smaller than that of silicon (3.9 versus 11.9). Based on the ring oscillator performance, the 130 nm SOI CMOS technology can achieve 25% faster speed or require 50% less power compared to a similar bulk technology.[20] SRAM, DRAM, CPU, and rf CMOS have all been successfully fabricated using SOI technology. Therefore SOI is a key candidate for the future system-on-a-chip technology, considered in the following section.

▶ EXAMPLE 5

For an equivalent oxide thickness of 1.5 nm, what will be the physical thickness when high-k materials nitride ($\varepsilon_i/\varepsilon_0 = 7$), Ta_2O_5 (25), or TiO_2 (80) are used?

SOLUTION For nitride,

$$\left(\frac{\varepsilon_{ox}}{1.5}\right) = \left(\frac{\varepsilon_{\text{nitride}}}{d_{\text{nitride}}}\right),$$

$$d_{\text{nitride}} = 1.5\left(\frac{7}{3.9}\right) = 2.69 \text{ nm}.$$

Using the same calculation, we obtain 9.62 nm for Ta_2O_5 and 10.77 nm for TiO_2. ◀

14.5.2 System-On-A-Chip

The increased component density and improved fabrication technology have helped the realization of the system-on-a-chip (SOC), that is, an IC chip that contain a complete electronic system. The designers can build all the circuitry needed for a complete electronic system, such as a camera, radio, television, or personal computer (PC), on a single chip. Figure 37 shows the SOC application in the PC's mother-board. Components (11 chips in this case) once found on boards are becoming virtual components on the chip at the right.[21]

There are two obstacles in the realization of the SOC. The first is the huge complexity of the design. Since the component board is presently designed by different companies and different design tools, it is difficult to integrate them into one chip. The other

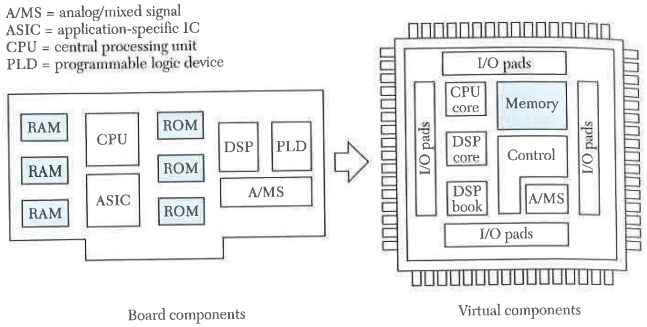


Fig. 37 System-on-a-chip of a conventional personal computer mother-board.[21]

is the difficulty of fabrication. In general, the fabricating processes of the DRAM are significantly different from those of logic IC (e.g., CPU). Speed is the first priority for the logic, whereas leakage of the stored charge is the priority for memory. Therefore, multilevel interconnection using five to six levels of metals is essential for logic IC to improve the speed. However, DRAM needs only two to three levels. In addition, to increase the speed, a silicide process must be used to reduce the series resistance, and ultrathin gate oxide is needed to increase the drive current. These requirements are not critical for the memory.

To achieve the SOC goal, an embedded DRAM technology is introduced, i.e., to merge logic and DRAM into a single chip with compatible processes. Figure 38 shows the schematic cross section of the embedded DRAM, including the DRAM cells and the logic CMOS devices.[22] Some processing steps are modified as a compromise. The trench-type capacitor, instead of the stacked type, is used so that there is no height difference in the DRAM cell structure. In addition, multiple gate oxide thicknesses exist on the same wafer to accommodate multiple supply voltages and/or combine memory and logic circuits on one chip.

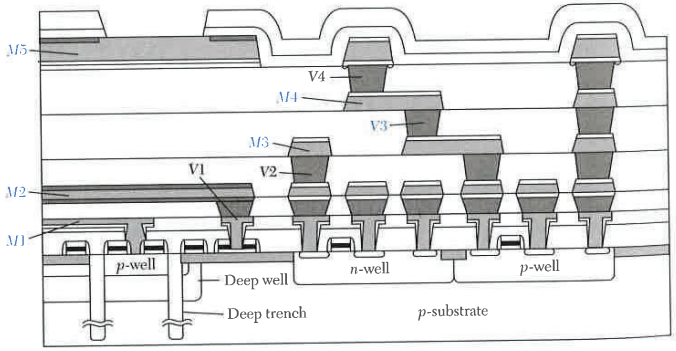


Fig. 38 Schematic cross section of the embedded DRAM including DRAM cells and logic MOSFETs. There is no height difference in the trench capacitor cell because of the DRAM cell structure. *M*1 to *M*5 are metal interconnections, and *V*1 to *V*4 are via holes.[22]

SUMMARY

In this chapter we considered processing technologies for passive components, active devices, and IC. Three major IC technologies based on the bipolar transistor, the MOSFET, and the MESFET were discussed in detail. It appears that the MOSFET will be the dominant technology at least until 2014 because of its superior performance compared with the bipolar transistor. For 100 nm CMOS technology, a good candidate is the combination of an SOI-substrate with interconnections using Cu and low-*k* materials.

Because the rapid reduction in feature length, the technology will soon reach its practical limit as the channel length is reduced to about 20 nm. What will be the device beyond the CMOS is the question being asked by research scientists. Major candidates include many innovative devices based on quantum mechanical effects. This is because when the lateral dimension is reduced to below 100 nm, depending on the materials and the temperature of operation, electronic structures will exhibit nonclassical behaviors. The operation of such devices will be on the scale of single-electron transport. This approach has been demonstrated by the single-electron memory cell. The realization of such systems with trillions of components will be a major challenge beyond CMOS.[23]

REFERENCES

1. For a detailed discussion on IC process integration, see C. Y. Liu and W. Y. Lee, "Process Integration," in C. Y. Chang and S. M. Sze, Eds., *ULSI Technology*, McGraw-Hill, New York, 1996.
2. T. Tachikawa, "Assembly and Packaging," in C. Y. Chang and S. M. Sze, Eds., *ULSI Technology*, McGraw-Hill, New York, 1996.
3. T. H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*, Cambridge Univ. Press, Cambridge, U.K., 1998, Ch. 2.
4. D. Rise, "Isoplanar-S Scales Down for New Heights in Performance," *Electronics*, **53**, 137 (1979).
5. T. C. Chen, et. al., "A submicrometer High-Performance Bipolar Technology," *IEEE Electron. Device Lett.*, **10**(8), 364, (1989).
6. G. P. Li et. al., "An Advanced High-Performance Trench-Isolated Self-Aligned Bipolar Technology," *IEEE Trans. Electron Devices*, **34**(10), 2246 (1987).
7. W. E. Beasle, J. C. C. Tsai, and R. D. Plummer, Eds., *Quick Reference Manual for Semiconductor Engineering*, Wiley, New York, 1985.
8. R. W. Hunt, "Memory Design and Technology," in M. J. Howes and D. V. Morgan, Eds., *Large Scale Integration*, Wiley, New York, 1981.
9. A. K. Sharma, *Semiconductor Memories—Technology, Testing, and Reliability*, IEEE, New York, 1997.
10. U. Hamann, "Chip Cards—The Application Revolution," *IEEE Tech. Dig. Int. Electron Devices Meet.*, p. 15, 1997.
11. R. D. Rung, H. Momose, and Y. Nagakubo, "Deep Trench Isolation CMOS Devices," *IEEE Tech. Dig. Int. Electron. Devices Meet.*, p. 237, 1982.
12. D. M. Bron, M. Ghezzo, and J. M. Primbley, "Trends in Advanced CMOS Process Technology," *Proc. IEEE*, p. 1646, (1986).
13. H. Higuchi, et al., "Performance and Structure of Scaled-Down Bipolar Devices Merge with CMOSFETs," *IEEE Tech.Dig. Int. Electron. Devices Meet.*, 694, 1984.
14. M. A. Hollis and R. A. Murphy, "Homogeneous Field-Effect Transistors," in S. M. Sze, Ed., *High-Speed Semiconductor Devices*, Wiley, New York, 1990.

15. H. P. Singh, et al., "GaAs Low Power Integrated Circuits for a High Speed Digital Signal Processor," *IEEE Trans. Electron Devices*, **36**, 240 (1989).

16. *International Technology Roadmap for Semiconductor (ITRS)*, Semiconductor Ind. Assoc., San Jose, 1999.

17. Y. Taur and E. J. Nowak, "CMOS Devices below 0.1 μm: How High Will Performance Go?" *IEEE Tech. Dig. Int. Electron Devices Meet.*, 215, 1997.

18. L. Peters, "Is the 0.18 μm Node Just a Roadside Attraction," *Semicond. Int.*, **22**, 46 (1999).

19. M. T. Bohr, "Interconnect Scaling—The Real Limiter to High Performance ULSI," *IEEE Tech. Dig. Int. Electron Devices Meet.*, p. 241, 1995.

20. E. Leobandung, et al.,"Scalability of SOI Technology into 0.13 μm 1.2 V CMOS Generation," *IEEE Int. Electron Devices Meet.*, p. 403, 1998.

21. B. Martin, "Electronic Design Automation," *IEEE Spectr.*, **36**, 61 (1999).

22. H. Ishiuchi, et al., " Embedded DRAM Technologies," *IEEE Tech. Dig. Int. Electron Devices Meet.*, p. 33, 1997.

23. S. Luryi, J. Xu, and A. Zaslavsky, Eds, *Future Trends in Microelectronics*, Wiley, New York, 1999.

▶ PROBLEMS (* DENOTES DIFFICULT PROBLEMS)

FOR SECTION 14.1 PASSIVE COMPONENTS

1. For a sheet resistance of 1 kΩ/□, find the maximum resistance that can be fabricated on a 2.5 × 2.5-mm chip using 2 μm lines with a 4 μm pitch (i.e., distance between the centers of the parallel lines).
2. Design a mask set for a 5 pF MOS capacitor. The oxide thickness is 30 nm. Assume that the minimum window size is 2 × 10 μm and the maximum registration errors are 2 μm.
3. Draw a complete step-by-step set of masks for the spiral inductor with three turns on a substrate.
4. Design a 10 nH square spiral inductor in which the total length of the interconnect is 350 μm; the spacing between turns is 2 μm.

FOR SECTION 14.2 BIPOLAR TECHNOLOGY

5. Draw the circuit diagram and device cross section of a clamped transistor.
6. Identify the purpose of the following steps in self-aligned double-polysilicon bipolar structure: (a) undoped polysilicon in trench in Fig. 13*a*, (b) the poly 1 in Fig. 13*b*, and (c) the poly 2 in Fig. 13*d*.

FOR SECTION 14.3 MOSFET TECHNOLOGY

*7. In NMOS processing, the starting material is a *p*-type 10 Ω-cm <100>-oriented silicon wafer. The source and drain are formed by arsenic implantation of 10^{16} ions/cm^2 at 30 keV through a gate oxide of 25 nm. (a) Estimate the threshold voltage change of the device. (b) Draw the doping profile along a coordinate perpendicular to the surface and passing through the channel region or the source region.

8. (a) Why is <100>-orientation preferred in NMOS fabrication? (b) What are the disadvantages if too thin a field oxide is used in NMOS devices? (c) What problems occur if a polysilicon gate is used for gate lengths less than 3 μm? Can another material be substituted for polysilicon? (d) How is a self-aligned gate obtained and what are its advantages? (e) What purpose does P-glass serve?

*9. For a floating-gate nonvolatile memory, the lower insulator has a dielectric constant of 4 and is 10 nm thick. The insulator above the floating gate has a dielectric constant of 10 and is 100 nm thick. If the current density J in the lower insulators is given by $J = \sigma\mathcal{E}$, where $\sigma = 10^{-7}$ S/cm, and the current in the other insulator is negligibly small, find the threshold voltage shift of the device caused by a voltage of 10 V applied to the control gate for (a) 0.25 μs, and (b) a sufficiently long time that J in the lower insulator becomes negligibly small.

10. Draw a complete step-by step set of masks for CMOS inverter shown in Fig. 23. Pay particular attention to the cross section shown in Fig. 23*c* for your scale.

*11. A 0.5 μm digital CMOS technology has 5 μm wide transistors. The minimum wire width is 1 μm and the metallization layer consists of 1 μm thick aluminum. Assume that μ_n is 400 cm^2/V-s, d is 10-nm, V_{DD} is 3.3 V, and the threshold voltage is 0.6 V. Finally, assume that the maximum voltage drop that can be tolerated is 0.1 V when a 1 μm^2 cross section aluminum wire is carrying the maximum current that can be supplied by the NMOS transistor. How long a wire can be allowed? Use a simple square-law, long-channel model to predict the MOS current drive (resistivity of aluminum is 2.7×10^{-8} Ω-cm).

12. Plot the cross-sectional views of a twin-tub CMOS structure of the following stages of processing: (a) *n*-tub implant, (b) *p*-tub implant, (c) twin-tub drive-in, (d) nonselective p^+-source/drain implant, (e) selective n^+-source/drain implant using photoresist as mask, and (f) P-glass deposition.
13. Why do we use a p^+-polysilicon gate for PMOS?
14. What is the boron penetration problem in p^+-polysilicon PMOS? How would you eliminate it?
15. To obtain a good interfacial property, a buffered layer is usually deposited between the high-*k* material and substrate. Calculate the effective oxide thickness if the stacked gate dielectric structure is (a) a buffered nitride of 0.5 nm and (b) a Ta_2O_5 of 10 nm.
16. Describe the disadvantages of LOCOS technology and the advantages of shallow-trench isolation technology.

FOR SECTION 14.4 MESFET TECHNOLOGY

17. What is the purpose for the polyimide used in Fig. 34*f*?
18. What is the reason that it is difficult to make bipolar transistor and MOSFET in GaAs?

FOR SECTION 14.5 CHALLENGES FOR MICROELECTRONICS

19. (a) Calculate the *RC* time constant of a aluminum runner 0.5 μm thick formed on a thermally grown SiO_2 0.5 μm thick. The length and width of the runner are 1 cm and 1 μm, respectively. The resistivity of the runner is 10^{-5} Ω-cm. (b) What will be the *RC* time constant for a polysilicon runner ($R_\square$= 30 Ω/□) of identical dimension?
20. Why do we need multiple oxide thicknesses for a system-on-a-chip (SOC)?
21. Normally we need a buffered layer placed between a high-*k* Ta_2O_5 and the silicon substrate. Calculate the effective oxide thickness (EOT) when the stacked gate dielectric is Ta_2O_5 (k = 25) with a thickness of 75Å on a buffered nitride layer (k = 7 and a thickness of 10 Å). Also calculate EOT for a buffered oxide layer (k = 3.9, and a thickness of 5 Å).

Appendix A

List of Symbols

Symbol	Description	Unit
a	Lattice constant	Å
$\mathscr{B}$	Magnetic induction	Wb/m^2
c	Speed of light in vacuum	cm/s
C	Capacitance	F
$\mathscr{D}$	Electric displacement	C/cm^2
D	Diffusion coefficient	cm^2/s
E	Energy	eV
E_C	Bottom of conduction band	eV
E_F	Fermi energy level	eV
E_g	Energy bandgap	eV
E_V	Top of valence band	eV
$\mathscr{E}$	Electric field	V/cm
$\mathscr{E}_c$	Critical field	V/cm
$\mathscr{E}_m$	Maximum field	V/cm
f	Frequency	Hz(cps)
$F(E)$	Fermi–Dirac distribution function	
h	Planck constant	J·s
$h\nu$	Photon energy	eV
I	Current	A
I_C	Collector current	A
J	Current density	A/cm^2
J_{th}	Threshold current density	A/cm^2
k	Boltzmann constant	J/K
kT	Thermal energy	eV
L	Length	cm or µm
m_0	Electron rest mass	kg
m_n	Electron effective mass	kg
m_p	Hole effective mass	kg
$\bar{n}$	Refractive index	
n	Density of free electrons	cm^{-3}
n_i	Intrinsic carrier concentration	cm^{-3}

(continued)